University of Nottingham
UK | CHINA | MALAYSIA

Systematic Purchase Behaviour in Big Transactional Data: Measurement Innovations and Applications

Thesis submitted to the University of Nottingham for the degree of **Doctor of Philosophy, June 2024.**

Roberto Javier Mansilla Lobos

20195506

Supervised by

Prof. Andrew Smith
Dr. Gavin Smith

Signature _____

Date _____ / _____ / _____

Abstract

With the rapid advancement in technology and the emergence of new sources of data, consumer buying behaviours have become increasingly dynamic. This has created a growing need for new measures and models to understand and predict complex buying patterns, aiming to enhance customer experience, satisfaction, and loyalty.

This thesis seeks to address this need through three interconnected studies that utilise a combination of traditional statistics and modern machine learning models and techniques. The objective is to explore systematic purchase behaviour (SPB) measurements and their real-world applications. Each study builds upon the previous one to provide a comprehensive understanding of SPB and its implications.

The primary objective is to develop a new measure of SPB by directly assessing the predictability aspect of basket composition. The thesis demonstrates the effectiveness of the proposed measure using both synthetic and real-world data sets. It also highlights the limitations of existing measures and introduces a new measure called *bundle entropy* (BE), which provides a precise indication of predictability, with zero denoting SPB and one indicating total unpredictability. The research also explores real-world applications of BE using two different large transactional datasets from leading UK retailers. The research delves into SPB at various levels of ag-

gregation, providing novel insights into consumers' choices across different retail settings. The final aim of the research is to provide a comprehensive understanding of the main drivers of SPB by analysing variables from historical transactional, demographic, and psychographic data. Machine learning models and variable importance methods are used to understand the influence of each group of variables on SPB.

This research endeavours to advance our understanding of consumer behaviour dynamics and to explore the broader implications of its findings. It emphasises the significant potential of big transactional data, particularly individual loyalty card data, in various aspects of consumer research, including forecasting buying behaviour and explanatory modelling. Additionally, the thesis acknowledges the different limitations encountered within the studies and the common challenges presented by big data. It concludes by offering actionable recommendations and suggesting potential areas for future scholarly inquiry in this field, thereby contributing to the ongoing discourse on consumer behaviour research.

List of Publications

Several cross-disciplinary papers have been published since I initiated my PhD program.

1. Mansilla, R., Smith, A., Smith, G. and Goulding, J. (2024). The relative power of behavioural, demographic, and psychographic variables as predictors of systematic purchase behaviour. British Academy of Management. [Accepted]

2. Mansilla, R., Long, G., Welham, S., Harvey, J., Evgeniya, L., Nica-Avram, G., Smith, G., Salt, D., Smith, A., and Goulding, J. (2024). Identifying iodine deficiencies from dietary transitions using shopping data. Scientific Reports (Nature).

3. Harvey, J., Long, G., Mansilla, R., Welham, S., Rose, P., Thomas, M., Milligan, G., Dolan, E., Parkes, J., Goulding, J. (2023). Who consumes anthocyanins and anthocyanidins? Mining national retail data to reveal the influence of socioeconomic deprivation and seasonality on polyphenol dietary intake. In 2023 IEEE International Conference on Big Data (Big Data).

4. Long G., Nica-Avram G., Harvey J., Mansilla R., Welham S., Lukinova E., and Goulding J. (2023). Predicting health related deprivation using loyalty card digital footprints. International Journal of Population Data Science.

5. Mansilla, R., Long, G., Welham, S., Harvey, J., Evgeniya, L., Nica-Avram, G., Smith, G., Salt, D., Smith, A., and Goulding, J. (2023). Identifying and understanding dietary transitions and their impact on nutrient deficiency: An exploratory analysis using loyalty card digital footprints. International Journal of Population Data Science.

6. Mansilla, R., Smith, G., Smith, A., and Goulding, J. (2022). Bundle entropy as an optimised measure of consumers' systematic product choice combinations in mass transactional data. In 2022 IEEE International Conference on Big Data (Big Data), pages 1044–1053.

7. Smith, G., Mansilla, R., and Goulding, J. (2020). Model Class Reliance for Random Forests. In Advances in Neural Information Processing Systems, pages 22305–22315.

8. Hurtado, D. E., Chavez, J. A., Mansilla, R., Lopez, R., and Abusleme, A. (2020). Respiratory volume monitoring: A machine-learning approach to the non-invasive prediction of tidal volume and minute ventilation. IEEE Access.

Under Review

1. Mansilla, R., Smith, A., Smith, G., and Goulding, J. (2023). Systematic purchase behaviour and healthy choices across online and offline channels: Insights from transactional data. Journal of Business Research.

2. Long, G., Nica-Avram, G., Harvey, J., Evgeniya, L., Mansilla, R., Welham, S., Engelmann, G., Dolan, E., Makokoro K., Thomas, M., Powell E., and Goulding, J. (2024). Machine learning on national shopping data reliably estimates childhood obesity prevalence and socio-economic deprivation. Food Policy.

Acknowledgements

"This thesis is dedicated with love and gratitude to my wife, my unborn baby Martina, my parents, and my brother. Each of you has played a significant role in my life and this journey, and for that, I am forever grateful."

I would like to express my gratitude to everyone who has been a part of my PhD journey because nobody conquers anything alone. It has been an incredible and challenging experience that I never thought I would undertake.

Firstly, I want to thank my brilliant supervisors for their unwavering guidance and support since day one. Despite challenging circumstances (COVID-19), they never made me feel alone. Andrew, you were the first person I met at Nottingham. Our conversation that day marked the beginning of everything, the moment I realised I had made one of the best decisions of my life. Thank you for accepting me into the MSc and PhD programs and for allowing me to be a part of what I now consider my second family, The N/LAB. Gavin, I will always remember our meetings in various coffee shops in Nottingham, where you patiently introduced me to the initial ideas of MCR. Thank you for your dedication and for pushing me beyond my limits in areas I never thought I could be a part of. Your passion for machine-learning topics has been a great source of motivation and inspiration.

James, even though you were not officially my supervisor, I cannot remember a single PhD meeting where you were not there for guidance and support. What I am truly grateful for is seeing the potential in myself that I couldn't see. Your encouragement to pursue this PhD (more than once) is something for which I will always be thankful.

Thanks to the N/LAB group — John, Vanja, Georgi, Maddy, Evgeniya, Lizzie, Gavin, Gregor, Jo, Sam, and all the new members for their support, great ideas, camaraderie and for making me feel at home even in a foreign land. Also, to my old friends Javier, Diego, Roberto, and Javier C. Thank

you all for a lifetime of invaluable friendship and for always wishing me the best. To my newest friends Mauricio, Camila, Paulina, Emilio, and Daniel, your kind support has shown me that friendship knows no borders.

I am deeply grateful to my wonderful parents, Roberto and Nolfa. Without your unwavering support, none of this would have been possible. Dad, your life story has been one of the greatest inspirations in my life. Thank you for being not only an exceptional father but also a true friend. Your guidance and the values you've instilled in me have been instrumental in helping me pursue my dreams and carve out my own path. Your pride in my every achievement means the world to me. Mom, you've been my number one cheerleader since I was a little kid. Your unconditional support and the comfort of your presence have been my rock. Thank you for always lending me your ear and offering me your wise counsel. I'd also like to extend my gratitude to my brother, Rodrigo, for his pure love and support. His presence and the cherished memories we share always make me feel close to home and family. Also, thank you and Romina for blessing Natalia and me with our dear Antonia, who brings immense joy and love into our lives.

A heartfelt thank you to my wife, Natalia. You are my partner, my friend, and my inspiration. Thank you for encouraging and supporting me in this greatest challenge every step of the way. Without you, this path would have been infinitely more challenging. Your words of support and motivation have been my guiding light during moments of anxiety and stress. Also, thank you for giving the most wonderful news of carrying our little, loving baby Martina; both of you have been the driving force behind my last push.

I conclude by thanking God for being the rock of my life and for granting me the strength, determination, and resilience required to overcome all the challenges presented during this journey and in my life.

—

"Esta tesis está dedicada con amor y gratitud a mi esposa, mi hija en camino Martina, mis padres y mi hermano. Cada uno de ustedes ha jugado un papel significativo en mi vida y en este desafío, y por eso, les estoy eternamente agradecido."

Quisiera expresar mi gratitud a todos los que han sido parte de mi viaje de doctorado porque nadie conquista nada solo. Ha sido una experiencia increíble y desafiante que nunca pensé que emprendería.

En primer lugar, quiero agradecer a mis brillantes supervisores por su inquebrantable guía y apoyo desde el primer día. A pesar de algunas circunstancias desafiantes (COVID-19), nunca me hicieron sentir solo. Andrew, fuiste la primera persona que conocí en Nottingham. Nuestra conversación ese día marcó el comienzo de todo, el momento en que me di cuenta de que había tomado una de las mejores decisiones de mi vida. Gracias por aceptarme en los programas de maestría y doctorado y por permitirme ser parte de lo que ahora considero mi segunda familia, The N/LAB. Gavin, siempre recordaré nuestras reuniones en varias cafeterías en Nottingham, donde pacientemente me introdujiste a las ideas iniciales de MCR. Gracias por tu dedicación y por empujarme más allá de mis límites en áreas en las que nunca pensé que podría participar. Tu pasión por los temas de aprendizaje automático ha sido una gran fuente de motivación e inspiración.

James, aunque no fuiste oficialmente mi supervisor, no puedo recordar una sola reunión de doctorado en la que no estuvieras presente para brindarme guía y apoyo. Pero lo que realmente te quiero agradecer es por ver en mí el potencial que yo no podía ver. Tu aliento para que persiguiera este doctorado (más de una vez) es algo por lo que siempre estaré agradecido.

Gracias al grupo N/LAB — John, Vanja, Georgi, Maddy, Evgeniya, Lizzie, Gavin, Gregor, Jo, Sam y todos los nuevos miembros por su apoyo, grandes ideas, camaradería y por hacerme sentir en casa incluso en un país extranjero. También, a mis viejos amigos Javier, Diego, Roberto y Javier C. Gra-

cias a todos por una vida de amistad invaluable y por siempre desearme lo mejor. A mis nuevos amigos Mauricio, Camila, Paulina, Emilio y Daniel, su amable apoyo me ha demostrado que la amistad no conoce fronteras.

Estoy profundamente agradecido de mis maravillosos padres, Roberto y Nolfa. Sin su apoyo inquebrantable, nada de esto hubiera sido posible. Papá, tú historia de vida ha sido una de las mayores inspiraciones en mi vida. Gracias por ser no solo un padre excepcional sino también un verdadero amigo. Tu guía y los valores que me has inculcado han sido fundamentales para ayudarme a perseguir mis sueños y forjar mi propio camino. Tu orgullo por cada uno de mis logros significa el mundo para mí. Mamá, has sido mi fan número uno desde que era un niño. Tu apoyo incondicional y el consuelo de tu presencia han sido mi roca. Gracias por siempre prestarme tu oído y ofrecerme tus sabios consejos. Quisiera extender mi gratitud a mi hermano, Rodrigo, por su amor y apoyo inquebrantables. Su presencia y los recuerdos compartidos siempre me hacen sentir cerca de casa y de la familia. Agradecerle también por que junto a la Romi nos bendecirnos a Natalia y a mí con nuestra amada Antonia, que trae una inmensa alegría y amor a nuestras vidas.

Un sincero agradecimiento a mi esposa, Natalia. Eres mi compañera, mi amiga y mi inspiración. Gracias por alentarme y apoyarme en cada paso de este gran desafío. Sin ti, este viaje habría sido infinitamente más difícil. Tus palabras de apoyo y motivación han sido mi luz y guía durante momentos de ansiedad y estrés. También, gracias por darme la maravillosa noticia de llevar en tu vientre a nuestra pequeña y amada Martina; ambas han sido mi fuerza detrás de mi último empuje.

Concluyo agradeciendo a Dios por ser la roca de mi vida y por concederme la fuerza, determinación y resiliencia necesarias para superar todos los desafíos presentados durante este viaje y en mi vida.

Contents

List of Tables

List of Figures

Acronyms

BLE Basket Level Entropy.

BRE Basket Revealed Entropy.

CRISP-DM Cross-Industry Standard Process for Data Mining.

MCR Model Class Reliance.

SHAP SHapley Additive exPlanations.

SPB Systematic Purchase Behaviour.

Glossary

basket level entropy Refers to an entropy-based measurement used to assess predictability at the basket level.. 8

basket revealed entropy Refers to an entropy-based measurement used to assess systematic purchase behaviour at the sub-basket level.. 8

big five Refers to five broad dimensions of human personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.. 4

big data Refers to large and complex datasets.. 3

cross-industry standard process for data mining Refers to the standard approach to guide the process of extracting useful insights from data. It typically includes business knowledge, data understanding, data preparation, modelling, evaluation, and deployment.. 184

demographic Refers to statistical characteristics of a particular population.. 4

machine learning Refers to algorithms and statistical models that enable computers to learn and make predictions or decisions without being explicitly programmed.. 7

model class reliance Refers to how much a machine learning model's predictions depend on a specific feature or set of features.. 11

psychographic Refers to lifestyles, values, attitudes, interests, and opinions of a particular population.. 4

SHapley Additive exPlanations Refers to a method for interpreting machine learning models' output by determining each feature's importance to the final model.. 11

systematic purchase behaviour Refers to a person's stability and regularity of product selections over a period of time.. 5, 29

variety-seeking Refers to the tendency of consumers to seek out different products or services.. 4

Chapter 1

Introduction

Contents

1.1 Theoretical motivation

Understanding and predicting the behaviour of individuals and households is more important than ever for academics, business professionals, and policymakers. The ability to forecast and comprehend individual-level behaviour has been used in diverse domains, including human mobility (Song et al., 2010; Smith et al., 2014), advertising (Krumm, 2010), retailing (Guidotti et al., 2017), service provision (Jung et al., 2010), intelligent agents (Froehlich and Krumm, 2008) and more. However, predicting individual behaviour has a more extensive scope than just segment or market-level extrapolations of potential individual probabilities. Understanding regularity in other behaviours, such as systematic purchasing patterns (Guidotti et al., 2015), can provide novel market insights.

However, the current approaches often fail to measure the regularity and probability of products bought in combination rather than those purchased sequentially as single items. Therefore, more advanced individual and household predictions are required to provide more detailed insights into customer behavioural choices over time. The potential for advanced individual and household prediction is enormous, given that large-scale transactional datasets are routinely collected as part of our digital footprint and processed as part of loyalty programs and online purchase platforms.

While repeat purchase rates and probabilities are useful measures based on market-level aggregate analysis (Frisbie, 1980), understanding individual and household regular behaviour can offer even more valuable insights into customer needs. This knowledge can help scholars and decision-makers address customer needs more efficiently. The potential to utilise behavioural Big data for both academic and practical purposes is evident (Hossain et al., 2020; Foxall, 2001), yet this area of research is still largely unexplored.

Consumer buying behaviour is a complex field, with the majority of research focusing on Demographic and Psychographic factors. Studies have demonstrated the effectiveness and limitations of using demographics and psychographics to understand and predict consumer buying behaviour, yielding mixed results (Sandy et al., 2013).

On the one hand, demographic variables have been instrumental in providing insights into consumer behaviour, but they fail to capture its complete complexity and are not universally predictive (Bellman et al., 1999). On the other hand, psychographics offers a deeper understanding by exploring individuals' attitudes, values, and beliefs. However, their effectiveness varies across industries, product categories, and target audiences (e.g. Mulyono and Rusdarti 2020, Bosnjak et al. 2007). Among psychographics, much emphasis has been placed on personality traits, with the Big five framework (agreeableness, openness, conscientiousness, extroversion, and neuroticism) being one of the most popular. Numerous studies have found that personality traits are significant predictors of various buying behaviours, such as impulsiveness (Mowen, 2000), Variety-seeking (Sharma et al., 2010a), and brand and product loyalty (Lin, 2010). Although most studies have identified a significant relationship between buying behaviours and personality traits, the specific traits that are significant vary depending on the context.

Comparative studies have yielded mixed results, with demographics performing better in predicting some purchases while psychographics are better for others (Sandy et al., 2013). Despite this, personality traits can complement and enhance traditional demographic metrics in understanding buying behaviours.

The advancement of technology and massive behavioural datasets offer new opportunities to understand and predict buying behaviours. While there is

already some evidence supporting the effectiveness of behavioural data in both traditional (Ehrenberg, 1988; Bellman et al., 1999) and online settings (Bosnjak et al., 2007), there is still much more to be explored.

For example, in the retail sector, big behavioural data such as transactional data can provide years of purchasing information, enabling novel empirical research and insights into various aspects of consumer behaviour, market dynamics, and organisational performance. Since big transactional datasets can contain years of historical purchases at the individual/household level through loyalty programs, the potential of understanding different long and short-term buying patterns is evident. For instance, researchers have explored the use of big transactional data to identify diet transitions with high risks of nutrient deficiency (Mansilla et al., 2024a), as well as food insecurity and deprivation (Nica-Avram et al., 2021) and plastic bag usage (Lavelle-Hill et al., 2020). However, there is still limited research on leveraging big transactional data to examine and predict buying regularities over time within the retail field.

A study by Guidotti et al. (2015) is one of the first attempts to leverage big transactional data to examine and predict purchase systematic product choices over time, highlighting some interesting practical implications for retailers. In addition to that, Guidotti et al. (2015) presented the idea of Systematic purchase behaviour (SPB). This concept relates to the regularity and frequency of product selections made by customers within a specific period. This term is consistently used throughout the thesis.

While demographics, psychographics, and behavioural drivers have been used to understand and predict buying outcomes in isolation (Van Trijp et al., 1996; Di Crosta et al., 2021; Ali et al., 2022), their interactions have never been modelled together due to different challenges related to big data.

Integrating these types of datasets would uncover their interactions and highlight the most relevant predictors of systematic purchase behaviour, which marketers could leverage to identify specific groups of individuals for actionable interventions.

1.2 Research gap

Previous use of transactional data in consumer research: During the 1980s, the application of transactional data in consumer research was constrained due to the absence of individual identifiers. Researchers relied on macro-level sales figures and market trends, which restricted their comprehension of individual buying behaviour. Despite this, early studies laid the groundwork for future research in this field (Ehrenberg, 1988; Bawa et al., 1989).

With advancements in computing power and data storage in the 1990s, personalised analysis became possible, bringing in a new era of consumer research (Fader and Lodish, 1990). Recently, transactional data has been used to uncover complex hidden purchasing patterns that were previously difficult to identify. This shift has led to a more diverse and precise understanding of individual consumer behaviour (Guidotti et al., 2015, 2018). However, as this is still a relatively new area in consumer research, there is still more work to be done and new standards to define for exploring and understanding individuals' behaviours from this type of big data.

The potential of transactional data in understanding consumer buying behaviour: The analysis of transactional data has greatly improved with the advent of loyalty card programs, allowing for a more comprehensive understanding of consumer buying habits (Boussofiane, 1996).

Thanks to advancements in predictive analytics and Machine learning, it's now easier to gain deeper insights into consumer behaviour with an unprecedented level of granularity (Boone and Roehm, 2002; Tian et al., 2018; Guidotti et al., 2015). Retailers can now integrate both online and offline data sources to obtain a holistic view of consumer behaviour across various channels (Smith et al., 2004). Utilising basket analysis and cross-category purchase behaviour exploration allows them to understand consumer preferences better, leading to targeted marketing campaigns and personalised recommendations (Russell and Petersen, 2000a; Mild and Reutterer, 2003). Retailers can also optimise inventory management and promotional efforts by analysing purchase patterns over time. By using individual-level transactional data, personalised marketing strategies can be implemented to tailor to each customer's historical purchase behaviour (Asniar and Surendro, 2019). The utilisation of innovative technologies like machine learning and data mining further enhances the analysis of transactional data, providing hidden insights into different aspects of consumer research and future directions in the field. Although the potential for novel insights in different areas of consumer research is huge, the utilisation of these new tools and technologies has yet to be explored deeply.

The need for a parsimonious measure of systematic purchasing behaviour: Extensive research has been conducted on consumer purchasing patterns using both latent and explanatory models (Ehrenberg, 1988; Allenby and Lenk, 1994). These models aim to represent observed behaviours through unobserved causal factors, providing a simplified representation of each concept of interest. However, these models have limitations, such as the assumption of stationary conditions and the requirement of estimated parameters (Bhattacharya, 1997). While some research has centred on representing consumers' purchase dynamics behaviour by eval-

uating different factors like marketing variables, heterogeneity across individual preferences, variety-seeking, and consistent purchase, there is still a lack of comprehensive understanding (Givon, 1984; Chintagunta, 1999). In recent years, there has been a growing focus on predictive models of buying behaviours (Van Den Poel and Buckinx, 2005; Lo et al., 2016). While some models prioritise predictive accuracy and performance, others have attempted to prioritise explainability and comprehensive understanding. These attempts include Basket revealed entropy (BRE) (Guidotti et al., 2015) and Basket level entropy (BLE) (Nicolas-Sans and Ibáñez, 2021).

As mentioned before, Guidotti et al. (2015) introduces the concept of systematic purchase behaviour (SPB), a term used across the whole thesis to describe a person's regular buying patterns over time. To better understand SPB, let's consider a hypothetical purchase sequence outlined in Table 1.1. The data in the table clearly illustrates that *Person 1* consistently exhibits high SPB by regularly purchasing products w, x, and y together. On the other hand, *Person 5* demonstrates the least predictable buying behaviour since he purchases different products each time, making future predictions challenging.

When reviewing the purchase patterns of *Person 2, 3 and 4*, it is not immediately evident which of them has the most systematic purchasing behaviour, as each of them has made multiple purchases of at least one item.

While it may seem simple to evaluate the probability of each product or the repeat purchase rate for v, w, x, y, *or* z, assessing the evolution of combinations is more challenging. Current measures such as BRE and BLE have made progress in quantifying the predictability and consistency of individual purchases. However, there are still challenges in accurately

and intuitively assessing the consistent patterns or systematic tendencies in consumer purchases at various levels, such as the individual basket or the entire purchase sequence. This is particularly important for businesses and researchers seeking to gain insights into consumer buying behaviour while minimising complexity in the assessment process.

Table 1.1: The predictability of different purchase scenarios.

Systematic Purchase Behaviour	**Person**	**Set of Baskets**
Extremely high	Person 1	$[(wxy), (wxy), (wxy)]$
High	Person 2	$[(wxy), (wxz), (wxv)]$
Medium	Person 3	$[(wxy), (wxv), (xyz), (xyv)]$
Low	Person 4	$[(wx), (xy), (yz), (zv)]$
Extremely low	Person 5	$[(wxy), (abc), (stu)]$

Limited understanding of within-subject buying behaviours differences across retail channels: Numerous studies have delved into the various dissimilarities between in-store and online shopping environments and the impact they have on consumer behaviour (Ratchford et al., 2022; Campo et al., 2021). However, the majority of these studies have primarily concentrated on recognising the distinguishing features of these channels (Cimana, 2013; Grewal et al., 2004b) and how customers perceive them (Wang et al., 2016). While a few studies have explored the differences in purchasing behaviours between offline and online shopping, they have only compared these behaviours across different consumer groups (e.g. Chu et al. 2010). Other studies, like the focus of the second study in this thesis, have explored differences in purchasing behaviour within subjects across different channels (e.g. Kulkarni et al. 2012), but their data is quite outdated and limited to high-involvement or low purchase rate products. Thus, there remains a need for a comprehensive understanding of buying differences, especially systematic buying behaviour, within-subject across channels in the fast-moving consumer goods sector.

Mixed findings on factors driving systematic behaviours and related concepts: Consumer behaviour research has relied on demographic and psychographic variables to understand different aspects of buying behaviours (Bell and Lattin, 1998b; Baumeister, 2002). On the one hand, demographic predictors have undergone extensive study. While their impacts may vary across contexts, income consistently influences loyalty and product variety tendencies (Carlson et al., 2015; Klopotan et al., 2016). Gender differences are well studied, with women making more frequent and impulsive purchases (Rich and Jain, 1968; Henry, 2002). The influence of household composition varies (Mittal and Kamakura, 2001), and education's impact has evolved, particularly in the online sphere (Wood et al., 1985; Ghafoor et al., 2015).

On the other hand, psychographic predictors are a complex realm encompassing personality traits and attitudes that offer profound insights. Traits such as neuroticism, extroversion, and conscientiousness are intricately linked to impulsive buying and loyalty both offline (Mulyono and Rusdarti, 2020; Sharma et al., 2010b) and online (Bosnjak et al., 2007). Attitudes towards emotional states wield significant influence over decisions (Lin and Lin, 2009). While the specific insights gained may vary depending on the context, it's difficult to deny their impact on various buying behaviours.

Although demographic and psychographic variables are informative, incorporating behavioural variables from transactional records can substantially enhance our comprehension of purchasing behaviours, such as repeated product selections. Moreover, combining variables from diverse sources can augment the predictive models of consumer behaviour. These underexplored areas of research hold the potential to offer a more comprehensive understanding of consumer behaviour.

The potential of machine learning models: Traditional methods, including latent, stochastic, and segmentation models, have been widely used to understand and validate theoretical models of drivers for various purchasing behaviours (Ehrenberg, 1988; Sun and Wu, 2014; Ali et al., 2022). To investigate the linear and non-linear relationships between predictors and behaviour outputs, researchers generally apply different versions of Partial Least Squares or regression models (Schaninger, 1981; Bellman et al., 1999; Sorce et al., 2006; Brunelle and Grossman, 2022; Ali et al., 2022). However, they are not designed to find these relationships in large amounts as is required nowadays (Faraway and Augustin, 2018).

The recent advancements in big data analytics and machine learning have been crucial in overcoming the limitations of traditional statistical approaches. Machine learning techniques, in particular, demonstrate exceptional proficiency in scrutinising vast, multi-dimensional datasets, unveiling both linear and non-linear patterns within the data (Faraway and Augustin, 2018), and accurately predicting complex behaviours, such as churn (Khodabandehlou and Zivari Rahman, 2017), food waste (Panda and Dwivedi, 2020), future purchase (Martínez et al., 2020), product demand (Huber and Stuckenschmidt, 2020), among others.

Although machine learning models have not traditionally been utilised for understanding consumer behaviours due to their emphasis on prediction over explanation, recent advancements in the field, including the integration of SHapley Additive exPlanations (SHAP) values (Lundberg et al., 2017) and Model class reliance (MCR) (Fisher et al., 2019; Smith et al., 2020), have significantly improved the capacity of some models to provide insightful explanations. This progress has led to exceptional outcomes across a range of domains (Lavelle-Hill et al., 2021; Dolan et al., 2023a; Rodríguez-Pérez and Bajorath, 2020; Ljevar et al., 2021). As a result,

these developments offer great potential for uncovering complex relationships between predictors of buying behaviours (Asniar and Surendro, 2019).

1.3 Research aims and rationale of current study

This thesis aims to thoroughly examine and understand the practical applications and key determinants that influence systematic purchase behaviour. It utilises a range of quantitative methodologies, empirical testing, and statistical and explanatory analysis, drawing upon real-world datasets with nationwide coverage in the UK to contribute comprehensive insights into the dynamics of consumer research.

The following sections briefly describe the three studies contained in this thesis, stating their aims and rationales.

1.3.1 Study 1: Measuring Consumers' Systematic Purchase Behaviour in Retail

Understanding and quantifying the predictability of consumer systematic purchasing behaviour holds significant value for both researchers and practitioners. While predictability measures such as entropy have been thoroughly examined and utilised in various sectors, their application in the retail sector, particularly with complex multi-dimensional data sets, is not as widespread.

Though a few methods exist for estimating systematic purchase behaviour, they do not align with intuition and lack intuitive interpretability, poten-

tially leading to misunderstandings between analysts and decision-makers.

This study addresses these limitations by developing and evaluating *bundle entropy*, a novel measure designed to assess the predictability of basket composition across multiple transactions.

Study 1a focuses on establishing the theoretical foundations of *bundle entropy* by identifying the desired properties of an ideal measure and comparing its performance against existing metrics. The study highlights the inability of traditional measures to align with intuitive reasoning and their practical shortcomings, emphasizing the need for a more robust approach.

Study 1b evaluates *bundle entropy* empirically using two comprehensive real-world datasets: transactional records from a prominent UK retailer and data from a leading data science firm. Together, these datasets include over 2,000 households observed over two years. The study assesses the ability of *bundle entropy* and established measures to categorize customers into three behavioural groups (systematic, standard, and unsystematic) using statistical methods.

The findings demonstrate that *bundle entropy* is the only measure that satisfies the desired properties, providing an interpretable and robust framework for understanding systematic purchase behaviour. It outperforms existing metrics by accurately capturing patterns in basket composition. By bridging theoretical rigour with practical applications, *bundle entropy* offers a valuable tool for advancing retail analytics and consumer behaviour research.

1.3.2 Study 2: A case study of bundle entropy across retail channels using mass transactional data

This study extends the application of *bundle entropy* introduced in *Study 1*. *Bundle entropy* captures the stability of choice sets across multiple transactions, accommodating purchases of varying complexity, from single items to extensive shopping baskets. Using transactional data from a prominent UK grocery retailer with over 3,000 physical stores, the study analyses a sample of 2,181 households who actively shop both online and offline. This dataset includes 228,488 baskets containing over 6.2 million items (45% sold online, 55% offline) over a 19-month period from 2014 to 2016, prior to the COVID-19 pandemic. This methodological approach enables a robust within-subject examination of consumer behaviour across online and offline shopping channels.

The research investigates systematic purchase behaviour at multiple levels, including entire baskets, product categories, and a focus on soft drinks. It also explores the relationship between *bundle entropy* and the healthiness of soft drink choices. Findings reveal that households demonstrate more consistent and predictable online purchase patterns than in-store shopping. Additionally, online purchases are healthier than those made in physical stores.

By expanding the application of *bundle entropy*, this study highlights how systematic buying habits intersect with preferences for shopping channels and health-related product choices. The results provide insights into the predictability of consumer behaviour and its implications for retailers and public health, emphasizing the utility of *bundle entropy* in understanding the complexities of consumer purchase dynamics.

1.3.3 Study 3: The relative power of behavioural, demographic, and psychographic variables as predictors of bundle entropy

This study bridges a critical gap in consumer research by integrating behavioural, demographic, and psychographic variables to predict and understand systematic purchase behaviour (SPB) through the novel measure of *bundle entropy*. While traditional studies often focus on demographic and psychographic factors, the contribution of behavioural data derived from transactional records has been underexplored. This research addresses this gap by combining real-world transactional data with survey-based demographic and psychographic information, providing a comprehensive view of the predictors of *bundle entropy*.

The study leverages data from a prominent UK grocery and pharmacy chain, covering January 2012 to November 2015, with pseudo-anonymized transactional records merged with survey data for a common cohort of 12,137 participants. The merged dataset enables the examination of a wide range of predictors, including shopping frequency, variety-seeking behaviour, age, and personality traits.

To model *bundle entropy*, the study compares traditional statistical regression methods with advanced machine learning approaches, including Random Forest and XGBoost. Using techniques such as SHapley Additive exPlanations (SHAP) and Model Class Reliance (MCR), the research also evaluates the relative importance of behavioural, demographic, and psychographic variables in predicting SPB.

The findings reveal that behavioural variables play a more prominent role than demographic and psychographic factors in predicting *bundle entropy*.

For example, shopping frequency emerges as a critical predictor, highlighting the importance of consistent customer engagement. Meanwhile, variables like age and variety-seeking behaviour underscore the complex interplay of behavioural and psychological traits in shaping SPB.

This study advances the understanding of systematic purchase behaviour by demonstrating the predictive power of behavioural variables and providing a novel framework for integrating diverse data sources to model consumer behaviour effectively.

1.4 Thesis contribution

This thesis offers three global contributions (See Table 7.1 in Section 7 for the complete list of contributions). Firstly, it proposes a novel and parsimonious measurement called *bundle entropy* to assess systematic purchase behaviour across multiple purchase sequences when faced with multiple options. Chapter 3 provides a detailed explanation of this measure, which aims to shed light on the dynamics of buying behaviours. *Bundle entropy* can be used to identify individuals with systematic and unsystematic behaviours, which can be valuable for both academics and practitioners.

For instance, academics can use *bundle entropy* to uncover fresh insights by cross-referencing it against other behaviours, as empirically demonstrated in *Study 2*. Practitioners can leverage *bundle entropy* to identify individuals or households with predictable buying behaviours and offer them appropriate bundle offers, products, and services according to their regular needs, improving customer experience and financial gains. For less predictable individuals/households, retailers might want to target them with suitable innovations, new products, or varied direct offers or recommen-

dations. *Bundle entropy* also contributes to the field of data mining by allowing the efficient and accurate mining of systematic product choices from massive transaction records. This was demonstrated by its publication at the IEEE Big Data 2021 conference (Mansilla et al., 2022).

Furthermore, in addition to the use cases in transactional data sets, bundle entropy could be used in other areas where there is a high degree of data. This can be used in supply chain management, for instance, to forecast demand, such as vehicle need (frequency and regularity) for third-party logistics. Bundle entropy can make fleet scheduling, inventory and route planning easier, by looking at the frequency and patterns within demand data. It can also be used to analyze procurement datasets that reflect activities of buying goods and services and provide a way to see the consistency and variation of suppliers. Such uses show the measure's adaptability and highlight how it can advance predictive analytics and decision-making across industries.

Secondly, this thesis furthers our understanding of within-subject buying behaviours across different channels by applying *bundle entropy* to explore empirically whether systematic purchase behaviour and healthy choices differ online versus offline. Chapter 5 showcases a comprehensive case study that illustrates the efficacy of *bundle entropy* in examining real-world data and identifying consistent and healthy purchasing patterns across different retail channels for an individual or household. Moreover, this thesis explores the wider implications of its findings on consumer welfare and identifies potential avenues for further research in this field, thereby adding to the ongoing conversation on consumer behaviour research. The findings of this study (Chapter 5) have been developed into an article, which is currently undergoing the second stage of review at the Journal of Business Research (Mansilla et al., 2024c).

Finally, modelling and understanding any type of consumer behaviour is a complicated task, as selecting the appropriate variables and methods can be challenging. Through the thesis, particularly *Study 3*, we have a unique opportunity to access a variety of data sources - demographic, psychographic, and behavioural information for each individual/household of a leading UK pharmacy and grocery retailer. This access enables us to explore a wide range of diverse predictors to understand and predict *bundle entropy*. Chapter 5's final study aims to contribute to our understanding of the factors driving *bundle entropy* by utilising a diverse range of predictors. This study also highlights the efficacy of big data analytics, machine learning models, and techniques as innovative approaches to predict and understand complex buying patterns in contrast to traditional statistical methods. These contributions were demonstrated by the acceptance of an academic paper derived from this study at the latest British Academy of Management conference (Mansilla et al., 2024b).

In short, the synergy of diverse datasets and novel approaches from different fields can offer a more comprehensive understanding of buying patterns, which are notoriously changing, making them more challenging to measure, predict, and understand.

1.5 Thesis outline

This section provides an overview of the thesis structure, which is visually represented in Figure 1.1. The thesis consists of three parts. The first part, covered in Studies 1a and 1b (Chapter 3), focuses on developing, comparing, and testing a new direct method to measure systematic purchase behaviours across multiple purchase events in the complex field of purchase

dynamics. The second part, covered in *Study 2* (Chapter 5), examines the accuracy and practical usage of the new measurement across different purchase contexts using real-world datasets. The final part, covered in *Study 3* (Chapter 6), aims to obtain a novel and comprehensive understanding of the primary factors influencing *bundle entropy* by exploring different machine learning models and techniques. Each study includes contextual background information and a brief discussion of the specific findings.

Finally, Chapter 7, the Discussion and Conclusion, reflect on the empirical findings from the different studies and how they contribute to answering the research questions and objectives of the thesis. The chapter closes by acknowledging some of the limitations and future research within consumer buying research.

Overall, the thesis highlights the usefulness of current sources of massive transactional data in understanding the dynamics of buying behaviours, specifically systematic purchase behaviour. However, novel data sources often require new approaches. Therefore, the thesis proposes a new approach to measuring systematic purchase behaviour, outlines the requirements, tests its properties, and compares it against existing measures. The thesis also emphasises the benefits of using machine learning models when investigating complex relationships between predictors derived from different data types, sources and sizes.

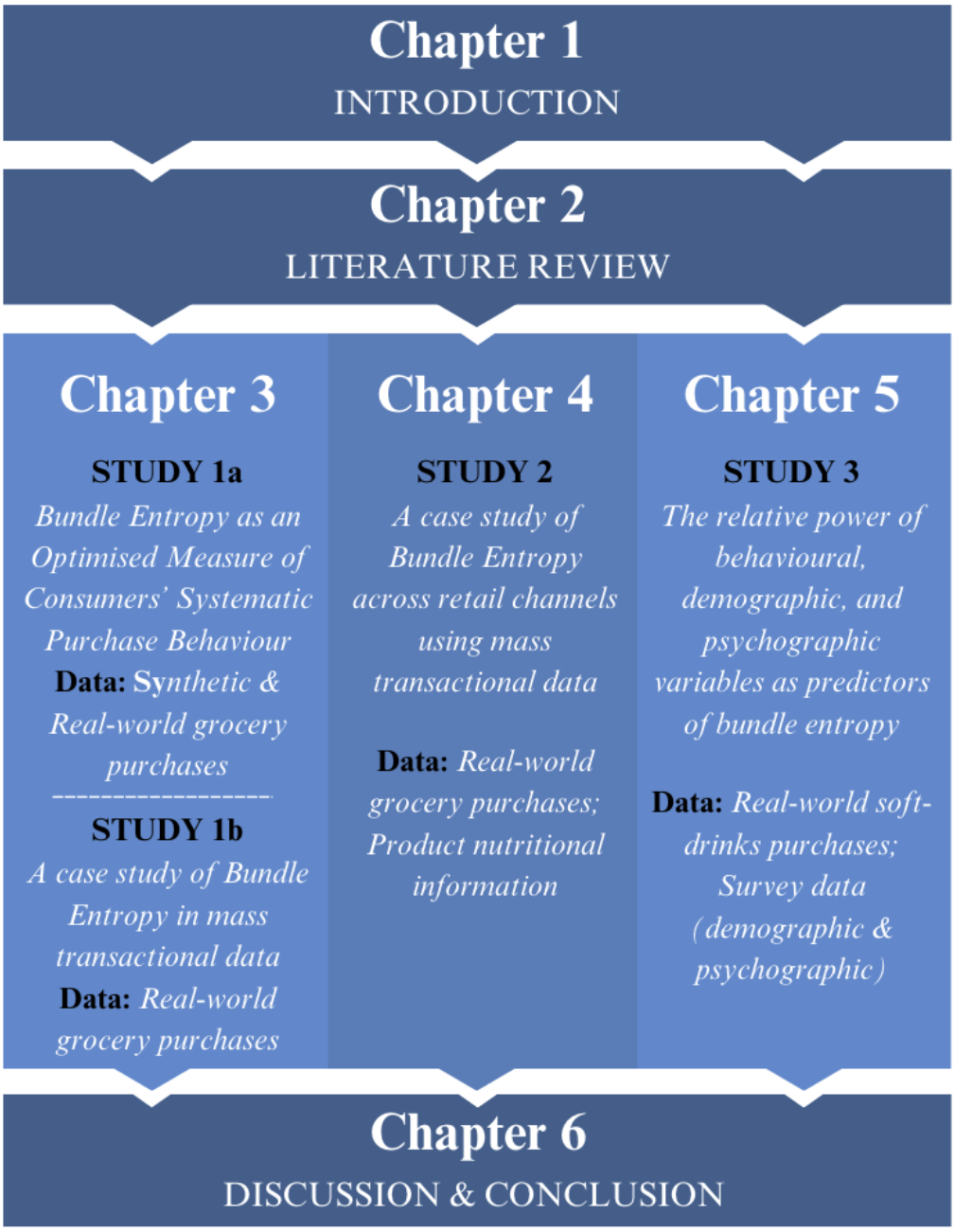


Figure 1.1: Schematic of the thesis structure

Chapter 2

Literature Review

Contents

In this chapter, the study aims to support the theoretical and methodological aspects of the thesis through three sections.

The first section provides a comprehensive overview of the most pertinent literature on the use of transactional data in consumer research. It delves into the past and present use of transactional data and explores its potential future applications. Moreover, the discussion highlights the potential benefits of transactional data in exploring and understanding hidden purchase and behavioural patterns across different retail scenarios. The section also identifies gaps in the literature, providing a basis for the theoretical, methodological, and practical contributions this thesis targets.

The second section of the chapter introduces the most relevant methods that have been applied to assess consumers' purchase predictability. These methods are examined in detail, with particular attention given to their capabilities and limitations. The discussion also highlights how the research in this work can complement existing ones by introducing a novel measure to assess purchase predictability over time.

In the third and final section of the chapter, the study reviews the most relevant literature and discusses different approaches to modelling consumer purchase behaviour outcomes. Additionally, it examines models from various fields that attempt to understand the anatomy of purchase behaviour. The discussion focuses on the variables considered in different models to understand their potential contribution to research. These predictive variables are grouped into two categories: demographic and psychographic. By exploring these various approaches, the study aims to provide a better understanding of the mechanisms that drive systematic purchase behaviour and how they can be leveraged for research and practical purposes.

2.1 Part I: Past and future usage of transactional data in consumer research

In the present times, the retail industry has evolved into a data-driven sector where loyalty programs, both online and offline, generate vast amounts of multi-dimensional data sets. These datasets provide retailers ample opportunities to examine and understand consumer behaviour by identifying concealed buying patterns and generating data-driven insights (Hossain et al., 2020). With unique individual/household identifiers incorporated into the data sets, tracking and monitoring consumer behaviour over extended periods has become feasible. This is made possible due to recent technological advances, particularly in data analytics, that have created new and efficient possibilities in the consumer behaviour field. As a result, retailers can now make better and more informed decisions by identifying and analysing hidden purchase patterns at different levels that were previously challenging to obtain.

The field of consumer buying behaviour has been the subject of extensive research through the lenses of demographic and psychographic variables. However, the use of transactional data in this field has undergone significant changes over time, leading to a transformation in its direction.

In the 1980s, studying consumer behaviour using transactional data was challenging due to the lack of individual identifiers. During this period, researchers focused on examining consumer buying behaviour by analysing demographic and psychographic attributes. There was a limited emphasis on behavioural data, and researchers primarily relied on macro-level sales figures, market trends, or general purchase behaviours.

Despite these limitations, researchers utilised transactional data to discern

broad market behaviours and preferences, laying the foundation for future developments. For instance, studies conducted during this period, such as Ehrenberg (1988) and Bawa et al. (1989), used transactional or shopping data to understand purchase behaviours on a larger scale.

Furthermore, transactional data was also used in the stock market to understand market trends, as seen in studies such as Harris (1986) and Wood et al. (1985). However, there were limited capabilities for studying individual buying behaviour at that time.

The 1990s marked a significant turning point in various fields, including consumer research, thanks to the introduction of advanced computing power and data storage capabilities. This technological revolution paved the way for more personalised analysis, leading to a better understanding of individual purchasing behaviour.

Influential work by Fader and Lodish (1990) explored the examination of individual transactions to understand their purchase preferences across more than 330 product categories. This early exploration of individual purchasing behaviour inspired other studies, such as those by Vilcassim and Chintagunta 1995, Gupta et al. 1996 and Bell and Lattin 1998a, which aimed to uncover further insights into consumer purchase behaviour, preferences, and needs.

To better understand consumer purchase behaviour, researchers began utilising household scanner panel data to develop various models. These models contributed to understanding different phenomena, such as brand choices, brand loyalty, price sensitivity, and repeat purchases (e.g. Fader and Schmittlein 1993, Allenby and Lenk 1994, Siddarth et al. 1995). The shift towards individual-level analysis allowed for a more precise understanding of consumer buying behaviour, which was previously impossible.

This new era of consumer research has enabled companies to provide better customer experiences and create more effective marketing strategies.

During the late 1990s, loyalty programs became increasingly popular among businesses and consumers alike. These programs gave businesses a unique opportunity to gather detailed transactional data on individual consumers (Hart et al., 1999; Smith et al., 2004). This breakthrough enabled retailers to delve deeper into complex purchase patterns and formulate tailored marketing strategies that could better target specific consumer groups (Uncles, 1994). The loyalty programs quickly became an invaluable tool for researchers, offering a wealth of transactional data that allowed for unprecedented insights into consumer behaviour at the individual level (Smith et al., 2004).

As technological advancements took place, retailers began accumulating transactional data on a larger scale linked to individual consumers, which allowed for a more sophisticated analysis of consumer behaviour (Bounsofiane, 1996) but at the same time, the start of the discussion of privacy issues (Evans, 1999; Long et al., 1999; Smith and Sparks, 2004). However, despite these advancements, the focus remained on individual transactions, and there was still much potential to explore insights that could be derived from the sequence of purchases.

Advancements in predictive analytics and machine learning propelled the field even further. Boone and Roehm (2002) demonstrated how algorithms like artificial neural networks could segment individuals based on transactional data analysis. By integrating cutting-edge technologies, we have gained a deeper understanding of consumer behaviour at an unprecedented level of granularity.

As the 21st century began, e-commerce platforms and technological ad-

vancements drove a new era for transactional data analysis. The integration of online and offline sources provided a holistic view of consumer behaviour across various channels, marking a crucial shift from transactional data being a mere recording of individual transactions to a tool for understanding temporal patterns and sequences of purchases (Smith et al., 2004).

Furthermore, the evolution of transactional data analysis has shifted towards understanding what products individuals buy and how those products are associated. The application of basket analysis, as exemplified by Julander (1992), Russell and Petersen (2000a) and Mild and Reutterer (2003), showcases the exploration of associations between products bought together across multiple product categories (cross-category purchase behaviour). This evolution allowed retailers to gain a more nuanced understanding of consumer preferences, which aided in improving the design of targeted marketing campaigns and personalised recommendations.

Transactional data, enriched with temporal information, also allows for exploring time-based purchase patterns. Analysing cross-selling effects or when certain products are frequently purchased or identifying seasonal variations in buying behaviour allows retailers to optimise, for example, inventory management and marketing strategies (Wong et al., 2005). Moreover, temporal buying patterns can reveal periods of stability and uncertainty that retailers can leverage to optimise promotional efforts (Smith, 2019). Furthermore, leveraging individual-level transactional data empowers retailers to implement personalised marketing strategies customised to each customer's historical purchase behaviour, such as tailored promotions, discounts, and product recommendations (Verhoef et al., 2016; Pathak et al., 2017).

Modern technology has revolutionised the way researchers use and analyse

transactional data. By incorporating innovative technologies, researchers can now gain a more comprehensive understanding of their customer's shopping behaviour, surpassing conventional market-level analysis (Li et al., 2019; Arasu et al., 2020). For instance, machine learning and data mining techniques can be used to analyse past purchase records and make predictions regarding customer churn (Khodabandehlou and Zivari Rahman, 2017; Ascarza et al., 2016), product choices and interactions (van Wezel and Potharst, 2007; Behe et al., 2020; Chen et al., 2021; Paolanti et al., 2020), promotion response (Shin and Cho, 2006), repeat purchase (Schwartz et al., 2014), store visits (Droomer and Bekker, 2020; Bian et al., 2023), and general buying preferences and intentions (Arasu et al., 2020; Martínez et al., 2020).

Furthermore, utilising transactional records and new technologies can facilitate a comprehensive analysis of various social aspects of consumer research. This includes identifying potential issues such as nutrient deficiencies (Mansilla et al., 2024a), food insecurity (Nica-Avram et al., 2021), and excessive use of grocery plastic bags (Lavelle-Hill et al., 2020). Researchers such as Ma and Sun (2020), Duarte et al. (2022), and Ngai and Wu (2022) have conducted extensive reviews on the application of artificial intelligence, data mining, and machine learning in the field of consumer research, shedding light on what we can expect in the future.

2.1.1 Knowledge gap

For a long time, transactional data has been overlooked in many fields, such as consumer research, due to limitations in collecting, processing and analysing data. However, with recent technological advancements and improvements in data analytics and related fields, transactional data has be-

come a great source of information. Despite this, there is still a significant gap in the literature as the full potential of transactional data in understanding consumer buying behaviours, particularly when combined with demographic and psychographic characteristics, has yet to be thoroughly explored.

One of the primary aims of this thesis is to address this gap by conducting three interconnected studies, with each study building upon the findings of the previous one. In addition to their specific objectives (provided later in this chapter), these studies aim to highlight the significant potential of big transactional data in enhancing our understanding of consumer purchasing behaviour, particularly Systematic purchase behaviour (SPB). Throughout this thesis, the term SPB denotes the regularity of product choices across multiple purchases over a period of time, as defined by (Guidotti et al., 2015). For a more in-depth exploration of this concept, please refer to section 1.2.

2.2 Part II: Basket predictability and heterogeneity

As previously mentioned, SPB refers to the consistency of product selections across a series of shopping trips. According to Guidotti et al. (2015), due to the stochastic nature of SPB, each product has a certain probability of being selected in any given purchase. This results in the entire basket or sub-baskets having associated probabilities. Understanding these probabilities is crucial for comprehending SPB.

However, despite the work of Guidotti et al. (2015), there is limited research

directly quantifying the predictability of consumer purchases from transactional historical records to identify individuals for actionable interventions. Most existing studies have focused on explaining product choice, often in a sequential manner, by fitting latent models (e.g. Ehrenberg (1988); Goodhardt et al. (1984); Fader and Schmittlein (1993); Russell and Petersen (2000b); Sharp et al. (2012)) or building predictive models to predict subsequent visit behaviour (e.g. Kim et al. (2003); Van Den Poel and Buckinx (2005); Lo et al. (2016)).

To comprehensively examine SPB, this section will delve into the most significant research on product and basket predictability alongside related purchasing behaviours, such as heterogeneity and variety. It begins by exploring previous research that aimed to explain product choices through latent models and predictive models for anticipating future visit behaviour, as these studies laid the foundation for current knowledge. This review aims to establish a groundwork for the contributions of this work, thereby enhancing the current understanding of systematic behaviours.

2.2.1 Related work

In the past, numerous researchers have conducted extensive studies to gain a deeper understanding of the factors that influence consumer product choices. These studies have often employed latent or explanatory models (e.g. Ehrenberg 1988, Allenby and Lenk 1994), aiming to represent observed behaviours by unobserved causal factors. The approach tries to reveal the anatomy of consumers' motivations toward a specific behaviour by using potential causal factors, which enables the exploration of a small number of hypothesised concepts of interest that drive or cause the observed behaviour. These factors provide a single value representation of

each concept of interest. Assuming the existence of multiple factors, these act together to explain the observed purchase behaviour under exploration.

One of the most notable studies within this approach is the Negative Binomial Distribution Dirichlet model on repeat buying by Ehrenberg (1988), which has inspired similar research (Frisbie, 1980; Goodhardt et al., 1984; Fader and Schmittlein, 1993; Uncles et al., 1995; Uncles and Hammond, 1995; Russell and Kamakura, 1997; Bhattacharya, 1997; Sharp and Sharp, 1997; Russell and Petersen, 2000b; Sharp et al., 2012). The Dirichlet model is designed to examine aggregate behaviour more than individual/household behaviour. The success of the model in predicting behaviour is hard to debate under certain conditions. However, there are several assumptions, problems, and acknowledged shortcomings, such as the assumption of stationary conditions (accepted by Ehrenberg himself and others - for example, Bhattacharya 1997). The model uses variables that it derives from the purchase data, so it is internally valid and has the advantage of not having to require exogenous explanatory variables. But this is also its weakness - it's a 'you put in what you get out model'.

Another model that attempts to represent consumers' varied behaviour is by McAlister and Pessemier (1982), who explored consumers' levels of satiation, distinction, interpersonal stimulation, and intrapersonal information at a sequence of purchase occasions. However, most input values (parameters) for each factor are either arbitrary or constant values estimated based on different theoretical frameworks and assumptions.

Still, other researchers (e.g. Seetharaman and Chintagunta 1998) have centred on representing consumers' purchase dynamics behaviour by nesting and evaluating different factors like marketing variables, heterogeneity across individual preferences, variety-seeking, and consistent purchase

(Givon, 1984). However, like the models mentioned above, some parameters have to be estimated under various specifications, leading to potential aggregation bias.

More recently, researchers have been using predictive models to understand better subsequent visit behaviour (e.g. Kim et al., 2003, Van Den Poel and Buckinx, 2005, Lo et al., 2016). Rather than exploring relationships between observed variables, these models aim to predict or forecast consumer behaviour with a focus on accuracy and performance.

However, there has been a lack of research that directly quantifies the predictability and consistency of human purchases from historical data. This is an important gap in the literature that needs to be addressed, as it can help businesses identify those individuals who are likely to make certain purchases. It can also provide insights into customers who regularly purchase the same bundle of products and those who do not. This information is especially valuable for large chains with a diverse range of products, allowing them to take actionable steps to influence consumer behaviour and boost sales.

In an attempt to measure the variety of products people buy, early researchers counted the number of distinct products that individuals purchased (Kahn and Lehmann, 1991). Still, this method was found to have limitations and did not accurately capture the complexity of human purchasing behaviour. A more accurate measure of purchasing behaviour was introduced through the concept of entropy, which measures the variety of products within a single group. This measure considers the predictability of behaviour through uncertainty and is more comprehensive than the earlier method (Akaika, 1985; Smith et al., 2014). Entropy has also been used as a proxy for diversity and variety in many other fields, such as psychology

(Stamps, 2002), economy (Straathof, 2007), and ecology (Jost, 2006).

Other methods were proposed, such as the Hirschman-Herfindahl (Nauenberg et al., 1997) and the Gini (Dorfman, 1979) coefficient. Still, entropy measures were found to be a more actionable link in this context, as they are directly linked with prediction (Straathof, 2007). Moreover, the entropy measures can encode desirable aspects related to the distribution, rareness, and commonness of the products contained in the group. As noted by Straathof (2007), entropy measures' utility is higher than other measures as they can provide a more complete understanding of human purchasing behaviour.

From the viewpoint of a single group, entropy is a measure that quantifies the uncertainty or difficulty one faces when trying to predict a single item from a specific group of items. Under this point of view, Nicolas-Sans and Ibáñez (2021) utilised the concept of entropy and proposed a measure to check the weighted variety of products that an individual buys across multiple baskets during a certain period of time. In his research, Nicolas-Sans and Ibáñez (2021) distinguishes his measurement from simply counting the distinct items across all baskets by considering the proportion of times the items were purchased. To illustrate his point better, imagine a hypothetical store that sells three different items: milk (m), cheese (c), and salmon (s). Two customers visit the store four times each, and their sequences of purchases are as follows:

$$\textit{Customer 1} = \{m, c, s\}, \{m, c, s\}, \{m, c, s\}, \{m, c, s\}$$
$$\textit{Customer 2} = \{m, m, s\}, \{m, m, c\}, \{m, m, m\}, \{m, m, m\}$$

If we count the unique items purchased by each customer from all the baskets, it is evident that both customers have bought three unique items.

This simple approach suggests that both customers buy the same range of products. However, Nicolas-Sans and Ibáñez (2021) argues that this is not entirely accurate in terms of diversity. *Customer 1* shows a more varied buying behaviour with purchases including 4 milk, 4 cheese, and 4 salmon, whereas *Customer 2* shows a less diverse product selection, predominantly purchasing 10 milk, 1 cheese, and 1 salmon. Nicolas-Sans and Ibáñez (2021) suggests that *Customer 1* exhibits a more diverse purchasing behaviour than *Customer 2*. This can be seen in Table 2.1, which displays the percentage of each item purchased in all baskets, calculated by dividing the quantity of each item by the total number of items bought.

	milk	**Cheese**	**Salmon**
Customer 1	(4/12) 33.3%	(4/12) 33.3%	(4/12) 33.3%
Customer 2	(10/12) 83.3%	(1/12) 8.3%	(1/12) 8.3%

Table 2.1: Percentage of each item purchased across all baskets relative to the total number of items bought.

It is crucial to recognise that when analysing multiple baskets of purchased items as a single group, entropy can be utilised to gauge the variety of the baskets (Nicolas-Sans and Ibáñez, 2021). This insight can prove beneficial in ascertaining the breadth of products that individuals buy from the entire selection. However, it's important to highlight that under this lens, entropy does not measure the predictability of baskets or bundles of items. This particular aspect is the central focus of study 1 in Chapter 4 of this thesis.

To understand this difference better, let's consider another example. Suppose an individual visits a store three times and buys milk (m), cheese (c), and salmon (s) each time. In this case, a shopping basket is understood as a set of items a customer purchases during their visit to the store (Boztuğ and Reutterer, 2008). Thus, the purchases result in a sequence of baskets containing the same three products $[\{m, c, s\}, \{s, c, m\}, \{c, s, m\}]$.

In this scenario, the order within the baskets does not alter the fact that the individual bought the same three products on each visit. However, predicting the next random item becomes almost impossible if we consider these purchases as a single unordered group ($[m, c, s, s, c, m, c, s, m]$). Similarly, predicting the following item is equally challenging if the products are considered a single ordered group ($[m, m, m, c, c, c, s, s, s]$). In such a scenario, entropy reflects this challenge by reporting maximum entropy.

However, if these purchases are considered at the basket level, it can be seen that the same basket is always purchased (m, c, s), and hence, it's 100% predictable, regardless of the order within the basket. Thus, the measure should have reported zero uncertainty. This simple example illustrates the importance of determining the level at which entropy is to be applied to obtain the correct insights.

The previous motivated Guidotti et al. (2015) to develop one of the first attempts to measure the unpredictability of an individual's basket composition. For this, they created a measure called Basket Revealed Entropy (BRE). Using real-world grocery transactions, the measure uses frequent patterns mined from customers' baskets to calculate entropy based on common sub-baskets. However, the direct application of entropy in this context is inappropriate since the definition of predictability plays a crucial role in formulating and taking subsequent actions in a business context.

BRE considers predictability as correctly predicting sub-baskets, ignoring any additional context. In contrast, Basket Level Entropy (BLE) defines predictability as the task of predicting an individual's entire basket composition. This is equivalent to the joint entropy of baskets, with all items represented by binary indicator variables indicating whether each item was present in the basket. In contrast to the previous application of entropy,

which was applied at the item level and resulted in the loss of item attribution to baskets, in this case, a symbol represents an item, and the prediction task is focused on predicting which item will be chosen to add to the basket at any given time.

It is crucial to consider the usefulness of different measures of "predictability" in their respective contexts. This thesis aims to understand customers' systematic purchasing behaviour in the fast-moving consumer goods sector. Although the discussion pertains primarily to this sector, the findings may have broader applicability.

2.2.2 Knowledge gap and rationale for Study 1 & 2

This section discussed and delved into an overview of the existing techniques that have been utilised to evaluate the predictability and heterogeneity of consumer behaviour when making purchases. The primary objective of this section was to explore the previous methods and understand their approaches to provide a comprehensive understanding of the requirements needed to develop a new measure that could better capture systematic purchase behaviour. A more detailed understanding of the most relevant measure is provided in *Study 1a* (Chapter 4) of this thesis.

The traditional methods of assessing predictability, heterogeneity, and related concepts, such as repeat purchases, have relied on basic metrics and statistical analyses. However, these methods may not capture the complex patterns and relationships in transactional data. As a result, more sophisticated techniques are needed to account for the dynamic nature of consumer buying behaviour and variations across multiple purchases and choices.

To address this gap, *Study 1* of this thesis proposes a novel and parsimonious measure that utilises the concept of entropy to quantify systematic choices across historical purchases. This new measure aims to capture the systematic purchase behaviour that has not been accurately captured by traditional and recent methods. This measure will be subjected to empirical validation and comparative analysis to demonstrate its effectiveness in enhancing basket predictability and understanding systematic choices over time.

Furthermore, *Study 2* of this thesis will empirically test the effectiveness of the proposed measure on big real-world data across different retail channels to explore the measure's usability in different purchase settings for actionable future interventions.

2.3 Part III: Predictors of buying behaviour outcomes

Examining consumer behaviour has always played a crucial role in marketing research. Demographic and psychographic variables have traditionally been used to comprehend different consumer buying outcomes (Bellman et al., 1999; Baumeister, 2002; Islam et al., 2017). While demographic factors such as *age, income*, and *social class* have provided valuable insights, they are known to be incomplete predictors not only in offline settings (Rich and Jain, 1968; Bellman et al., 1999) but in the online as well (Li and Russell, 1999). On the other hand, psychographics, which analyses deeper psychological traits and attitudes, offer additional layers of understanding but exhibit varying degrees of effectiveness in predicting purchase outcomes (Van Trijp et al., 1996).

Despite extensive research, there is still an ongoing debate regarding the relative power of demographics versus psychographics, with studies showing mixed results depending on the context and industry (Van Trijp et al., 1996). Furthermore, there is limited research done on exploring behavioural variables derived from historical transactional records in modelling purchase behaviours.

2.3.1 Demographics predictors

Retailers have always been interested in understanding the profile of individuals who frequently buy to plan and target the right audience. Researchers have approached this from various perspectives, but all share a common goal of comprehending the underlying demographics of different purchasing behaviour outcomes.

Demographic variables, such as *age, income, gender*, and *education*, are critical determinants in the modelling of consumer buying behaviour. Many studies suggest that *age* plays a significant role in various buying behaviours and contexts, although its impact is not universal. Early research by Rich and Jain (1968) used *age* as a metric for the life cycle, which showed that younger women shop more frequently than older women. Similarly, Henry (2002) found that younger individuals prioritise non-functional purchases more than older individuals. Research focusing on *age* has evolved over time in the online shopping landscape. Studies from the late 1990s suggested that *age* had either a weak or insignificant association with online purchasing behaviour (e.g. Li and Russell 1999, Bellman et al., 1999).

However, as online shopping experiences became more mature, subsequent research indicated that *age* indeed serves as a significant predictor of online

buying behaviour. For instance, Wood (2002) demonstrated that younger consumers show greater interest in adopting new technologies throughout their purchasing decisions (Sorce et al., 2006). Moreover, younger consumers tend to embrace enjoyment and impulsivity during the online shopping process (Kanwal et al., 2022), whereas older consumers lean towards risk avoidance and adherence to traditional purchasing patterns (Lian and Yen, 2014).

Extensive research has been conducted on the impacts of *income* and related factors, such as *social class*, on purchasing behaviour. According to research conducted by Slocum and Mathews (1970), Myers et al. (1971), and Peters (1970), *income* has a more significant impact on buying behaviour than *social class*. Nonetheless, some scholars have put forth the idea that the significance of *income* versus social status may differ depending on the type of product being considered. This seems to hold especially true for groceries, as highlighted by Schaninger (1981). More recent research has found a positive relationship between individuals' *income* and their attitude towards loyalty (Klopotan et al., 2016) as well as with their tendency towards less varied choices (Carlson et al., 2015), suggesting systematic behaviours as *income* increases.

Numerous studies have explored the impact of *gender* on buying behaviour. While some studies, such as those conducted by Li and Russell (1999) and Kim and Forsythe (2008), have not found significant differences in buying behaviour between genders, many others have identified clear *gender* disparities in a range of contexts. For example, research has revealed that women tend to opt for food with fewer calories (Skatova et al., 2019), are more inclined to make impulsive purchases, particularly in the fashion and online shopping domains (Brunelle and Grossman, 2022), and make purchases more frequently than their male counterparts (Verplanken, 2006).

Kanwal et al. (2022) thorough analysis of gender-based disparities in consumer buying patterns revealed that distinctions between genders are more pronounced than similarities.

There is an ongoing discussion regarding the impact of *household composition* on consumer purchasing behaviour, with a particular focus on brand loyalty and repurchasing habits. Mittal and Kamakura (2001) conducted a study on automobiles and found a positive correlation between *household size* and repurchase rates. However, other empirical studies, such as those by Ailawadi et al. (2008) and Koschate-Fischer et al. (2014), have produced varying results, with no significant relationship found between household size and loyalty across different grocery data sets. In a recent study by Koll and Plank (2022) on grocery buying, it was concluded that the correlation between *household size* and repurchasing was not significant. These findings suggest that the impact of *household size* may be specific to the product's level of involvement. While *household size* may play a significant role in high-involvement products, it may not have the same impact on low-involvement purchases.

Researchers have also explored the connection between *education* levels and buying behaviour, with results evolving over time. Early studies, such as Bellman et al. (1999) and Li and Russell (1999), indicated that higher *education* was positively correlated with online purchasing. This was attributed to the idea that those with higher *education* were more digitally literate and comfortable with e-commerce transactions, making them more likely to shop online.

In addition, Wood (1998) suggested a link between higher *education* and impulsive buying decisions. However, recent research has challenged these findings as online shopping has become more accessible and mainstream.

Contrary to previous studies, Nayyar and Gupta (2011) found no significant association between *education* and online purchases, questioning the assumption that higher *education* leads to higher online shopping engagement. Furthermore, Rana and Tirthani (2012) examined the demographics influencing impulse buying in Indian consumers, including *education*. Their findings showed a negative correlation between *education* and impulse buying behaviour. These results are supported by Ghafoor et al. (2015), who suggest that higher educational attainment may actually reduce impulsive tendencies rather than increase them.

2.3.2 Psychographic predictors

Recent research indicates that a person's unique personality traits can impact consumer behaviour. Early studies revealed that extrinsic factors like *sales, promotions, out-of-stock* situations (Holbrook, 1984), or different product displays (Deng et al., 2016) influenced individuals seeking variety in their purchases (Van Trijp et al., 1996). However, intrinsic motivations, such as *curiosity, satiation*, or *uncertainty* regarding future choices (Simonson, 1990), were also associated with those seeking variety in their purchase decisions (Van Trijp et al., 1996). Additionally, research has focused on the correlation between different levels of *self-control* and buying behaviour, particularly impulsive buying, which can directly impact product choices (Mulyono and Rusdarti, 2020).

Moreover, individuals tend to lose *self-control* towards the end of the day, increasing the likelihood of impulsive purchases (Baumeister, 2002). Sharma et al. (2010b) study also corroborates that individuals with high *self-control* are less prone to impulsive buying. However, they may still seek variety, highlighting the impact of *self-control* on systematic product

choices, whether through impulsive or planned decisions.

Researchers have also argued that attitudes towards certain states can also affect decision and buying behaviours. For instance, a study conducted by Sorce et al. (2006) suggests that our attitudes can significantly influence our buying behaviour, particularly in online settings. The study also highlights how our attitudes affect our purchasing decisions and influence how we evaluate and perceive products.

Another study by Garg et al. (2007) found that individuals who experience positive emotional states, like *happiness*, tend to be more mindful of their food choices. This implies that happy individuals are less likely to make impulsive purchases since they are more likely to consider the emotional consequences of their food choices and avoid items that may lead to regret later on. This finding is supported by Lin and Lin (2009), who indicates that individuals who experience negative emotional states, such as *sadness*, are more likely to purchase a greater variety of snacks compared to those who are happy.

Other research studies have delved into the connection between consumer purchasing behaviour and personality traits (Brunelle and Grossman, 2022), with a focus on the Big Five personality model (Costa and McCrae, 1992) or derived frameworks (e.g. 3M hierarchical mode (Mowen, 2000)). This model assesses traits such as *neuroticism, extroversion, openness, agreeableness,* and *conscientiousness*. The findings from these studies reveal that certain personality traits can positively or negatively affect different systematic behaviours, such as impulsive buying, brand loyalty, and repeat purchases.

For instance, Mowen (2000) found that *agreeableness* and *neuroticism* are positively associated with impulsiveness in purchasing behaviour. Consis-

tent with Mowen (2000), Pirog and Roberts (2007) found that *neuroticism* was linked to impulsivity among a group of 254 students. On the other hand, He also found that *extroversion* shows a negative association with impulsiveness. Interestingly, he did not find a significant correlation between *agreeableness* and compulsive buying, as in previous studies.

In addition, further research has shown that *conscientiousness* can also impact impulsive behaviour. A study by Verplanken and Herabadi (2001) reported a negative correlation between *conscientiousness* and impulsive buying. Similarly, in the context of online shopping, personality traits such as *neuroticism* (Wang and Yang, 2008) and *conscientiousness* (Sun and Wu, 2014) were negatively related to impulsive buying.

Moreover, a recent study discovered that a buyer's personality matching that of a seller in an offline context might either enhance or reduce impulsivity in purchasing behaviour (Ali et al., 2022). For example, if a buyer and seller share similar levels of *agreeableness* and *openness*, it can lead to more compulsive buying. Conversely, if there is a similarity in *neuroticism*, it can lead to less compulsive buying.

Furthermore, gender also plays a crucial role in impulsive buying, with women being more susceptible to exhibiting compulsive buying behaviour driven by *neuroticism, extroversion*, and *openness* to experience than men (Tarka et al., 2022). Additionally, shopping mission also influences impulsiveness, with *conscientiousness* and *agreeableness* having a direct and adverse relationship with impulsive buying when shopping for pleasure, entertainment, or emotional gratification. This relationship was found to be stronger for women than for men (Tarka et al., 2022).

Individuals' loyalty towards products, brands, and stores can be considered a reflection of their consistent choice of a particular thing or systematic

purchase behaviour. This area of study in consumer behaviour is of great importance. Research on the relationship between personality traits and store loyalty is still relatively limited. Existing studies have so far provided only weak evidence to suggest any significant correlation between the two.(Bove and Mitzifiris, 2007). However, the correlation between personality traits and brand or product loyalty has received more attention. Early research suggests a positive connection between the two. Guo (2003) found that all five dimensions of personality (*agreeableness, openness, conscientiousness, extroversion*, and *neuroticism*) were significantly positively related to brand personality. However, other scholars, such as Chow et al. (2004), have only found significant positive relationships with *extroversion* and *openness* among their participants.

Studies have also focused on the hedonic aspect of products that can significantly impact a consumer's loyalty towards products that offer hedonic value. Research indicates that certain personality traits can influence an individual's loyalty towards such products. For example, individuals who possess traits like *openness* and *extroversion* are more likely to develop brand loyalty for hedonic products (Matzler et al., 2006). These types of products, which include gourmet ice creams, premium wines, artisan cheeses, and indulgent chocolates, are commonly available in grocery stores with a wide range. An extroverted person may develop a solid attachment to a specific brand of ice cream that offers innovative flavours and vibrant packaging, as it aligns with their desire for novelty and enjoyment.

Moreover, Lin (2010) conducted a study that confirmed previous research and added a new perspective to the topic. The study found that friendly, cooperative, and empathetic consumers are more likely to associate positive emotions with specific products while shopping. A shopper with *agreeable* personality traits may develop brand or product loyalty towards a line of

premium coffees known for their ethically sourced beans and sustainable practices, such as fairtrade certification, as it aligns with their values of empathy and cooperation within global supply chains. This suggests that individuals with *agreeable* personalities will tend to view an exchange between a firm and themselves as honest, decent, and trustworthy. Therefore, speciality coffees, artisan chocolates, and organic wines are just a few examples of hedonic items that satisfy consumers' desires for pleasurable and enjoyable grocery store experiences.

A recent study by Di Crosta et al. (2021) suggests that personality traits may have a more significant influence on consumer behaviour during times of uncertainty, such as the ongoing COVID-19 pandemic, than during periods of certainty. The study found that individuals with high levels of *openness*, who are more open to change, diversity, and new experiences, tend to purchase hedonic products related to new hobbies. This may indicate that people sought out these products to satisfy their desire for exploration and variety, given the limitations on travel and socialising. On the other hand, *conscientious* individuals who prioritise practicality and functionality were less likely to purchase hedonic products. This illustrates how personality traits shape consumer behaviour and how people cope with stress and uncertainty in different ways (Al Hamli and Sobaih, 2023).

There is also evidence that personality traits affect consumers' loyalty behaviour in the online setting. Bosnjak et al. (2007) conducted a study on 808 internet users and found that emotional factors play a stronger role than rational factors in the decision to purchase online. Additionally, three of the five personality traits — *neuroticism, openness*, and *agreeableness* — were found to have a small but significant impact on the intention to make online purchases.

In a subsequent study by Islam et al. (2017) involving college students, it was discovered that extroverted individuals had a positive and significant connection with consumer purchasing engagement. *Neuroticism*, *openness*, and *agreeableness* also had a noticeable impact on online purchasing behaviour. In contrast, the study found that *conscientiousness* had a negative association with online engagement towards purchasing.

Numerous research studies have also explored the impact of personality traits on consumer loyalty across various industries, including mobile services and tourism. For instance, a study by Jani and Han (2014) analysed 529 frequent guests at 5-star hotels and found that individuals with *extroverted*, *agreeable*, and *neurotic* personality traits have a significant relationship with satisfaction and ultimately with overall loyalty towards the hotel. Similarly, Smith (2020) found that customers in the mobile industry who exhibited *agreeable*, *neurotic*, and *open* personality traits were more content with mobile services than other personality types.

2.3.3 Knowledge gap and rationale for Study 3

Overall, the previous section reviewed extant literature that examines a wide range of predictors of different buying behaviours, focusing on both demographic and psychographic variables. While demographic factors like *age, income, gender, education*, and *household size* have traditionally been used to understand consumer behaviour, their effectiveness as predictors varies across contexts and industries. Psychographic variables, which delve into deeper psychological traits and attitudes, offer additional insights but also exhibit varying degrees of effectiveness. The literature highlights an ongoing debate regarding the relative power of demographics versus psychographics in predicting purchase outcomes, with mixed results depending

on the context, channels, categories, and even products.

The gap in the literature lies in the lack of exploration of behavioural variables derived from historical transactional records in modelling purchase behaviours, specifically buying behaviours related to repeat product choices.

The last study of this thesis aims to bridge this gap by proposing a comprehensive model that integrates transactional, demographic and psychographic variables to predict and improve the understanding of systematic behaviours across multiple purchases (assessed by *bundle entropy*). This is achieved by leveraging traditional statistical and machine learning methods (please refer to section 6.1.1 of *Study 3* for a thorough justification of the method selection) to accurately predict *bundle entropy*. Additionally, the study aims to uncover hidden patterns and pertinent correlations within the predictor variables through the use of novel modelling techniques (SHapley Additive exPlanations values (Lundberg et al., 2017) and Model Class Reliance (Fisher et al., 2019; Smith et al., 2020)).

Therefore, through empirical validation and comparative analysis, *Study 3* aims to demonstrate the efficacy of these methodologies in predicting *bundle entropy* and enriching the comprehension of systematic choices.

Chapter 3

Methodology

Contents

3.1 Research philosophy

Various research philosophies can be applied to consumer behaviour research depending on the nature and approach of the study. For many years, positivism and interpretivism have been the most common philosophies in consumer research (Belk, 1986). Positivism assumes that knowledge is derived from observable phenomena and empirical evidence (Kuhn, 1997). On the other hand, interpretivism is based on subjective interpretations of experiences (Mackenzie and Knipe, 2006). Post-positivism emerged as an approach that focuses on the objectivity of the phenomenon and gives equal significance to the experiential and meaningful aspects that underpin it. This later philosophy acknowledges the complex, social, and often unpredictable nature of consumer behaviour (Venkatesh, 1992).

With the emergence of big data and data analytics, scientific inquiry is being reshaped, offering new avenues for data generation, collection, processing and analysis. Data-driven approaches, which blend abduction, induction, and deduction, are gaining dominance over time due to their ability to harness the potential of vast datasets (Kitchin, 2014). Social science research can benefit from the vast array of rich data sources available (Ruppert, 2013). However, the epistemological implications of big data are still a topic of debate (Kitchin, 2014).

As mentioned before, this thesis comprises three interconnected studies that collectively contribute to the overall research. The first study aims to develop a new method to measure individuals' systematic purchase behaviours across multiple purchases. The second study evaluates the proposed measure's accuracy and usefulness in uncovering hidden relationships with other observable behaviours. The final study examines the impact of different predictors derived from demographic, psychographic, and be-

havioural datasets on systematic purchase behaviour (via *bundle entropy*). Although advanced data analysis techniques and big data are utilised in this research project, a positivist philosophy is ultimately adopted. This decision is based on the current discussions surrounding a new research philosophy that is driven by big data, which may become clearer in the near future. By utilising a quantitative methodology approach, positivism provides a robust framework for exploring the topic at hand (Hair, 2009).

Furthermore, the adoption of this philosophy is significant for the research's motivation, intention, and expectations (Mackenzie and Knipe, 2006). It allows for an objective investigation of systematic purchase behaviour and the development of novel scientific insights. Positivism is rooted in the realist ontology assumption, which states that an objective reality exists independently of human perception. Furthermore, this reality can be understood through an epistemological position of systematic observation and empirical analysis (Shelby, 1991). In the context of this thesis and the studies therein, positivism serves as a suitable philosophical principle for the rigorous exploration of consumer buying behaviour patterns and the identification of underlying regularities in purchasing decisions.

The positivist approach emphasises empirical evidence as the foundation of scientific inquiry. This work seeks to objectively uncover patterns and trends in consumer behaviour by analysing real-world, large-scale transactional datasets. The aim is to develop reliable and valid insights into the drivers and dynamics of systematic purchase behaviour (via *bundle entropy*).

Positivism prioritises objectivity and minimises bias and subjectivity in research (Popper, 2005). To achieve this, *Study 1* develops and proposes a direct measure to assess the behaviour under study directly from past pur-

chases. Additionally, this work employs quantitative methodologies such as statistical analysis, machine learning models, and data mining techniques. These quantitative approaches enable this work to objectively analyse large volumes of data, identify meaningful patterns and complex relationships, and draw evidence-based conclusions. By adopting a quantitative approach to big data sets, this work can ensure the reliability and validity of the findings, thus enhancing the credibility and generalizability of the research outcomes.

Positivism is known for its emphasis on systematic observation and rigorous research methods (Bryman, 2016). In this study, novel standards protocols and methodologies for data analysis have been applied (e.g. Cross-industry standard process for data mining (Shearer, 2000)). By adhering to rigorous research standards, the study aims to minimise the potential for bias and error and produce robust and replaceable findings in all three studies. Through this, the study seeks to contribute to the cumulative body of knowledge in the field of consumer behaviour research.

Moreover, positivism encourages the use of theory, hypothesis testing or research questions to guide research inquiry (Kuhn, 1997). This work draws upon existing theoretical frameworks and empirical research findings to formulate research questions about the relationships between systematic purchase behaviour and other measurable buying behaviours. These research questions serve as testable propositions that can be evaluated using empirical data, allowing this thesis to assess the validity of our theoretical assumptions and refine our understanding of consumer behaviour dynamics in different contexts (e.g., online vs in-store).

While positivism provides a solid foundation for empirical research, it is not without its limitations. Positivism tends to prioritise quantitative data

and may overlook the subjective or qualitative aspects of human experience, especially in grocery shopping where multiple external (e.g., offers, discounts, etc.) and internal factors (e.g., psychographics) can play a huge role. Additionally, positivist research can be constrained by data availability and quality and statistical techniques' limitations. To address these limitations, the thesis worked with several large real-world transactional datasets with national coverage from different retailers to maximise the reliability of the findings as much as possible. Additionally, when modelling *bundle entropy* in *Study 3*, a complementary approach is adopted to integrate not only behavioural data from past purchases but also psychological traits and attitudes, allowing for a broader understanding of *bundle entropy* and ultimately systematic purchase behaviour.

While the research follows a positivist approach, it's important to acknowledge that the psychographic data used in *Study 3* is derived from a survey based on the Big Five personality framework. Although the survey was initially developed for a different study (Lavelle-Hill et al., 2020), it provides valuable psychological insights into the same customer cohort for whom we have transactional data. The Big Five framework is widely recognised and validated within the field of psychology, providing a robust foundation for understanding personality traits (John and Srivastava, 1999; McCrae and Costa, 2004).

Even though positivism typically emphasises observable and quantifiable data, integrating psychographic information, grounded on an established psychological theory, into the transactional data enables a more nuanced and comprehensive analysis of consumers' systematic purchase behaviour. This integration upholds the positivist commitment to empirical rigour while recognising the complexity of human behaviour that such frameworks aim to capture (Gosling et al., 2003).

By treating the Big Five personality data as reliable, this thesis makes use of a well-established psychological construct to enhance the depth and validity of its findings, thereby bridging the gap between quantitative data analysis and the potential richness, explanatory power of psychological insights as shown in several consumer research studies (e.g. Sandy et al., 2013, Bosnjak et al., 2007, Islam et al., 2017, Brunelle and Grossman, 2022, Di Crosta et al., 2021).

In summary, the positivist philosophy underpinning this study provides a rigorous framework for the investigation of systematic purchase behaviour. By emphasising empirical evidence, objectivity, and methodological rigour, the work aims to advance our understanding of consumer behaviour and contribute to the broader body of knowledge in the field. Through novel but rigorous research standards and interdisciplinary inquiry, the thesis seeks to uncover valuable insights that can inform both academic scholarship and practical decision-making in the retail industry.

3.2 Data provenance and technical framework

The real-world transactional data sets employed in the three studies of this thesis were sourced from two major retailers and one service company and complemented by survey data. These transactional data sets are rich in scale and scope, allowing for longitudinal and granular analyses of consumer purchase behaviour covering different time frames. The rationale and the description of the data sets used in each study are as follows.

Study 1 utilises historical transactional (purchase) data from two real-

world data sets that collectively enable a comprehensive examination of SPB and the development of the *bundle entropy* metric. These datasets were selected for their ability to complement one another in terms of scale, granularity, and contextual relevance, ensuring the metric's validity and applicability across diverse retail scenarios.

The study utilises historical transactional (purchase) data from two real-world data sets. The first is an open source dataset from *Dunnhumby* called *The Complete Journey*[1] (See Appendix A.0.1 for a description of the data set). Over a period of two years, this dataset includes grocery purchases at a household level from 2,500 frequent shoppers, providing a cohort for tracking SPB over time. It comprises over 2.5 million entries documenting household-level transactions, including detailed information on purchased items, quantities, purchase locations, and timestamps. All pertinent code to replicate experiments conducted with this dataset has been made available in Appendix B and on Github[2].

The second data set[3] comprises a vast collection of transactional records from 1,130,262 unique customers of a major UK-based grocery retailer. This dataset covers over 20 months between 2014 and 2016. The dataset captures details such as the type of product purchased, the quantity bought, the store location, and the time of purchase. Each transaction is linked to a specific customer through their loyalty card, enabling a more comprehensive analysis of individual buying patterns and consistency. In order to get a representative sample of regular customers, an inclusion criteria of a minimum of 5 purchases were established in both of the channels that the dataset includes (online and in-store stores). Out of the entire raw data set,

[1]The dataset is available at https://www.dunnhumby.com/source-files/

[2]https://github.com/rmansillal

[3]This second dataset is unavailable for public release due to a non-disclosure agreement

only 2,181 customers met the criteria, making it a relatively small sample. Nevertheless, this sample holds significant value as it provides insights into the systematic consumer purchase choices.

The combination of these datasets provides a framework for developing and testing the *bundle entropy* metric. The Dunnhumby dataset served as an ideal pilot environment for metric development, leveraging its controlled scale and public availability to ensure replicability and transparency. In contrast, the UK retailer's dataset introduced a greater diversity of shopping contexts and behaviours, enabling validation across a much larger and more varied customer base.

For study 2, the transactional and nutritional dataset comes from a leading UK grocery chain that identifies customers/households via online and in-store loyalty card IDs. This particular chain has an extensive physical and virtual presence across the country, providing customers with a wide range of product categories and well-known brands. The chain has over 3,000 physical stores in five different formats, which are spread out across various regions of the UK. Additionally, the chain has been operating its website since the late 1990s, which offers an online channel with nationwide coverage. Online products are sourced directly from physical stores, not regional distribution centres. As a result, a shopper can expect a consistent and comparable product offering both online and in-store, where they will be served by their local store for online deliveries.

In addition to transactional data, the study incorporates detailed nutritional information for each soft drink product gathered by utilising the Application Programming Interfaces (APIs) provided freely by the retailer. The data collected included the values for energy, saturated fat, sugar, sodium, fibre, and protein present in 100 ml of the drinks. This infor-

mation was then accurately linked to our transactional data (explained in more detail in section 5.1.5), which is maintained at the product level. This comprehensive data collection process allows us to provide a fairly complete picture of the nutritional content of each soft drink product.

As mentioned before, the transactional records come from the retailer's loyalty card program. Hence, members who have purchased online or in-store. The retailer has anonymized and processed the data according to confidentiality, privacy requirements, and standard ethical protocols. The data comprises sales records of more than 1 million members for a recent period comprising 19 months between 02/10/2014 and 31/05/2016 (before COVID). The *when*, *where*, *what*, and *how* each individual bought is known for each acquisition. Likewise, descriptors like the *quantity*, *price*, *product category/subcategory*, and a unique *product ID* that will associate them with their respective transactions are also known for each item acquired.

As previously stated, the business has different store formats, catering to both online and in-store channels. However, for this study, the website service was selected for the online channel, and three store formats were selected for the in-store channel. These formats were chosen based on their similar shopping missions and product flows, thus making them ideal for comparative analysis. As such, stores located in transport hubs and petrol stations were excluded to ensure a more accurate and relevant study.

This dataset was chosen for its ability to provide a comprehensive view of consumer behaviour across channels, supported by the richness of transactional and nutritional data. Its dual-channel structure facilitates robust comparisons between online and in-store SPB, shedding light on the influence of retail context on purchasing patterns. Moreover, the dataset's granularity supports a detailed examination of basket-level SPB and its

nutritional implications, highlighting its utility for addressing real-world retail and public health challenges.

Study 3 The data used in Study 3 comprises two complementary sources—a transactional dataset provided by a leading UK grocery and pharmacy chain and a survey dataset capturing socio-demographic, psychological, and behavioural information. Together, these datasets offer a rich, multi-dimensional framework for exploring systematic purchase behaviour (SPB) predictors and the relative importance of psychological, demographic, and behavioural factors.

The transactional dataset covers purchases made by survey participants across online and in-store channels, tracked through loyalty card IDs. Covering the period from January 1, 2012, to November 4, 2015, this dataset records detailed information on "what," "where," and "how much" individuals bought, similar to the structure of the transactional data used in *Studies 1 and 2*. However, this dataset includes additional product-level details, such as product names and sizes, enabling a finer-grained analysis of consumer purchasing patterns. The dataset was pseudo-anonymized by the retailer and processed in compliance with strict confidentiality, privacy, and ethical protocols, ensuring adherence to data protection standards such as GDPR. By leveraging this dataset, the study extends the *bundle entropy* metric to analyse purchase regularities within a diverse retail context, incorporating both grocery and pharmacy categories.

The survey dataset was initially designed to explore the impact of personality and psychological factors on purchase behaviours in the retail context. The complete survey is available in Appendix H.0.1. Although the survey was originally developed for understanding plastic bag usage (Lavelle-Hill et al., 2020) and not specifically designed to investigate *bundle entropy*, it

still provides valuable information on 12,835 common participants, including their socio-demographics, shopping behaviours and motivations, and psychological characteristics, which are described below:

- **Socio-demographics:** Some of the socio-demographic variables contained in this data are *age, gender, marital status, occupation, household income*, *household composition*, *education level*, among others.
- **Psychological characteristics:** The psychological variables in this dataset are self-control, impulsiveness (BAS) and personality traits, such as *extroversion, conscientiousness, agreeableness,* and *openness.*
- **Shopping behaviours and motivations:** The data contain purchase and motivation variables, such as *frugality, shopping impulsivity,* and *variety-seeking.*

The Survey data assessed these psychological and behavioural variables utilizing established Likert scales to capture individual subjective perceptions. While SHapley Additive exPlanations (SHAP) and Model Class Reliance (MCR) do not directly address the subjective nature of these variables, they facilitate an evaluation of each variable's influence on the model's outputs. This method allows the systematic interpretation of the impact of these variables without undermining their intrinsic subjectivity.

By anchoring these measurements in validated frameworks, the study aligns with a positivist approach that emphasizes empirical rigour while also acknowledging the variability introduced by self-reported data. Through SHAP and MCR, the model offers insights into the relative importance of these subjective variables, ensuring that both individual perceptions and broader trends in consumer behaviour are effectively captured within

its predictive framework. A list of all the variables used in this study is provided in section 6.1.5.

3.2.1 Technical framework

Across all three studies, a combination of programming languages, tools, and software platforms was employed to manage, preprocess, analyze, and visualize the data. These tools were selected for their robustness, scalability, and compatibility with large, multi-dimensional datasets.

- **Programming Languages:** SQL: Used extensively for data extraction, transformation, and integration from the relational databases provided by the retailers. Python: Served as the primary language for data preprocessing, feature engineering, statistical analyses, and machine learning tasks.

- **Software Tools:** PostgreSQL: This open-source relational database management system was used for structured data storage, querying, and transformation, ensuring efficient handling of large datasets from multiple sources. Google Colaboratory: Provided an interactive environment for conducting data exploration, statistical analyses, and machine learning experiments. It facilitated transparency and reproducibility in the analytical workflow. GitHub: Hosted the codebase for replicating experiments with the Dunnhumby dataset, ensuring that methods and analyses were accessible for independent validation.

- **Python Libraries:** Pandas and NumPy: Enabled efficient data manipulation and computation, particularly for multi-dimensional array data. SciPy and scikit-learn: Supported statistical analysis and

machine learning tasks. SciPy provided tools for hypothesis testing and feature selection, while scikit-learn powered the Random Forest models and Model Class Reliance (MCR) analysis. SHAP (SHapley Additive exPlanations): Applied in Study 3 to interpret the importance of individual predictors in the machine learning models, offering transparency in model outputs. Matplotlib and Seaborn: These are used for data visualization, including feature importance charts and cross-channel comparisons.

- **Specialized Tools:** Retailer APIs, in Study 2, APIs provided by the retailer were used to gather detailed nutritional information for soft drinks, ensuring accurate data.

3.3 Research Methods Integration

The thesis follows a logical and systematic progression, beginning with the development of a novel measurement for SPB, moving to its empirical application, and finally exploring its predictors. This structured methodology showcases the versatility of the bundle entropy as both an analytical tool and a conceptual anchor that guides the overall research design. Essentially, the logical flow of this research illustrates that bundle entropy not only addresses specific analytical inquiries but also provides a cohesive framework that connects various facets of consumer behaviour research, encompassing measurement, application, and interpretation.

In *Study 1*, the focus is the development of a propensity measurement called *bundle entropy*, which addresses limitations in existing measures such as Basket Level Entropy (BLE) and Basket Revealed Entropy (BRE). While useful for understanding basket-level repetition and basket 'predictabil-

ity', these existing metrics are insufficient for accurately capturing the predictability of multi-item basket compositions over time that accords to intuition (this is explained later in section 4.1.1). Bundle Entropy was designed as a normalized, parameter-independent measure that evaluates the regularity of product combinations over time. By utilizing datasets with both synthetic and real-world characteristics, Study 1 established the theoretical soundness and empirical robustness of the metric. This foundational work was critical in demonstrating the utility of Bundle Entropy for analysing SPB in complex retail environments.

Building upon the foundations established in *Study 1*, *Study 2* employed *bundle entropy* to examine SPB across both online and in-store retail channels. The dual-channel structure of the dataset facilitated a within-subject comparison, providing valuable insights into how the retail context influences purchasing behaviours. The study revealed significant differences in SPB between the two channels, underscoring the impact of environmental factors on consumer product choices. Additionally, the incorporation of nutritional data allowed for an analysis of SPB in relation to health-related products, such as the nutritional profiles of soft drinks. By demonstrating the practical relevance of *bundle entropy* in real-world retail settings, Study 2 broadened the metric's applicability and highlighted its potential for addressing challenges in public health and marketing.

Study 3 made significant strides in the research by integrating survey and transactional data to examine the predictors of SPB. In this study, *bundle entropy* was treated not only as a dependent variable—representing the outcome of systematic purchasing behaviour, but also as a framework for exploring its underlying drivers. By incorporating demographic, psychographic, and behavioural variables from the survey dataset, Study 3 offered a comprehensive perspective on the factors influencing SPB. Ad-

vanced machine learning techniques, including Random Forest and SHAP, were employed to assess the relative importance of these predictors. These methods provided valuable insights into the intricate relationships between individual traits, purchasing motivations, and systematic behaviours.

The integration of these three studies establishes a cohesive methodological narrative. Collectively, they reflect a rigorous and iterative research process rooted in scientific reasoning and methodological innovation. This thesis progresses from metric development to empirical application and finally to predictive analysis, demonstrating how the proposed measure can serve multiple purposes, ranging from explaining systematic purchasing patterns to identifying their key drivers.

3.4 Research Ethics

This thesis adheres to standard ethical conduct, considering the principles of transparency, reproducibility, and social responsibility in data protection, collection, and analysis. Each stage of the research process of each study was designed to ensure compliance with legal non-disclosure agreements with the providers of the data and broader ethical norms, particularly in the context of handling shopping consumer data.

A cornerstone of this ethical approach was the rigorous anonymization of all datasets used in the studies. The transactional data provided by major UK retailers was pseudo-anonymized by the retailers themselves before being made available for research purposes. This ensured that no personal identifiers, such as names or contact information, were included in the datasets, thereby safeguarding participant privacy and preventing the risk of re-identification, for example, by reverse engineering. All data processing

and analysis adhered to the General Data Protection Regulation (GDPR).

Informed consent was another key ethical consideration, particularly in Study 3, which integrated survey data with transactional records. Participants in the survey provided explicit consent for their data to be used in research and for it to be linked to their purchasing behaviours through loyalty card records.

The research emphasized social responsibility by tackling questions that yield tangible societal benefits. For instance, examining SPB with nutritional choices supports public health by providing actionable insights into dietary patterns. Furthermore, identifying the demographic and psychological predictors of SPB can aid in designing targeted interventions to foster healthier and more sustainable consumer behaviours. By aligning its research objectives with broader societal priorities, the thesis underscores its ethical commitment to generating knowledge that contributes to the public good.

Finally, the thesis adhered to principles of transparency and reproducibility throughout the research process. Key methodologies and analyses were documented and made publicly accessible, as referenced in the footnote on page 55. A public dataset, such as the Dunnhumby Complete Journey dataset, was employed to validate the results independently. For proprietary datasets, comprehensive descriptions and justifications for their use were provided in section 3.2, ensuring that the research is both accessible and comprehensible to academic and non-academic alike.

Chapter 4

Measuring Consumers' Systematic Purchase Behaviour in Retail

This chapter is based on work published at the IEEE International Conference on Big Data in 2022:

R. Mansilla, G. Smith, A. Smith and J. Goulding, 'Bundle entropy as a novel measure of consumers' systematic product choice combinations in mass transactional data,' 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 1044-1053, doi: 10.1109/BigData55660.2022.10021062.

Contents

4.1 Introduction

Understanding and measuring human behaviour predictability is increasingly valuable to scholars, as well as commercial and policy decision-makers. The significance of predictability measures in domains such as human mobility has been well studied (Song et al., 2010; Smith et al., 2014), with applications ranging from advertising and service provision to intelligent agents (Krumm, 2010; Jung et al., 2010; Froehlich and Krumm, 2008). However, there is a comparative lack of study on predictability in behaviours like purchasing patterns, despite the routine collection and processing of large-scale transactional data sets.

The retail industry possesses extensive and intricate datasets through loyalty programs (both in-store and online) and online purchase platforms. Many of these datasets contain unique identifiers that enable the tracking of individual consumers or households over time. This provides numerous opportunities for data-driven insights and management (Hossain et al., 2020; Foxall, 2001). For example, systematic or predictable consumers can be provided with relevant offers more efficiently, while unpredictable consumers may benefit from targeted innovations or varied direct offers. Assigning a predictability score to households or consumers can enhance retail segmentation and predictive analytics, creating greater opportunities for personalising responses and offers (Wen et al., 2018). Furthermore, consumer marketing communications are increasingly influenced by behavioural and propensity scores to ensure they align with consumer needs.

Driven by the significant potential for behavioural academics, retailers, and policymakers, this first study aims to provide more insights into consumer buying behaviour by developing a novel measure for assessing systematic purchase behaviour (SPB) from transactional big data. While prior studies

such as Guidotti et al. (2015) have touched upon this topic by presenting a measure for SPB based on basket predictability, this study aims to take a more in-depth approach and consider related measurements of basket variety (Straathof, 2007), heterogeneity (Nicolas-Sans and Ibáñez, 2021) and diversity (Jost, 2006; Budescu and Budescu, 2012) based on entropy. However, there are certain limitations in existing methods, which either fail to align with a more intuitive definition of basket predictability or are parameter-dependent/unstable.

These issues could significantly impact the practical applications of these methods on real-world data, a point that is elaborated in detail in section 4.3.3 of this study. To provide a more comprehensive understanding of the focus of this study, Table 4.1, in section 4.1.1, shows some synthetic examples of different purchase patterns that illustrate the goal of this study as well as the limitations of existing approaches.

4.1.1 Introducing the flaws of current predictability measures

As previously mentioned, existing methods for measuring basket predictability are ineffective in providing accurate and intuitive results. Although the Guidotti et al. (2015) measure performs intuitively in some scenarios; it relies heavily on manually defined parameters set by the user. Consequently, results may differ from user to user, affecting the comprehensive understanding of basket predictability.

The following examples are synthetic purchasing patterns of five different customers. Each basket (set) in a purchase history represents a unique shopping trip that includes one or more of the following items: *milk (m),*

butter (b), pasta (p), salmon (s), cheese (c), and yoghurt (y):

$C1 = \{m, b, y\}, \{m, b, y\}, \{m, b, y\}, \{m, b, y\}, \{m, b, y\}$

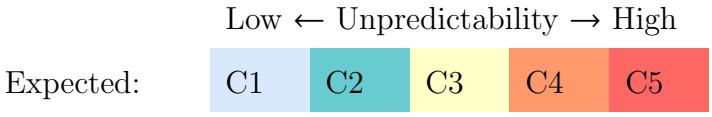
$C2 = \{m, b, p\}, \{m, b, p\}, \{m, b, p\}, \{m, b\}, \{m, b\}$

$C3 = \{m\}, \{m, b\}, \{m, p\}, \{m, s\}, \{m, c\}$

$C4 = \{m, b\}, \{b, p\}, \{p, s\}, \{s, c\}$

$C5 = \{m\}, \{b\}, \{p\}, \{s\}, \{c\}$

After analysing the examples above, it can be deduced that customer $C1$ is the most predictable among all the customers, as it tends to buy the same items during each visit. Similarly, $C2$ is also relatively predictable since it consistently purchases milk and butter but only occasionally buys pasta. However, the basket composition of $C3$ is more unpredictable, as this individual only purchases milk regularly but switches the second items on every visit. $C4$ and $C5$ are the hardest to predict, and although $C4$ has some commonality across their baskets, it is still difficult to determine their purchasing pattern. Therefore, to measure the predictability of these customers, a measure that produces the following ordering is required.



Nonetheless, as shown in Table 4.1, current measurement techniques do not accurately reflect this intuition. For instance, the Item Entropy (IE) measure does not capture the expected ordering of the data. Similarly, both Basket Level Entropy (BLE) and Guidotti et al. (2015) measure, Basket Revealed Entropy (BRE), are unable to distinguish the critical fact that the values of $C1 < C2 < C3 < C4 < C5$, even across various parameterizations.

The inability of these measures to match applied intuitions, even in simple synthetic examples, has highlighted the need for this work. To ensure clarity, the specific formulas utilized for calculating these measures are detailed in Section 4.3.3, where a thorough discussion of their limitations is also provided.

Table 4.1: Examples of consumer purchasing behaviour where the effectiveness of current approaches to measure basket predictability are insufficient.

	Normalised score from 0 to 1				
Measure	C1	C2	C3	C4	C5
Item Entropy (IE)	1.0	0.98	0.81	0.97	1.0
Basket Level Entropy (BLE)	0.0	0.97	1.0	1.0	1.0
Basket Revealed Entropy (BRE-low param)	0.0	0.97	1.0	1.0	1.0
Basket Revealed Entropy (BRE-med param)	0.0	0.97	0.0	1.0	1.0
Basket Revealed Entropy (BRE-high param)	0.0	0.0	0.0	1.0	1.0

This study acknowledges that measuring predictability at the basket level is also a focus of other behavioural studies. Similar to Guidotti et al. (2015), this study determines someone's predictability based on the composition of their basket or a sub-basket. This is of utmost importance because it allows us to gain certainty that a customer's next basket will include specific items, increasing the overall utility value. Furthermore, it reflects real-world conditions where baskets with the same content are rare due to factors such as availability, variety-seeking, promotions, cross-retailer shopping, and group purchasing. Neglecting to factor in the randomness of a few items in a basket might result in labelling numerous customers as unpredictable when they are fairly predictable. Therefore, it is imperative to consider and factor in regularity within baskets while measuring predictability to ensure accurate results.

To visually illustrate the differences between BE and existing metrics, let's consider the following set of three purchases (baskets) for a hypothetical customer ($C1$):

$C1 = \{A, B, C\}, \{A, B, D\}, \{A, B, E\}$

In this example, BE accurately reflects the customer's partial predictability. The consistent purchase of products A and B across all baskets suggests a stable purchasing pattern, while the variation in products C, D, and E introduces some unpredictability. BE assigns a value of 0.23 (see Table 4.1), which reflects this combination of stability and variation. In contrast, other measures either overestimate or underestimate the predictability.

IE, for example, calculates entropy (using equation 3 in Appendix A) at the individual product level without considering basket composition, leading it to report a value close to 1. In this case, IE measures the frequency of individual products but fails to account for the fact that some items (A and B) are consistently bought together. BLE, which assesses each entire basket as a single unit (see equation 4 in Appendix A), also assigns a value of 1, misrepresenting the evident regularity in products A and B. BLE treats each basket as entirely different because the third product in each basket varies, ignoring the common sub-basket.

Similarly, BRE measures predictability by identifying frequent sub-baskets through an algorithm that groups common items, depending on the parameterization (see equation 5 in Appendix A). In this example, BRE at thresholds like 10%, 24%, and 70% either returns a value of 0 (assuming full predictability) or 1 (assuming full unpredictability), missing the balanced predictability of A and B paired with the variable products C, D, and E. BE, however, uses a set similarity approach that measures both shared and unique items across baskets, penalizing unpredictability while accounting for the consistent sub-basket of A and B. By striking this balance, BE provides a more accurate and intuitive measure of predictability, recognizing

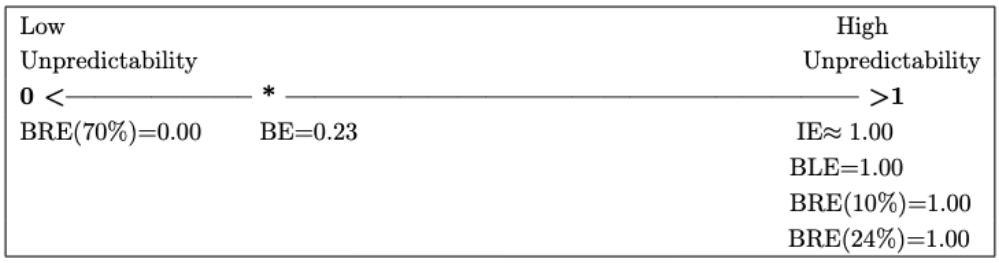


Figure 4.1: The accuracy of Bundle Entropy against existing metrics in accurately measuring the stability of product choice combinations across purchases.

both the regularity in certain product choices and the variation in others.

This study aims to quantify the predictability of purchases from transactional data. This section provides an overview of related measures that attempt to achieve this, highlighting the different definitions of predictability that these measures encode. It is argued that none of these accurately and consistently quantifies the predictability of basket purchases.

Section 4.3.2 lists the properties that such a measure should follow. The following three sections, 4.3.3, 4.3.4, and 4.3.5, detail the shortcomings of existing approaches against these properties. This study then introduces a novel measure called *bundle entropy* and theoretically demonstrates its utility.

In the final section 4.5, the study compares *bundle entropy* to existing approaches on two real-world data sets, empirically demonstrating its effectiveness. The study concludes with discussions and conclusions in section 4.5.

4.2 Current work

Understanding consumer purchasing behaviour and accurately predicting it holds significant value. Although entropy measures have proven effective in fields such as physics and information theory, their application to multi-dimensional retail datasets, like those utilized in this thesis, remains uncommon. These datasets capture multiple dimensions of consumer behaviour, including the composition of individual shopping baskets (the items purchased), the temporal dimension of the transactions (e.g., time and frequency), and spatial factors (e.g., store locations).

While a few existing methods for analysing this data are available, it is crucial that they are intuitive to minimize potential misunderstandings between analysts and decision-makers. Consequently, this study seeks to introduce a new measure of predictability that can effectively evaluate the predictability of individual basket compositions while ensuring both consistency and intuitiveness.

This study is divided into two, *Study 1a* and *1b*. *Study 1a* introduces and evaluates the new measure using synthetic and real-world data sets. On the other hand, *Study 1b* is a case study that explores the utility of the proposed measure to understand consumers' basket predictability compared to current measures using real-world data.

This study demonstrates that current methods are inconsistent and parameter-dependent and can lead to misinterpretation. In this context, *Study 1a* suggests the properties for such a measure (introduced in section 4.3.2). Additionally, it shows how current methods do not accord to these properties (exhibited in section 4.3.3). It introduces a novel measure called *bundle entropy* (section 4.3.4), which directly estimates the predictability

of basket composition. *Bundle entropy* assigns values between zero and one (when normalised) to denote total predictability and total unpredictability, respectively. *Study 1a* finalises by demonstrating how it met the proposed properties using two real-world transactional data sets, which are also then used in *Study 1b* to assess the measure's effectiveness (see section 4.4.3). Each dataset contains over 2,000 households of frequent shoppers for two years. Overall, the study (*1a* and *1b*) demonstrates the following:

- That *bundle entropy* is the only measure that meets the desired properties
- The study provides empirical evidence (shown in section 4.3.2) that *bundle entropy* is distinct from other measures.
- The study then analyses and introduces some use cases and discusses the practicality of *bundle entropy* in the retail industry, further explored and discussed in study two.

4.3 Study 1a: Bundle entropy as a novel measure of consumers' systematic purchase behaviour

4.3.1 Study Design

Based on the predictability intuition and practical use of predictability measures explained in section 4.1.1 and Table 4.1, the study proposes three desired properties (P0, P1, and P2) for a measure that would address the shortcomings of current measures. The current measures were then eval-

uated against these properties to determine their effectiveness. Using two real-world retail data sets, the proposed measure was introduced mathematically and evaluated against the desired properties. To ensure accuracy, strict inclusion criteria were established to extract frequent customers from the initial data set.

4.3.2 Methods

Measuring the predictability of basket composition

Study 1a proposes a measure of purchasing behaviour involving three key properties. These properties aim to capture basket composition's predictability while aligning with real-world applications. By considering these properties, the measure can better reflect consumers' shopping habits and provide valuable insights into purchasing behaviour (see Table 4.1).

Let $\mathcal{B} = \{b_0, b_1, \ldots, b_n\}$ represent an individual's set of baskets, where baskets are sets of unique items $b_x = \{\gamma_0, \gamma_1, \ldots\}$, and $\mathcal{M}(\mathcal{B})$ is the value of the measure assigned to a given $\mathcal{B}$. Then:

P0: If all baskets in a sequence contain the exact same items, then that sequence should receive a score of *zero*. On the other hand, if no basket in a sequence shares any items with any other basket, then that sequence should receive a score of *one*. In cases where normalisation is required, the maximal value should be considered (normalisation is discussed in more detail later on). Thus:

$\mathcal{M}(\mathcal{B}) = 0 \quad \text{if } b_0 = b_1 = \ldots = b_n$

$\mathcal{M}(\mathcal{B}) = 1 \quad \text{if } b_0 \cap b_1 \cap \ldots \cap b_n = \emptyset$

P1: The resulting score decreases when a purchasing sequence contains

more common sub-baskets. Conversely, the resulting score increases when the purchasing sequence contains fewer common sub-baskets. Formally, if Γ is any arbitrary combination of items that were not previously present in every basket of $\mathcal{B}$ (i.e. $\exists b_k - \forall b_k : \Gamma \not\subset b_k$), and we add Γ to each basket to produce $\mathcal{B}'$, then so long as $\mathcal{B}$ wasn't already totally predictable:

$\mathcal{M}(\mathcal{B}') < \mathcal{M}(\mathcal{B})$ if $\mathcal{B}' = \{b_0 \cup \Gamma, b_1 \cup \Gamma, \dots\}$

P2: Sequences with larger systematic sub-baskets should have a lower score than sequences with smaller sub-baskets (relative to any basket size) unless the sequence is already fully predictable (meeting property $P0$).

$\mathcal{M}(\mathcal{B}^*) < \mathcal{M}(\mathcal{B}')$ if

$\mathcal{B}' = \{b_0 \cup \Gamma, b_1 \cup \Gamma, \dots,\}$

$\mathcal{B}^* = \{b_0 \cup \Gamma \cup \Gamma', b_1 \cup \Gamma \cup \Gamma', \dots,\}$

In this context, Γ and Γ' represent two distinct combinations of items. The addition of Γ' to Γ in $\mathcal{B}^*$ increases the predictability of the system, as Γ' introduces additional systematic patterns. This enhanced predictability reduces the Bundle Entropy, reflecting a more structured and orderly system compared to the original combination with Γ alone. This demonstrates the measure's sensitivity to systematic sizes, aligning with the principle that greater predictability should result in lower entropy.

The properties described above are based on the expectations from BLE. Property $P0$ defines the expected behaviour in the most extreme cases. Predicting the next random basket is very easy if all baskets are identical. There is no uncertainty; in such cases, the measure should equal zero. However, if no baskets share any items, predicting the next basket becomes

difficult and uncertain. This is because there is maximal uncertainty as to what the subsequent basket should be. In such cases, the measure should reflect this level of uncertainty.

These properties are consistent with those encoded by BLE when the task degrades into predicting a set of identical baskets, where symbols and composition are unimportant. At the same time, the properties are also consistent with predicting a set of unique baskets, where symbols and composition are unimportant.

Compared to BLE, $P1$ provides a more flexible approach to handling uncertainty in cases where a consumer has repeated sub-baskets. BLE considers predicting the entire basket as the sole task and does not account for the predictability gains when a consumer's purchase history contains repeated sub-baskets. In contrast, $P1$ is designed to predict a sub-basket that is significant to a decision-maker, thereby allowing for more relaxed and context-specific predictions.

Furthermore, $P2$ recognises that larger predictable sub-baskets provide more valuable insights into real-world applications. The more predictable components in a customer's purchase history, the more certain one can be about their future purchasing behaviour and overall spending. $P2$ formalises this relaxation even further, allowing for even more accurate predictions that consider the predictability gains that occur when a customer's purchase history contains repeated sub-baskets.

It is worth noting that measures of predictability based on entropy are usually normalised. This accounts for the fact that as the number of possible outcomes increases, the complexity of predicting them also increases. If we use non-normalised versions of these measures to compare individuals, we conflate their uncertainty within a given set of options (how predictable

they are relative to what they have access to) with their access to a larger set of options. In the case of consumer goods purchasing, factors like household size and income often drive and constrain decision-making Li and Russell (1999); Bellman et al. (1999). Therefore, it's generally desirable to have a measure that is independent of these factors, which can be achieved by normalising against some measure of choice set size (such as the number of unique baskets or items). We follow the approach of Guidotti et al. (2015) and normalise the proposed measure by dividing it by the number of unique baskets. It is worth noting that other normalisation methods could also be used to achieve invariance to different aspects or definitions of choice group sets. The normalisation method used in this study aligns with that used in BLE and BRE, as it enables better evaluation of the proposed measure's properties ($P0$-$P2$). If desired, a practitioner could also use the non-normalised version of the proposed measure while maintaining all motivating properties, as their associated proofs still hold (see Appendix C.0.1).

4.3.3 Failure of existing methods

To delve deeper into the measures mentioned above, the study introduces some notation for entropy at the basket and item level as well as BRE.
Let $\mathcal{B} = \{b_0, b_1, \ldots, b_n\}$ represents an individual's set of baskets, where baskets are sets of unique items $b_x = \{\gamma_0, \gamma_1, \ldots\}$. Let $p(\gamma)$ denotes the probability of γ occurring in a basket in $\mathcal{B}$. Additionally, let $p(b)$ be the probability of b, an observed basket in $\mathcal{B}$, and B be the set of unique baskets in $\mathcal{B}$. Finally, let $I = \bigcup_{b\in\mathcal{B}} \bigcup_{\gamma\in b}$ be the collection of distinct items bought (and in all cases, $0 \log_2 0$ is taken to be 0 as per convention).
At the item level, normalised entropy is defined as follows:

$$IE(\mathcal{B}) = -\frac{1}{\log_2 |I|} \sum_{\gamma \in I} p(\gamma) \log_2 p(\gamma) \tag{4.1}$$

At the basket level, normalised entropy (which the study refers to as BLE) is defined as follows:

$$BLE(\mathcal{B}) = -\frac{1}{\log_2 |B|} \sum_{b \in B} p(b) \log_2 p(b) \tag{4.2}$$

On the other hand, BRE, proposed by Guidotti et al. (2015), involves creating a new list of baskets called $\mathcal{B}' = \{b'_0, b'_1, \ldots, b'_n\}$. This new list replaces each basket $b \in \mathcal{B}$ with a common sub-basket b' using a specific algorithm described below:

1. To begin with, the algorithm needs to identify a group of potential *common sub-baskets* using the Apriori algorithm Agrawal and Srikant (1994). This must be done by setting a minimum support parameter defined by the user. The Apriori algorithm is a fundamental data mining technique used for frequent item set mining and association rule learning. It identifies patterns or item combinations that appear frequently in a dataset, such as customer transaction records. The algorithm relies on the downward closure property, which states that if an item set is frequent, all its subsets must also be frequent. Steps

of the Apriori Algorithm:

(a) Generate Candidate Itemsets: Start with single items (1-itemsets) and count their occurrences in the dataset. Retain only those that meet the minimum support threshold (*minsup*). Combine frequent itemsets to generate larger candidate itemsets (e.g., 2-itemsets, 3-itemsets, etc.).

(b) Prune Non-Frequent Itemsets: If any subset of an itemset is not frequent, the itemset itself cannot be frequent (based on the downward closure property). This reduces the search space.

(c) Repeat Until No More Frequent Itemsets: Continue generating and pruning itemsets until no new frequent itemsets can be identified.

(d) Output: The remaining frequent itemsets are patterns that meet the *minsup* threshold.

2. Next, each basket is replaced by a single *common sub-basket* based on the following rules (expanding the set of *common sub-baskets* as required):

(a) **RULE 1:** The longest *common sub-basket* contained in the basket is selected to be the *common sub-basket*. There are additional rules for tie-breaking, which can be found in Guidotti et al. (2015).

(b) **RULE 2:** If there is no *common sub-basket* contained in the basket, which may happen depending on the minimum support value set by the user, then the full basket is considered, and a new symbol is added to the list of *common sub-baskets*.

Let $\boldsymbol{B}'$ be a set of unique baskets in $\mathcal{B}'$.

BRE is then defined as:

$$BRE(\mathcal{B}') = -\frac{1}{\log_2 |B'|} \sum_{b' \in B'} p(b') \log_2 p(b') \tag{4.3}$$

When considering BRE as a measure of basket entropy of assigned *common sub-baskets* and examining its behaviour as its parameterization, the Apriori algorithm's minimum support (*minsup*) is an essential factor to consider. By varying the *minsup*, two key points emerge:

1. When *minsup* approaches zero, all baskets become part of the candidate *common sub-basket* set, and all $b'_x = b_x$ (due to RULE 1).

2. When *minsup* approaches one, it depends on the data. If a *common sub-basket* exists in all baskets (i.e., $\{m, b, p\}, \{m, s, c\}, \{m, b, j\}$), then RULE 1 will apply, leading to all b'_x being all the same (m in the example), though in most real-world cases this will not be true (i.e., $\{m, b, p\}, \{m, s, c\}, \{s, c, j\}$). In this case, the candidate *common sub-basket* set will contain no candidates (as *minsup* is close to 1) and all $b'_x = b_x$ via RULE 2.

After computing the entropy by considering each unique *common sub-basket* as a symbol, the BRE degenerates to entropy at a basket level (BLE) for the most common two of the three cases.

Property violations in existing measures

In this section, the study examines the current limitations of predictability measures in relation to the intuitive properties described in section 4.3.2.

One of the primary limitations of IE is its inability to meet the $P0$ property

in theory, which is unlike BLE and BRE. To better understand the inadequacy of IE in meeting this property, we can analyse the simple example presented in Table 4.1. For instance, when all baskets in a dataset are identical (e.g. $C1$ in Table 4.1), such that there is only one basket $B = \{b\}$, and $p(b) = 1$, BLE meets the $P0$ requirement by returning zero. This is because the data has no uncertainty since all baskets are the same. In contrast, IE fails to meet this property since it calculates the entropy of each item in the basket separately. In this case, all items have an entropy of 1, leading to the maximum entropy for the basket (See IE for $C1$ in Table 4.1). On the other hand, when all baskets are distinct (e.g. $C5$ in Table 4.1), BLE is optimised trivially, with every item in each basket having $p(b) = \frac{1}{|B|}$. In this scenario, IE does meet property $P0$ just because it considers the entropy of every unique item within the basket. Once the baskets start to contain items that disrupt a consistent pattern or add a distinct purchase pattern, neither IE nor BLE provides the correct score. For instance, if we consider Table 4.1, $C2$ includes an item (p) that is bought only some of the times making it less predictable than $C1$, while $C4$ has some items that are bought more than others making it slightly more predictable than $C5$.

BRE satisfies property $P0$ regardless of the parameterization used. When all baskets share the same items under any parameterization, the Apriori algorithm always identifies the entire repeated basket composition as a *common sub-basket*. This means that if a particular itemset appears in multiple baskets, it is considered as a single entity. On the other hand, when all baskets are unique, there cannot be any *common sub-baskets*. In such cases, the original basket is always used according to RULE 2, which states that if there is no common subset, the original subsets should be used. In both scenarios, BRE computation proceeds with the same input as BLE, ensuring consistency in those extreme scenarios (e.g. BRE for $C1$

and $C5$ on Table 4.1).

Further, IE violates $P1$, with the inclusion of systematic behaviour at the basket level having the potential to lead to an increase in IE in some cases. To illustrate this, let's consider the following example: $\{m, b\}, \{b, p\}, \{b, p\}, \{b, p\}$ and $\{m, b\}, \{m, b, p\}, \{m, b, p\}, \{m, b, p\}$. The second case clearly has a higher presence of systematic sub-baskets. The IE in the first case, where there are 4 b's, 3 p's, and 1 m, is 0.887. However, after adding item m into each of the systematic sub-basket that did not contain it, the IE, instead of decreasing, increases to 0.992, where there are 4 b's, 3 p's, and 4 m's.

BLE also violates $P1$. To illustrate further, let's consider a simple example where we have a basket sequence consisting of $\{m, b, p\}, \{m, b, s\}, \{m, b, c\}, \{m, b, j\}$. Increasing the presence of systematic sub-baskets ($P1$) would not change the fact that every basket in the sequence would remain distinct at the basket level. As a result, the BLE score would be maximal.

In the case of BRE, the exact score depends on the parameter that is manually chosen by the user. Let's consider the following two scenarios:

- *Scenario 1*: $\mathcal{B} = \{p, s\}, \{p, s\}, \{c, j\}, \{c, j\}$
- *Scenario 2*: $\mathcal{B}' = \{p, s, m, b\}, \{p, s, m, b\}, \{c, j, m, b\}, \{c, j, m, b\}$ (where $\{m, b\}$ has been added to each basket systematically).

Because $\{m, b\}$ has been systematically added to each basket to form *Scenario 2*, this last is intuitively more predictable than *Scenario 1*. However, for BRE, when the minimum support (*minsup*) is set such that both $\{p, s\}$ and $\{c, j\}$ are mined as frequent patterns, in *Scenario 1*, $\{p, s\}$ becomes a distinct symbol, *common sub-basket 1* (X), and $\{c, j\}$ becomes a distinct symbol, *common sub-basket 2* (Y). By RULE 1, BRE is evaluated as the

BLE of $\{X, X, Y, Y\}$.

For the same minimum support threshold in *Scenario 2*, many two-item frequent patterns exist, but so do the longer *common sub-baskets* $\{p, s, m, b\}$ and $\{c, j, m, b\}$. By RULE 1, it is these, and only these, that will be selected to represent the baskets (as distinct symbols), and again, BRE is evaluated as the BLE of $\{X, X, Y, Y\}$. This violates $P1$ without any decrease reported by the measure. If the minimum support threshold for *Scenario 2* is increased to the point where $\{m, b\}$ is the only frequent pattern, BRE will return a score of zero due to the absence of uncertainty.

These examples serve as a clear demonstration that BRE is inadequate in accurately assessing the joint presence of structure and randomness in a dataset for a given threshold value. Although these may seem like artificial scenarios, these are not uncommon real-world occurrences, as will be demonstrated later through empirical evidence. Additionally, the examples emphasise the susceptibility of BRE to its parameter, as the resulting score can vary significantly depending on the chosen parameter.

Finally, let's consider $P2$, which is an extension of $P1$. $P2$ provides a better understanding of the expected behaviour based on the repeated application of $P1$. This means that measures which fail $P1$ cannot satisfy $P2$. Therefore, IE, BLE, and BRE cannot fully comply with $P2$.

4.3.4 Introducing Bundle Entropy

The previous section discusses the limitations of IE, BLE, and BRE by using various *minsup* parameter values and examining their failure to meet certain properties desired to assess SPB. The main goal of this study is to provide a practical description of human predictability regarding con-

sumers' purchase patterns. We have observed that the connection between BRE and entropy is not fully explored in the study conducted by Guidotti et al. (2015) due to their definition-by-algorithm approach, which somewhat obfuscates their measure's definition of predictability. To address these issues, we propose a new method called *Bundle entropy* that satisfies properties $P0$, $P1$, and $P2$. *Bundle entropy* is an extension of BLE that conceptualises bundles as a collection (set) of products bought simultaneously. Additionally, we reframed BLE's formulation to the (normalised) mean information for all baskets.

Let $\mathcal{B} = [b_0, b_1, \ldots, b_n]$ represents an individual's list of baskets, where each basket is a set of unique items. Additionally, let $B = set(\mathcal{B})$ and $p(b_k)$ denote the empirical probability of basket b_k given $\mathcal{B}$.

Then:

$$I(b_k) = -\log_2(p(b_k)) \tag{4.4}$$

Where $I(b_k)$ is the well-known measure of self-information, measuring the amount of *surprise* we receive when b_k is observed given we expected b_k with probability $p(b_k)$. Given $I(b_k)$, BLE is then:

$$\begin{aligned} BLE(\mathcal{B}) &= \frac{1}{\log_2 |B|} \sum_{b \in B} p(b) I(b) \\ &= \frac{1}{\log_2 |B|} \times \frac{\sum_{b \in \mathcal{B}} I(b)}{|\mathcal{B}|} \end{aligned} \tag{4.5}$$

Note the abovementioned distinction between B and $\mathcal{B}$. Ignoring the normalisation term, the final line highlights that non-normalised BLE accurately represents the average self-information over observed data, which is typically assumed to represent population statistics.

Let's delve deeper into the concept of self-information. Essentially, $I(b_k)$ measures the level of surprise we would experience upon correctly predicting b_k and observing the set of baskets ($\mathcal{B}$) with accuracy only $p(b_k) \times |\mathcal{B}|$ times. This underscores the importance of accurate predictions when making decisions. Alternatively, we can interpret $I(b_k)$ as the average level of dissatisfaction we would feel if we were to predict b_k indefinitely, assuming that the empirical probability $p(\cdot)$ is the true generative distribution. Therefore, the calculation of $I(b_k)$ depends on the empirical probability, which can be represented as follows:

$$p(b_k) = \frac{\sum_{b_q \in \mathcal{B}} \mathbb{1}(b_k = b_q)}{|\mathcal{B}|} \tag{4.6}$$

Leveraging the prediction point of view, the study aims to capture sub-baskets' predictability within the context of Basket Entropy. The study acknowledges that even if the prediction is not *precisely* accurate, it can still be considered a positive outcome if it calculates the prediction's partial value based on an anticipated utility measure. The assumption is that correctly predicting sub-baskets generates utility, and the utility increases proportionally to the sub-basket's size that is accurately predicted. Since the study views baskets as collections of distinct items, it employs a set similarity measure such as Jaccard (Niwattanakul et al., 2013) or Overlap (Lawlor, 1980). These measures align with the exact match similarity

function at the two extremes (zero: no partial match, one: exact match). The proposed approach in the study is to use a variant of the Overlap coefficient, specifically:

$$\mathcal{S}(b_k, b_q) = \frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)} \tag{4.7}$$

The measure is defined as the proportion of items that are common between the predicted and actual (truth) sets. The numerator of this measure is the number of shared items between these two sets. This measure is similar to the Jaccard and Overlap measures regarding the numerator but differs in how the denominator is computed. On the one hand, Overlap employs the proportion of the smaller basket ($\frac{|b_k \cap b_q|}{min(|b_k|,|b_q|)}$), potentially neglecting over-predictions. On the other hand, Jaccard uses the number of matched and unmatched items between the prediction and truth ($\frac{|b_k \cap b_q|}{|b_k \cup b_q|)}$), which may result in double-counting incorrect predictions and failing to penalise over-predictions. To tackle this issue, the overlap variant in Equation 4.7 counts incorrect predictions only once, effectively penalising over-predictions. This approach is a reliable means of accurately evaluating the shared portion of items between predicted and true sets.

We can replace Equation 4.7 in Equation 4.4 using Equation 4.6. By replacing the exact match indicator function ($\mathbb{1}(b_k = b_q)$) with the similarity function ($\frac{|b_k \cap b_q|}{max(|b_k|,|b_q|)}$), we can define the bundle self-information as an alternative measure, which we call *regret* ($R(b_k)$). This measure does not quantify how surprised one is when observing b_k, but rather how much regret one might feel if one assumed that b_k was going to occur.

$$\mathcal{R}(b_k) = -\log_2\left(\frac{\sum_{b_q \in \mathcal{B}} \frac{|b_k \cap b_q|}{max(|b_k|,|b_q|)}}{|\mathcal{B}|}\right)$$

$$\mathcal{R}(b_k) = -\log_2\left(\sum_{b_q \in B} p(b_q)\frac{|b_k \cap b_q|}{max(|b_k|,|b_q|)}\right) \tag{4.8}$$

The regret-based *bundle entropy* (BE) that has been proposed can be defined as follows:

$$BE(\mathcal{B}) = \frac{1}{\log_2 |B|} \times \sum_{b_k \in B} p(b_k)R(b_k) \tag{4.9}$$

The proposed measure *bundle entropy* has been thoroughly evaluated against the properties $P0-P2$, and all of them have been successfully satisfied. The detailed proofs are presented in Appendix C.0.1. By conforming to these properties, the measure $BE(\mathcal{B})$ performs consistently and reliably in the given examples, as demonstrated in Figure 4.2.

	Low ← Unpredictability → High				
Expected	C1	C2	C3	C4	C5
Bundle Entropy	0.0	0.25	0.32	0.6	1.0

Figure 4.2: Examples of how bundle entropy more accurately measures purchase predictability across customers C1 to C5.

4.3.5 Evaluation and Discussion

This section will explore two parts that provide empirical evidence for the *bundle entropy* measure's superiority over the other measures. The first part demonstrates how the *bundle entropy* measure fulfils the desirable properties outlined in section 4.3.2, while other measures do not. The

second part presents empirical evidence that highlights the significant differences between the proposed measure and other measures, emphasising how a practitioner's choice of measure can lead to varying and inconsistent conclusions.

Each of the parts described above compares and contrasts the effectiveness of *bundle entropy* against other measures in predicting consumer buying behaviour from transactional data. The study focuses on three measures described in previous sections, namely IE, BLE, and BRE using different *minsup* values of 10%, 24%, and 70%. The *minsup* value directly affects the *common sub-baskets* used to represent purchase history, as explained in section 4.3.3. While this study used *minsup* value of 24%, based on recommendations from Guidotti et al. (2015), it also tested the performance of BRE with *minsup* values of 10% and 70% to account for variations in dataset size and context.

The evaluations are conducted using the two datasets that were discussed in the previous section 3.2.

Quasi-synthetic Data

In this particular section, we shall delve into the alignment of *bundle entropy*, BLE, BRE, and IE with the desired properties $P1$ and $P2$ that were stipulated in section 4.3.2. The study will not consider $P0$ as it deals with the predictability of edge cases where IE is known to fall short. However, it's important to mention that IE does not accord with $P0$ since it assesses predictability at a different level, focusing on individual items rather than baskets or sub-baskets in a set of purchases. On the other hand, the remaining measures successfully meet $P0$.

$P1$ states that baskets with systematic sub-baskets should result in a lower score than those without. In order to evaluate the effectiveness of the proposed and the current measures, the study used Dunnhumby's dataset by incorporating systematic sub-baskets into each household's basket. In this approach, each original basket was modified to include 10 randomly selected items from the entire dataset, allowing for the computation of each measure. This step is essential for satisfying Property 1, which asserts that a measure's score should decline when systematic sub-baskets are present, thereby supporting the expectation of higher predictability (see Figure 4.3 for an illustrative example).

Measures aligned with property $P1$ consistently will produce the lowest measure value compared to the initial basket collection [1].

Table 4.2 summarises the results of the analysis on how various measures align with properties $P0$ and $P1$. The data reveals that all measures, with the exception of IE, satisfy property $P0$. As mentioned earlier, the study evaluated each measure before and after adding systematic sub-baskets to the data set. This is reflected in Table 4.2, which presents the percentage of households that decreased their initial score, indicating compliance with property $P1$, for each measure.

As anticipated, the proposed *bundle entropy* consistently satisfies property $P1$ for every household, resulting in a perfect score of 100%. Similarly, the IE aligns with property $P1$ with only a few minor exceptions, achieving a score of 99%. However, the BLE score does not match due to its basket-level approach to property $P1$. Hence, despite implementing systematic baskets, the number of unique baskets remains unchanged, resulting in no

[1] Unless an individual's basket sequence is already maximally predictable, meeting $P0$, something that did not occur in this data set. See the definition of $P1$ in section 4.3.2 for more information.

Table 4.2: Measures vs. Properties 0 & 1 and the percentage of households considered as fully predictable.

	Property accorded to:		% Households measure considered fully predictable
Measures	**P0**	**P1 (% Households)**	
Bundle Entropy	✓	100.0	0.0
Item Entropy	✗	99.0	0.0
BLE	✓	0.0	0.0
BRE 10%	✓	70.9	5.2
BRE 24%	✓	63.2	5.1
BRE 70%	✓	99.8	98.8

reduction in the BLE score (See illustrative examples in Figure 4.3).

According to the details provided in section 4.3.2, it is important to note that the fulfilment of $P1$ by BRE depends on the data and threshold utilized. The data presented in column three of Table 4.2 highlights that the violation of $P1$ is not uncommon in real-world scenarios, especially when the *minsup* levels are lower. For example, when the *minsup* level is set at 10%, BRE meets property $P1$ for only 70.9% of households that satisfy it. Similarly, when the *minsup* level is set at 24%, BRE meets it for 63.2% of households. However, when the *minsup* level is set at 70%, BRE satisfies property $P1$ for 99.8% of cases.

To further investigate this behaviour, the study examined individual household scores and selected three representative cases to illustrate the findings, as shown in Figure 4.3. The results generally aligned with expectations, with one notable exception: when systematic sub-baskets (represented by the lower orange dots) were added to the BRE algorithm using a *minsup* of 70% (and, in the case of household 2, with a *minsup* of 24%), the addition was handled incorrectly. Specifically, instead of adjusting the score to reflect the added systematic pattern, the households were classified as 100%

predictable. While this outcome technically satisfies Property 1, since the addition of systematic sub-baskets reduces entropy, it does so in an extreme manner. This extreme behaviour arises because the algorithm effectively ignores the variability within the systematic sub-baskets, treating the household's entire basket sequence as entirely predictable, regardless of the size or composition of the added sub-baskets.

To assess the extent of this issue, the study calculated the percentage of households affected for each measure, with results summarized in column four of Table 4.2. The findings reveal a clear trend: as the *minsup* parameter of BRE increases, the measure increasingly satisfies Property 1 in this flawed manner. At higher *minsup* levels, the algorithm predominantly identifies only the most frequent sub-baskets, leading to an oversimplification of household purchasing behaviour and ultimately classifying a disproportionately high number of households as fully predictable. This severely limits the practical usefulness of BRE at high *minsup* values, as it undermines the nuanced evaluation of systematic purchase patterns and fails to differentiate between genuine predictability and artificially induced outcomes.

$P2$ states that when a sequence of baskets contains larger systematic sub-baskets, it should result in a comparatively lower score than sequences of baskets with smaller sub-baskets relative to the size of the basket. To investigate the empirical performance of the measures regarding this property, the study randomly selected 1,000 households[2] and added systematic bundles of varying sizes (ranging from one to ten items) to their baskets. After each iteration, the mean score per measure across all households was calculated and presented the results in Figure 4.4. As predicted, the *bun-*

[2]A 1,000 households was the maximum sample that could be taken due to computational costs.

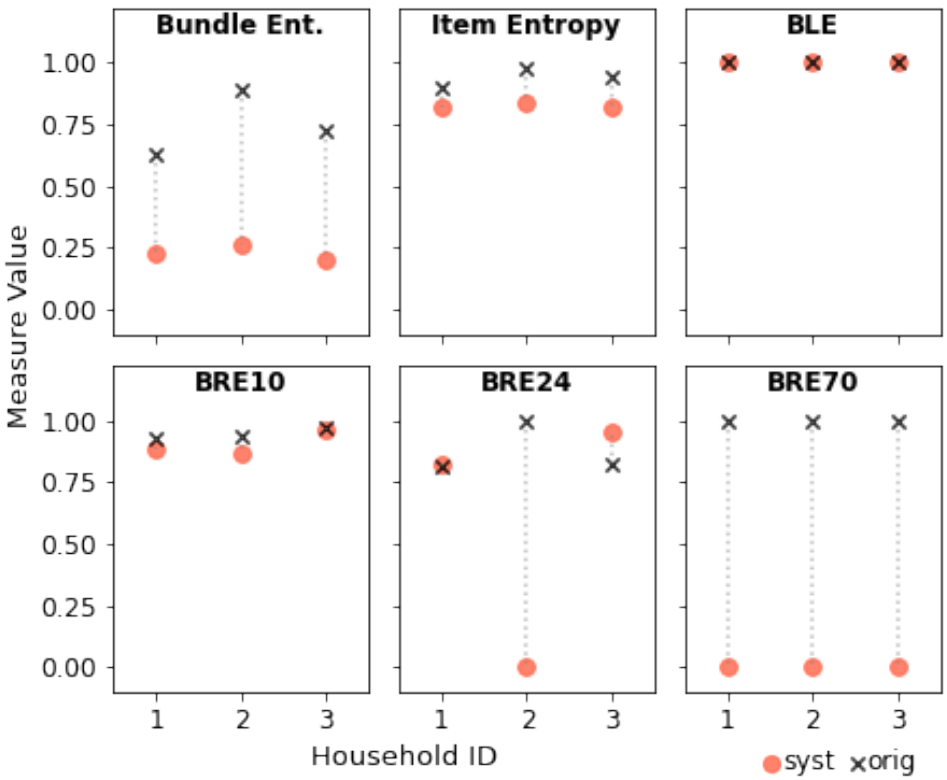


Figure 4.3: Illustrative examples of three household's scores for the evaluated measures when adding systematic bundles to the household's purchases.

dle entropy adheres to property $P2$, decreasing its score as the size of the systematic bundle added to each household's basket increases.

On the contrary, the performance of BLE, BRE10, BRE24, and BRE70 towards the addition of bundles at different size levels appears indifferent. The reason behind these performances requires further explanation. For the BRE cases, the performances can be attributed to how the *minsup* threshold affects the mining of *common sub-baskets*, particularly with the introduction of a new sub-basket component to all baskets at each point on the x-axis.

For instance, in the case of BRE70, the *minsup* threshold is high enough to prevent the discovery of sub-baskets from the original data. Thus, the sub-basket of length one is almost always mined as the *common sub-basket*

as soon as it is introduced. As a result, all baskets are represented by this sub-basket, leading to an almost entirely predictable behaviour (scores = 0.0).

However, when a *minsup* of 24% or 10% is set for BRE, other sub-baskets already considered common are extended with these new sub-basket components. The extended sub-baskets are usually longer than the original ones but have the same level of support since they exist within the same proportion of baskets. As a result, no changes are made to the symbol set, and the baskets are mapped to, with no alteration to the score computed in the subsequent BRE calculations. Similar to a BRE with a very low *minsup*, BLE is indifferent to any added systematic sub-baskets regardless of size due to its basket-level approach.

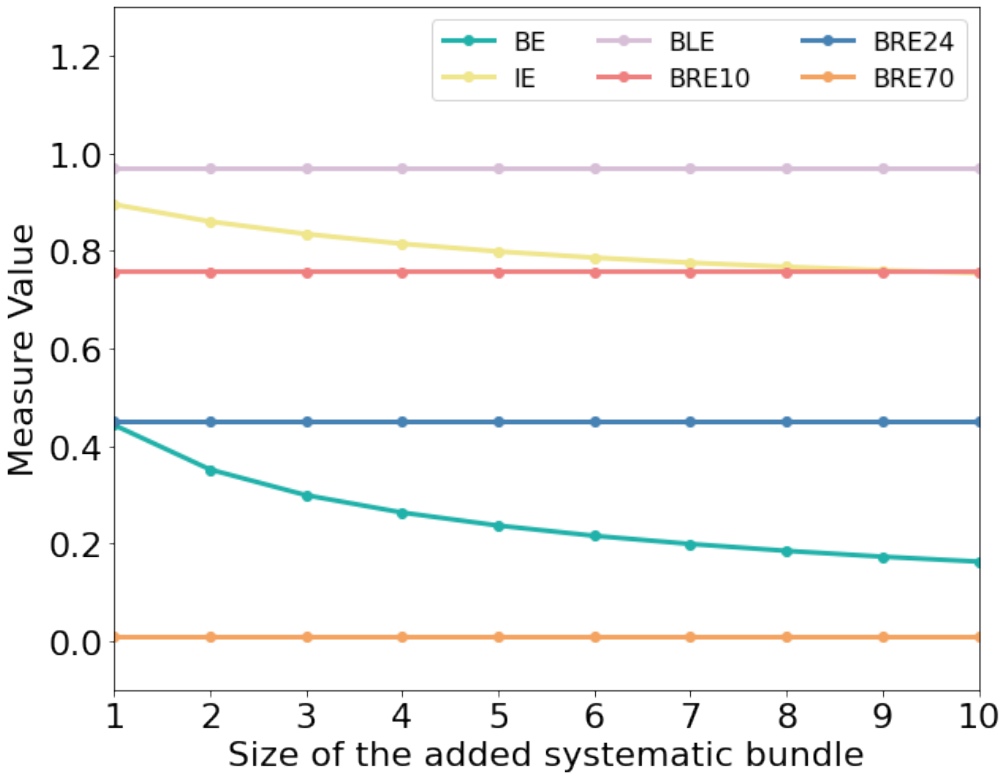


Figure 4.4: Comparing measures by increasing the size of systematic bundles added to each household's baskets. Figure abbreviations: Bundle Entropy (BE), Item Entropy (IE), Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

It is worth noting that IE is the only measure with comparable performance to *bundle entropy* throughout the entire x-axes of Figure 4.4. Nevertheless, this is due to its item-level aggregate approach, where the addition of larger systematic bundles also raises the overall probabilities of each added item, lowering the overall entropy. It is important to reiterate that, in this case, what is being conceptually measured by IE significantly differs from what *bundle entropy* is measuring.

After this rigorous evaluation of various predictability measures to meet the desired properties, the study determined that *bundle entropy* is the only suitable measure that fulfils our requirements for assessing the predictability of systematic choices (bundles) within a collection of purchases (baskets), in other words, SPB. The findings have demonstrated that *bundle entropy* provides a more comprehensive evaluation of purchasing patterns, accounting for the likelihood of specific items being purchased together and the frequency of their occurrence. This discovery enhances the feasibility of using *bundle entropy* in diverse purchase scenarios, including online retail and brick-and-mortar stores. It strengthens our rationale for employing it to assess SPB, thus supporting its application in *Study 1b*.

4.4 Study 1b: A case study of bundle entropy in mass transactional data

4.4.1 Study Design

This case study delves into the concept of *bundle entropy* and how it can be practically applied using real-world data. The study aims to showcase how *bundle entropy* can be used as a valuable measure of systematic purchase

behaviour (SPB). The study compares findings with those of the relevant parts of a previous case study from Guidotti et al. (2015). Additionally, the study compares the results of its findings with other measures used in *Study 1a*, including item entropy (IE), basket level entropy (BLE), and basket revealed entropy (BRE). To accomplish this, the study conducts the following steps for both data sets, *Dunnhumby* and the *UK Retailer* (Refer to 3.2 for data description):

- *Frequent customer selection:* To gain a deeper understanding of individuals' systematic choice combinations, this study focuses solely on customers who exhibit repeat purchasing behaviours that are representative of their overall food consumption patterns and meet minimum basket and spend criteria. This ensures that the sample size is appropriate and the data collected is reliable.

- *Data Preprocessing:* This is an essential step to ensure that the raw is converted into clean data before performing any analysis. Doing so makes the data more consistent, accurate, and reliable. To accomplish this, the study first checks for missing data and duplicates. Then, it transforms data features to the appropriate type and finally removes any outliers.

- *Data Engineering:* To replicate the relevant results from Guidotti et al. (2015) work, the study creates three features (using the cleaned data) for each individual: *Average Spend per Basket*, *Total Spend per Month*, and the *Total Number of Visits*.

- *Data Analysis:* After the previous steps, the study calculates several metrics (*bundle entropy*, IE, BLE, BRE10, BRE24, and BRE70) for each person in the data set. A distribution analysis is then conducted to determine the score group of each metric. It is important

to note that the study uses the normalised version of each metric for analysis. To compare how individuals rank in each metric based on their assigned scores, the study employs the Kendall-Tau Rank Agreement and Mean Rank Difference methods (Abdi, 2007). To ensure further comparability, the study replicates the categorisation of individuals into three segments (systematic, standard, and casual customers) from Guidotti et al. (2015) and compares the percentage of each metric that each segment represents. Finally, a Pearson Correlation analysis is conducted on the generated features of spending and visiting patterns, as described in the previous point.

The method of executing each step is described in detail below, along with how they lead to a comparison of *bundle entropy* with other measures of systematic choice combinations.

4.4.2 Methods

Frequent Customer Selection

As mentioned previously, the study only considers customers who exhibit repeat purchasing behaviours that accurately represent their overall food consumption patterns to investigate individuals' systematic choice combinations. To achieve this, the study establishes inclusion criteria to identify *frequent customers* from each data set. Only two inclusion criteria are established for both datasets to ensure consistency. Taking into account that both datasets dated between 2014 and 2016, the inclusion criteria are as follows:

- At least one shopping visit every month (for the whole period).

- Average spend per basket greater or equal to £3.

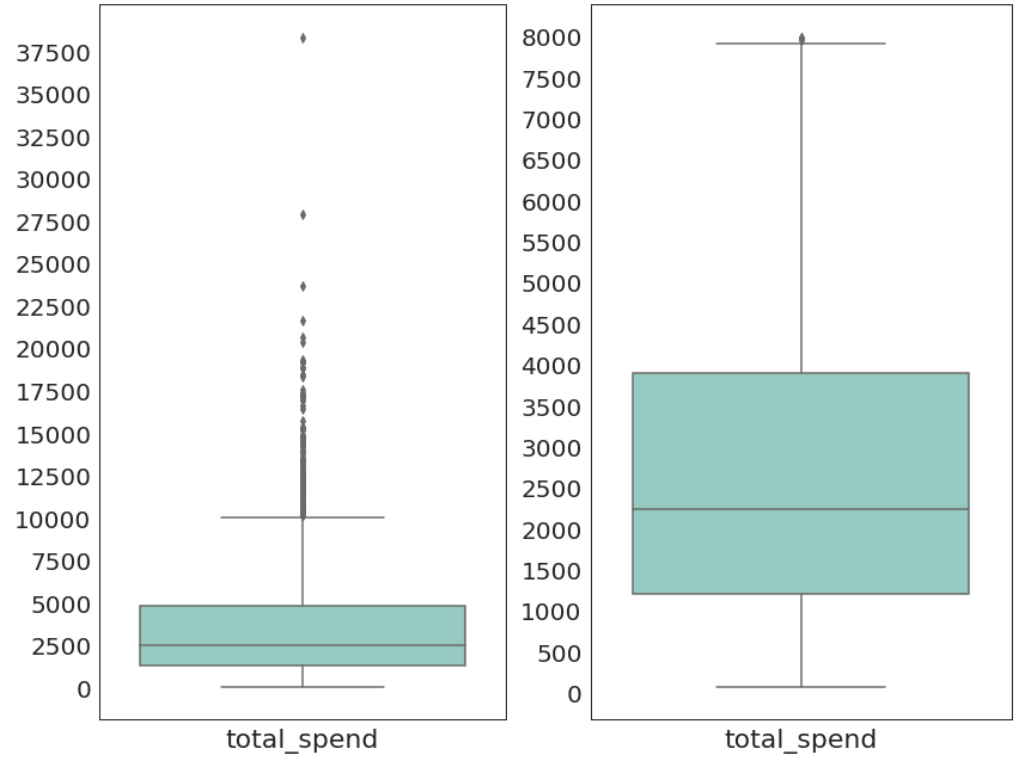
Within the *Dunnhumby* data set, a total of 2,213 *frequent customers*, out of the initial 2,500, meet the inclusion criteria, with more than 2.5 million items sold and 273.005 transactions (or number of purchases). In the *UK Retailer* data set, 2,181 meet the inclusion criteria, with 409,688 items sold and 52,102 transactions.

Data Preprocessing

After analysing both data sets, it was discovered that neither contained any missing or duplicate values. Based on data from the Office for National Statistics in the UK, the average annual spending on groceries is approximately £4,000 (Office for National Statistics, 2023), which sets a threshold of £8,000 for the two-year period covered by both data sets. A box-plot analysis was conducted to confirm this threshold, as shown in Figure 4.5. The study identified 218 individuals from the *Dunnhumby* dataset who exceeded this threshold, spending over £8,000. These individuals were removed, leaving a final sample of 1,995 individuals. No individual in the *UK Retailer* dataset exceeded the £8,000 threshold during the entire two-year period.

Feature Engineering

In order to compare the practicality of different measures, both datasets have been enriched with three new features. The first feature is called *Total Spend* and represents the total amount of money each customer spent during the entire period. The second feature, named *Average Spend per Month* is calculated by dividing the sum of each customer's *Total Spend* by



(a) With outliers (b) Without outliers

Figure 4.5: Box-plot of the *Dunnhumby* dataset before and after removing outliers.

the total number of months in the data set. Both *Total Spend* and *Average Spend per Month* use a unique identifier assigned to each customer and basket.

Lastly, the *Total Number of Visits* is the total number of transactions that each customer made during the entire period. This measure was calculated using the unique identifier associated with each transaction.

4.4.3 Results

Having shown that the proposed measure accords in theory and practice with the desired properties while other measures do not, the study now demonstrates that the selection of *bundle entropy* over the other measures will have a notable real-world impact on analysis and subsequent actions.

To demonstrate this, the study first illustrates each measure's distribution. Figure 4.6 depicts each measure distribution, where a clear difference can be seen from one another. IE, BLE, and BRE70 are the measures with the most narrow distributions, ranging between 0.9 and 1.0. BRE24 follows this with a median score of 0.9 and a range between 0.7 and 1.0. BRE24 is the only measure, aside from *bundle entropy* (BE), that has a bigger score range. However, it is heavily skewed towards the high values (above 0.8). The proposed measure (*bundle entropy*) has a wider score range, with a minimum value of 0.5 and a maximum value of 1.0. From this, it is evident that just by analysing their distributions, each measure will provide different insights and, consequently, different real-world implications.

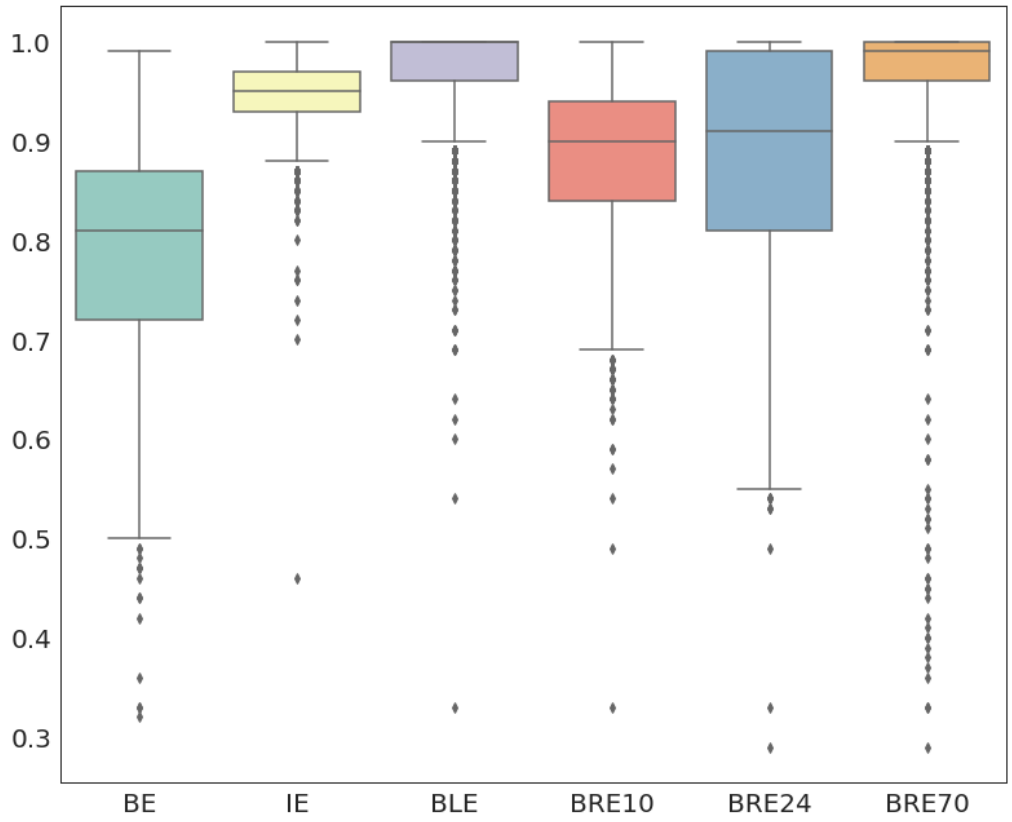


Figure 4.6: Distribution of the proposed and current measures of systematic purchase behaviour evaluated in this study. Figure abbreviations: Bundle Entropy (BE), Item Entropy (IE), Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

To further explore the differences between the measures, the study conducted a thorough analysis by comparing household and customer rankings in the previously described two data sets. To accomplish this, the study computed the previously discussed measures for all households in the *Dunnhumby* dataset and customers in the *Large UK grocery retailer* data set, resulting in a ranked list for each measure.

Afterwards, the study compared the two ranked lists for each measure in both data sets, utilising the *Kendall Tau Rank Agreement* and the *Mean Rank Difference*. The *Kendall Tau Rank Agreement* measures the difference between the probability of pairs of households or customers being in the same rank order according to both measures and the probability of them having a different rank order (Abdi, 2007). In contrast, the *Mean Rank Difference* provides a simpler indication of rank similarity. It involves matching the two lists by households or customers, taking the rank differences before computing their mean. The results of both comparisons are displayed in Figure 4.7.

The study's findings suggest that while there is some correlation between the measures, their dependence on the dataset is noticeable. As anticipated (see 4.3.3), the BRE measure with a high *minsup* (70%) is strongly associated with the BLE measure (0.91). However, despite significant agreement between the measures, notable differences exist. For example, the proposed *bundle entropy* measure shows a mean rank difference of 194 to 583 (out of 1,995 households) for *Dunnumby* and 287 to 411 (out of 2,181 customers) for the large *UK grocery retailer*.

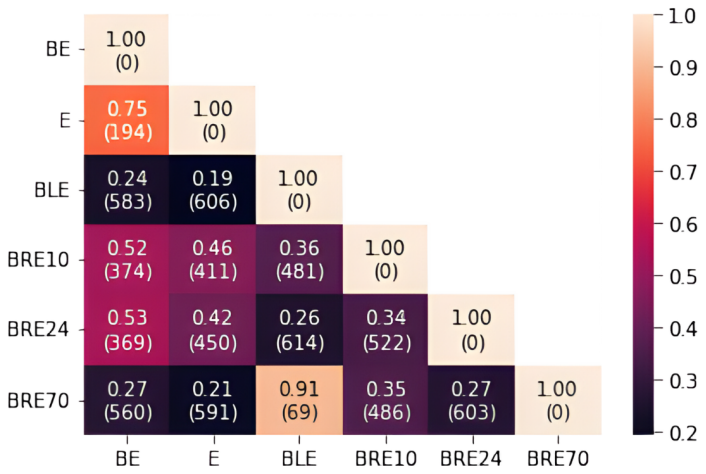
To illustrate, in the *Dunnumby* dataset, *bundle entropy* and *BRE24* show a Kendall Tau value of 0.53 and a mean rank difference of approximately 369, indicating only moderate agreement. A Kendall Tau of 0.53 suggests

that, while there is some alignment in the rankings provided by these two measures, many households are ranked differently by *bundle entropy* and *BRE24*, implying distinct dimensions of predictability are being captured by each measure. The mean rank difference of 369 further highlights this divergence, indicating that a household ranked, for example, 500th by *bundle entropy* could be ranked as high as 131th or as low as 869th by *BRE24*. This sizeable discrepancy suggests that choosing different measures could lead to different conclusions about household predictability, impacting practical decisions, such as identifying target households for loyalty programs or personalized offers.

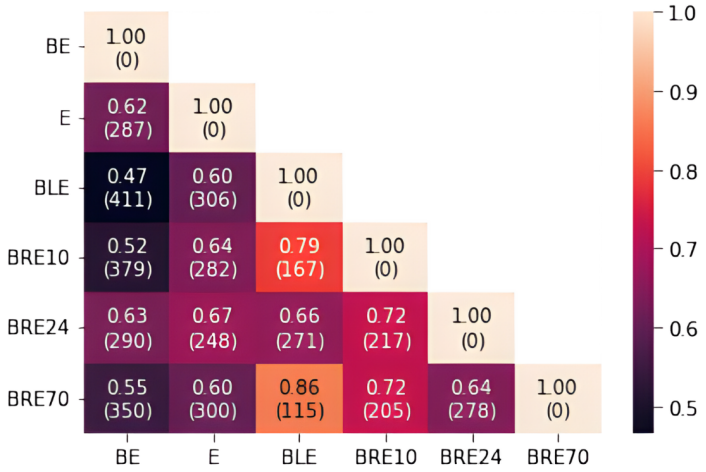
The results highlight the sensitivity of the BRE measure to the *minsup* parameter, with different choices directly influencing analytical outcomes. In contrast, the proposed *bundle entropy* measure provides clear interpretability and consistent theoretical properties. Therefore, it is crucial to consider the properties of each measure and the impact of different parameterizations when conducting any analysis based on them.

how different measures categorise customers based on their behaviour, this study follows the methodology used by Guidotti et al. (2015). The approach involves classifying customers into three categories: 'systematic', 'unsystematic', and 'standard'. Customers who fall within the lowest 10% of the distribution are labelled as 'systematic'. On the other hand, customers who fall within the top 10% of the distribution are labelled as 'unsystematic'. The remaining customers are classified as 'standard' consumers.

The study used these thresholds to compute those groups for each measure. Similarly to Guidotti et al. (2015), the study defines 'systematic' as consumers with highly predictable choice combination patterns over time, 'unsystematic' as consumers with unpredictable patterns, and 'standard'



(a) Dunnhumby



(b) Large UK grocery retailer

Figure 4.7: Kendall Tau Rank Agreement (Mean Rank Difference) of relative household/customer *predictability* for pairs of measures. Figure abbreviations: Bundle Entropy (BE), Item Entropy (IE), Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

as everyone in between.

Once each consumer is classified into one of the three groups, the study compares them by computing the percentage of shared consumers. Figure 4.8 reports the rate of shared consumers for *bundle entropy* on *Dunnhumby's* dataset (A not normalised version is provided in Appendix D.0.1). Since the middle row (standards) includes 80% of the data due to the thresholds, it is unsurprising that percentages are slightly high across some measures. The results clearly show that *bundle entropy* shares different percentages across all measures. The measure that shares the highest percentages with *bundle entropy* is IE, with 72.2% matching the customers that both classify as systematic customers, 86.9% as standard and 94.3% as unsystematic customers. In general, the measures share a bigger percentage of customers classified as unsystematic. However, the percentage of share customers classified as systematic is relatively low. Similar results were found when testing the data from the *UK grocery retailer*. See Appendix E for both normalised and non-normalised versions.

These results prove that *bundle entropy* differs from entropy, joint entropy, and especially from BRE, which measures how unpredictable a household's basket is without apparent properties to test. Instead, with bundle entropy, the study claims to measure how predictable or systematic a household's basket composition is, based on specific properties. *Bundle entropy* assigns lower scores to consumers with more systematic bundles than less or non-systematic bundles. Moreover, *bundle entropy* distinguishes between consumers with equal systematic bundles but different sizes. Therefore, it will lower the household's score with the larger systematic bundles because it provides more information for predicting their next basket. As tested in §4.3.2, the other measures do not address these key properties for measuring frequent choice combinations across baskets. BRE is the closest

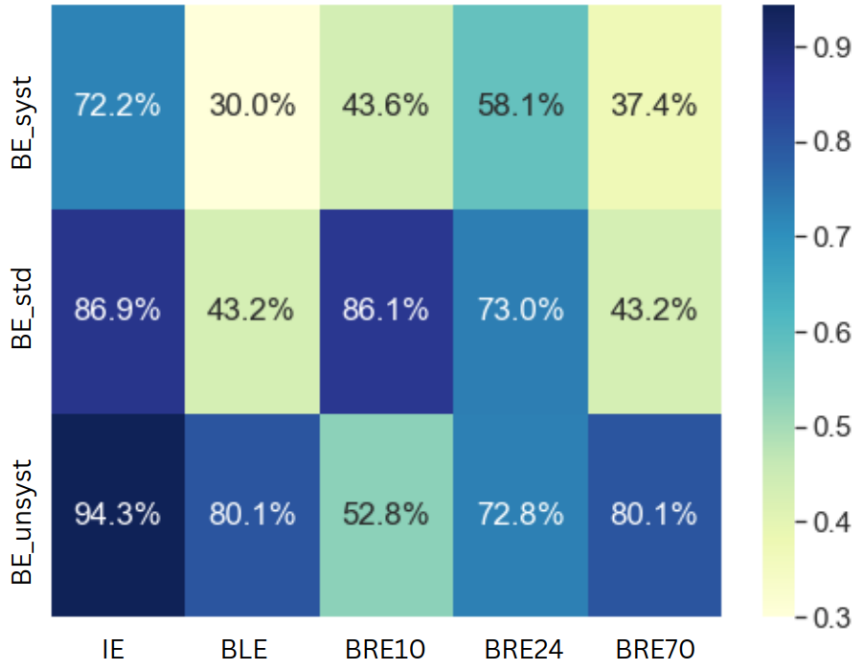


Figure 4.8: Percentage of customers share with respect to bundle entropy purchase patterns classifications (Large UK grocery retailer). Figure abbreviations: Bundle Entropy (BE), Item Entropy (IE), Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

measure to bundle entropy. However, it is parameter-dependent, and so are its results.

Figure 4.9 emphasises the previous argument by evaluating BRE24 on *Dunnhumby's* dataset (A not normalised version is provided in Appendix D.0.2). The results indicate that although BRE10, BRE24, and BRE70 are essentially the same measure with different *minimum* parameterisations, none of them has a 100% overlap in common customers for each customer classification (systematic, standard, and unsystematic). For instance, only 39.8% of systematic customers are shared between BRE24 and BRE10, and with BRE70, this number drops even further to 32.7%. As for unsystematic customers, BRE24 shares 64% with BRE10 and 50.3% with BRE70. While the percentages of unsystematic customers are higher than those of systematic customers, they are still relatively low, given that the

measures are the same.

This indicates, once more, that different results can be expected depending on the BRE parametrization. In other words, this can potentially lead to classifying consumers into different groups depending on the parameter selected. It is important to highlight these differences between the measures, to not expose researchers and practitioners to approaches that do not match what they measured, but more importantly, to clarify differences with our approach.

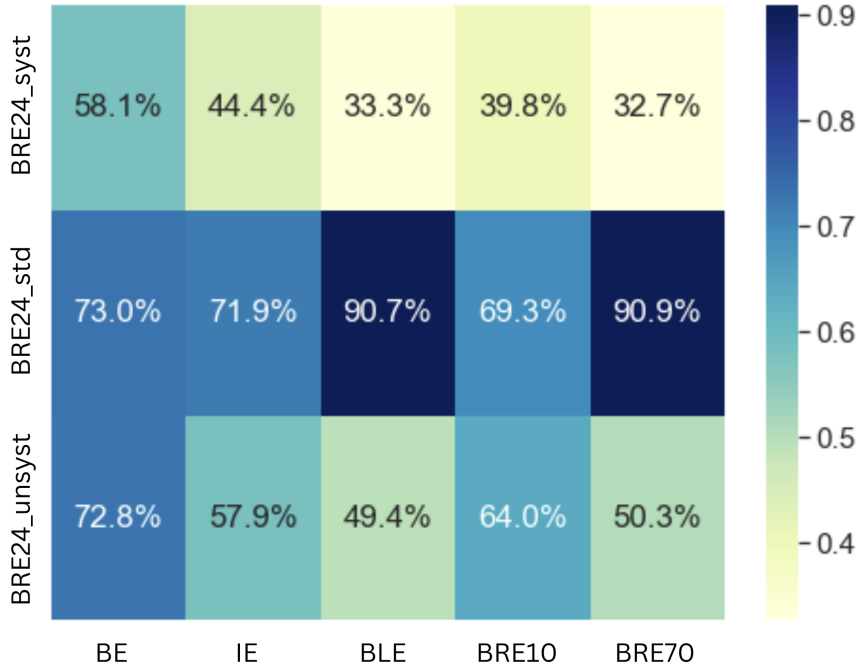


Figure 4.9: Percentage of customers share with respect to BRE24 purchase patterns classifications (Dunnhumby). Figure abbreviations: Bundle Entropy (BE), Item Entropy (IE), Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

Finally, the study evaluates the practical value of the proposed measure. The measure's usefulness in gauging the predictability or consistency of an individual's baskets can be categorised into two primary groups.

The first group refers to *bundle entropy* application as an explanatory variable for describing an individual or segment. When an individual's *bundle*

entropy score is computed and found to be low, it indicates that their purchase behaviour is highly predictable. This means that retailers can use this information to tailor their communication strategies and improve their marketing efforts. For this to be effective, the measure must align with the practitioner's intuition and understanding, which is why the study has examined it in the previous sections.

The second analysis involves utilising specific measures to establish a correlation between the measure, potentially in conjunction with other measures, and an output variable within a predictive framework. This approach may be motivated by commercial interests or other factors, such as promoting social good or enhancing consumer welfare. This type of application is akin to the one examined in Guidotti et al. (2015), which investigated the link between systematic customer behaviour and profitability in the supermarket retail sector. This study replicates the assessment of Guidotti et al. (2015) proposed BRE measure, excluding the evaluation related to the complementary Spatio-temporal measure that could be used in conjunction with *bundle entropy*, as presented in this study.

In the study conducted by Guidotti et al. (2015), the relationship between a customer's shopping behaviour predictability and a supermarket's profitability was explored. Two variables, specifically the *average spend per visit* and the *number of visits*, were analysed independently against the predictability measure (BRE) using a single-variable linear regression model for each case. The authors' findings revealed a negative correlation (-0.3253 and -0.3249) between customer predictability and profitability. This suggests that systematic and predictable customers tend to spend more per visit, making them more profitable.

To showcase the effectiveness of the *bundle entropy* measure, this study

compared it to the BRE, IE and BLE measures by computing the Pearson Correlation for all measures, along with the *average spend per visit* and the *number of visits* for the *Dunnhumby* and *UK grocery retail* data sets. Additionally, this study examined the measures' correlation with the *average monthly spend*, which indicates a customer's potential lifetime value. Table 4.3 displays the results of their analysis.

The findings of this analysis support the conclusions presented by Guidotti et al. (2015). The results indicate that there exists a negative correlation between the *number of visits* and the *average basket spend* for both data sets. Furthermore, depending on the parameterisation, the analysis also reveals either positive or negative relationships between the parameterised BRE measure for *mean basket spend* and the *number of visits* for *Dunnhumby's* data set. This highlights the fact that the interpretation of the BRE measure is heavily dependent on the *minsup* threshold, which can be a challenging parameter to set if there is no previous knowledge of the data set.

Moreover, this study shows that item-level and basket-level entropy measurements are two distinct concepts. This study founds non-significant correlations between the *mean basket spend* and item-level entropy measurements on both datasets and only a slight negative correlation for BLE for the *Dunnhumby* data set. This relationship flips to a slight negative correlation for the second data set.

Notably, the various parameterizations all demonstrate a negative correlation with the *mean monthly spending*. However, when using the BRE parameterization, there were inconsistencies in how the relationship between the two variables was affected across different data sets. Nevertheless, the data provides evidence to suggest that the *mean spend per month*, which is

a proxy indicator of lifetime value, is generally linked to the predictability of baskets and items. This suggests that information is being shared at multiple levels, both within and between baskets and items, which could be of great value for future studies.

The use of two distinct datasets in this study enhances the reliability and generalizability of the BE measure across different consumer segments. One dataset is obtained from *Dunnhumby*, while the other is from a prominent *UK grocery retailer*. Each dataset embodies unique transactional characteristics, market positioning, customer demographics, and purchasing behaviours. The disparities between these datasets suggest that different types of purchasing behaviour exist, underscoring the necessity for measures that are clearly interpretable and founded on stable theoretical principles.

The results for BE are consistent across both datasets, revealing significant negative correlations with spending patterns. While item and basket-level entropy measures exhibit limited or inconsistent relationships, BE remains robust, showing stable relative magnitudes across the datasets. For instance, the BRE70 metric demonstrates a negative and statistically significant relationship with *mean basket spend* in the *Dunnhumby* dataset, whereas, in the second dataset, it shows a positive and non-significant correlation.

Overall, the integration of two datasets validates BE as a reliable and adaptable metric capable of capturing systematic behaviours across various retail contexts. This approach not only emphasizes the robustness of BE but also supports its broad applicability in consumer behaviour research, providing a solid foundation for future studies in diverse retail settings.

Table 4.3: Pearson Correlation between the measures and spending and visiting factors. * denotes statistical significance. $p-values$ were adjusted using the Benjamini–Hochberg false discovery procedure with a p-value of 0.05 Benjamini and Hochberg (1995). Table abbreviations: Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

	Correlation with:					
	Dunnhumby			**UK grocery retailer**		
	Mean Basket Spend	Mean Spend per Month	Number of Visits	Mean Basket Spend	Mean Spend per Month	Number of Visits
Bundle Entropy	-0.187*	-0.475*	-0.374*	-0.215*	-0.401*	-0.371*
BRE 10%	0.009	-0.494*	-0.439*	0.066*	-0.226*	-0.365*
BRE 24%	-0.290*	-0.340*	0.108*	-0.001	-0.262*	-0.342*
BRE 70%	-0.134*	-0.268*	-0.182*	0.002	-0.152*	-0.223*
Item Entropy	0.027	-0.380*	-0.456*	0.000	-0.236*	-0.315*
BLE	-0.034*	-0.268*	-0.323*	0.088*	-0.138*	-0.270*

4.5 Discussion and Conclusion

This study investigated the currently available measures for SPB in the retail context. The study proposed a new measure called *bundle entropy*, which adapts the binary similarity function from joint entropy to a non-binary approach. The study compared the proposed measure against the other pertinent measures, showing how different and more actionable *bundle entropy* is in various purchase scenarios. The study confirmed Guidotti et al. (2015) findings regarding the spending patterns but not the conclusions about visiting patterns between household groups since results can flip depending on the normalisation approach and the parameterization set-up.

The study also provided a theoretical definition of the proposed measure and its properties. This enables the formulation of specific questions/-tasks in the retail context. Furthermore, *bundle entropy* was empirically tested on both synthetic and massive real-world purchase scenarios, showing the measure's unique ability to act according to the desired properties to measure basket composition's predictability in a robust way. Overall, the results demonstrated that (1) the proposed measure accords, in theory and practice, with the desired properties while the others do not, (2) the measures are notably different and should not be used interchangeably, and (3) the proposed measure has higher utility in practice, providing a consistent, parameter-less measure that accords to well-defined, intuitive properties allowing practitioners to efficiently and correctly interpret and action the outcome of analytics and insights based on the measure.

4.5.1 Practical implications

From a retailer's point of view, *bundle entropy* can be utilised in various ways, expanding its applications beyond consumer behaviour analysis. For instance, it can serve as a tool for optimising inventory management, helping retailers understand the diversity of product bundles in demand, thereby enhancing stock levels and reducing wastage. Additionally, *bundle entropy* could be integrated into dynamic pricing strategies, where products with higher entropy may be priced differently based on their bundling patterns, offering potential profit optimisation opportunities (Saberi et al., 2019).

Further exploration could involve examining the impact of *bundle entropy* on customer loyalty and lifetime value, providing insights into long-term consumer relationships. Moreover, considering the rise of e-commerce and personalised marketing, *bundle entropy* might be employed to enhance recommendation algorithms, tailoring product suggestions based on individual customers' bundling preferences and consistency over time.

The versatility of *bundle entropy* extends its potential applications to various fields, contributing to a broader understanding of complex systems. In the realm of supply chain management, *bundle entropy* can be harnessed to optimise logistics and distribution networks. By identifying patterns of bundled products, retailers can streamline their supply chains, reducing transportation costs and minimising environmental impact (Saberi et al., 2019).

Furthermore, in the context of data science and machine learning, *bundle entropy* may find applications beyond retail analytics. Its principles can be adapted to enhance anomaly detection systems, where deviations from

typical patterns could signal potential issues or opportunities. Integrating *bundle entropy* into anomaly detection models may improve the identification of irregularities in diverse data sets, ranging from cybersecurity to healthcare, thereby enhancing the overall robustness and reliability of anomaly detection frameworks (Chandola et al., 2009).

This interdisciplinary approach highlights the measure's adaptability across different fields, emphasising its significance in advancing knowledge and innovation beyond the scope of traditional retail research.

4.6 Subsequent Studies

This study contributes to the field of consumer analytics within the retail domain. It specifically examines SPB by analysing the predictability of a customer's shopping basket composition and introduces a new measure called *bundle entropy*. The study highlights the importance and reliability of this measure in comparison to existing ones. It also highlights the consequences of biases within the existing measures, which can range from inaccurate customer segmentation to inefficient marketing strategies across different channels. Nevertheless, the study also acknowledges certain limitations inherent to the analysis, which will be discussed in section 7.3 at the end of this thesis.

The outcomes of the current study have set the ground for the following studies, which aim to delve deeper into the practical usage, applications and extensions of *bundle entropy* within the dynamics of consumer purchase behaviour. The current study also examines the correlation between *bundle entropy* and output variables, such as the amount spent per basket and the total number of visits. However, these variables are transactional

in nature. Therefore, they just reflect common transactional patterns associated with monetary value and frequency of purchases. This suggests that subsequent studies should explore the relationship between *bundle entropy* and other types of consumption patterns, such as healthy choices. This will provide a more comprehensive understanding of its usability in various scenarios, which will be explored in *Study 2*. More specifically, the subsequent study aims to address specific research questions regarding customer spending, systematic buying patterns and healthy choices across different retail settings.

Chapter 5

Consumers' Systematic Purchase Behaviour and Healthy Choices

This chapter is based on work currently under review (2nd stage) to the Journal of Business Research in 2023:

Mansilla, R., Smith, G., Smith, A., and Goulding, J. (2023). Systematic purchase behaviour and healthy choices across online and offline channels: Insights from transactional data. *Journal of Business Research.*

Contents

5.1 Study 2: A case study of bundle entropy across retail channels using mass transactional data

5.1.1 Introduction

This study aims to contribute to the understanding of online versus offline purchases through an empirical application of a propensity measure known as *bundle entropy* (Mansilla et al., 2022) on real-world purchase data. *Study 1* introduced this metric, which evaluates Systematic Purchase Behaviour (SPB), a term originally coined by Guidotti et al. (2015). However, its pragmatic value and integration with other metrics to offer valuable and actionable insights have not been thoroughly tested on real-world transactional data.

To maximise the impact of *bundle entropy* application, the study explores the differences in SPB in online and offline purchase sets for given households (within-subject) in terms of the overall variance in products bought, as well as at the product category level, and also explores one category (soft drinks) in detail. In addition to this, the study investigates the relationship between household-level SPB propensities and the health/nutritional score of soft drinks in order to establish if there is a significant variation in the choice outcomes across channels (online versus offline). Since there is evidence that purchasing and consumption habits are closely linked, they can be used to estimate an individual's diet and nutrient intake reliably (e.g. Eyles et al. 2010, Appelhans et al. 2017, Vepsäläinen et al. 2022). The aim is to explore locked-in habits in food consumption across different retail channels from the consumers' purchase history, which can provide insight

into targeted nutritional interventions to reduce bad eating habits, such as overweight and obesity, considered global public health priorities for more than a decade (Organization and others, 2017; Organization, 2000). For instance, in the United Kingdom (UK), 64% of adults are overweight or obese (Conolly et al., 2019), while 9.9% of children (aged 4-5 years) are obese. These numbers are alarming since they increase yearly, both in the UK (Apperley et al., 2022) and worldwide (World Health Organization, 2020).

The research utilised a representative sample of 2,181 households to examine their online and offline purchasing habits. The sample consisted of 62,403 online transactions and 166,085 offline transactions (each representing a unique basket) that occurred two years before the outbreak of COVID-19 (2014-2016). This extensive dataset enabled a direct comparison among households, which is uncommon at this scale and breadth of data. Furthermore, the reach of the data is national (United Kingdom – UK).

5.1.2 Background

As mentioned before, SPB is a term originally introduced by Guidotti et al. (2015). To illustrate what is meant by SPB at a very basic level, consider the following hypothetical purchase sequences. *Person A* buys *a* and *b* together every time, which indicates an entirely systematic buying behaviour. *Person B* buys products *a*, *b*, and *c* in combination, then *a* and *b*, and next time *a* and *c*. *Person C* buys *a* and *c*, then *d*, then *b*, *c*, and *d*. How systematic are *Person B* and *Person C*? Assessment of probability at the individual item level or the repeat purchase rate for *a, b, c, or d* is straightforward. However, assessment of the evolution of the combinations is more

tricky and best assessed by measures based on joint entropy. Mansilla et al. (2022) make the comprehensive technical case for a joint entropy-based measure; the logic for this is also raised and revisited more in detail in section 5.1.5 (Measuring Systematic Purchase Behaviour). It is essential to state that choice set stability prediction (via *bundle entropy*) serves a different purpose than basket analysis based on association rule mining (e.g. Kaur and Kang (2016)). Basket analysis seeks to identify frequently co-purchased items, while *bundle entropy* predicts the stability of specific product combinations over time. Nonetheless, *bundle entropy* provides a valuable link between basket analysis and investigations of purchases over time.

There have been many attempts to understand and predict why consumers stick with certain products (known as 'inertia' or 'loyalty') and why they switch to others (known as 'variance', 'churn', or 'variety seeking'). These attempts are diverse and use different measures and models. However, they are all part of a global effort to understand how consumers make choices over time, both within and outside retail settings. Basic indicators like repeat purchase rates or simple probability models (e.g. Uncles and Hammond (1995)) can provide some insight into inertia and variance. However, many are not well equipped to assess if a household or individual's purchase is systematic at the basket level over a specific period of; they are not designed to assess systematic choice in terms of products bought together.

Guidotti et al. (2015) assess SPB via an entropy measure of purchase over time at the basket level; they term this basket revealed entropy (BRE). The term and the concept are distinct from predictability, which has a universal meaning but is too disparate to be entirely appropriate for the problem at hand. Mansilla et al. (2022) provide a better alternative to BRE in the form of *bundle entropy*.

Bundle entropy is capable of measuring variance in single item sequences (e.g. grocery store patronage or single items bought one at a time), the variance at the basket level, and at the sub-basket level over time – they term this the bundle level (e.g. a given category or topic within the basket). Many fast-moving products are bought in bundles, for example, fresh fruit and vegetables or soft drinks. Multi-item bundles are products in a given merchandise category or 'topic' bought together in a single purchase episode (basket or mission). For example, different beers, soft drinks, fruit, vegetables, or snacks bought during an offline or online grocery mission. The term 'bundle' entropy emphasises that the measure can assess the systematic nature of product combinations at the basket, sub-basket, or individual item level (a 'bundle' of 1 item). *bundle entropy* provides a single value measure, which can be easily compared with other co-variates, such as nutritional scores.

This study is organised as follows: First, it evaluates the technical and conceptual structure, pertinent literature, and research inquiries. Subsequently, it details the data set, sample selection, and broader methodology. Finally, the study concludes with a review of the significant results before moving on to a discussion and conclusion section.

5.1.3 Conceptual Framework and Research Questions

Many factors can drive consumers' purchase behaviour. Among these, environmental factors (e.g. store layout, structure and appearance) can be as relevant as the stability of the purchase context (Koll and Plank, 2022) or consumers' inherent motivations and characteristics (Seetharaman and Chintagunta, 1998). For example, online and offline store environments usually fulfil the same final consumers' purpose but in different ways. Each

channel provides unique characteristics that can influence how consumers interact and behave (Ratchford et al., 2022; Campo et al., 2021). For instance, the offline channel has several benefits, including real-time product evaluation, instant gratification, and store experience (Cimana, 2013; Morrison et al., 2011). It allows customers to touch and feel products, try them out, and get immediate feedback from sales associates. However, shopping offline also has drawbacks. It can be time-consuming and expensive due to commuting costs. Moreover, customers may have to deal with crowds of people, limited space to carry products, and out-of-stock shelves, which can be frustrating (Grewal et al., 2004a).

On the other hand, online shopping offers the convenience of easy product search and comparability (Jadhav and Khanna, 2016). Customers can find and compare products from the comfort of their own homes, and online shopping allows for greater time flexibility. Delivery services make it possible for customers to receive their products at their doorstep, thereby saving time (Huang and Oppewal, 2006). However, online shopping has its downsides, such as the inability to physically inspect the products before purchasing (Moshrefjavadi et al., 2012). Additionally, customers often have to pay high prices for delivery services, which can be a disadvantage when resources are limited (Gil et al., 2020; Xu et al., 2022). Finally, returning a product can be frustrating, no matter the reason. Whether the customer received a faulty item or changed his mind, the process can be tedious and time-consuming.

The distinctions in these channels' propositions/contexts and how customers perceive them (Wang et al., 2016) may influence consumer purchase behaviour (Hult et al., 2019). Moreover, the channel will likely affect consumers' cognition and decision-making when purchasing online versus offline. Smith et al. (2021) offers a novel theoretical framework for the po-

tential effect of retail channel blends on cognition, spending, and product choices. However, this framework remains untested. Other extant research can help us to generate various potential antecedents of differential choice online, but these often eschew the outcomes since they do not look at actual purchase data (André et al., 2018; Schneider and Leyer, 2019; Candrian and Scherer, 2022; Puntoni et al., 2021).

Online purchase is self-evidently different from the various stimuli experienced in a physical store. Some research has suggested explanations for the differential outcomes (e.g. Huyghe et al. 2017, Anesbury et al. 2016). Still, these insights are a) based on the analysis of online decision-making only and are not comparative (e.g. within-subject/same household paired as per the transactional analysis by Chu et al.2010); b) they avoid the core effects of the cognitive and information search differentials for the same decision-maker/household online and offline. Other research is quite dated and relates to consumer durable/high involvement purchases (e.g. Kulkarni et al.2012); the data for that paper was collected in 1999.

The moderating role of retail venues spending behaviour

Numerous studies have examined customers' purchasing behaviours across online and offline venues and their significance to management and marketing outcomes (Degeratu et al., 2000; Shankar et al., 2003; Danaher et al., 2003; Andrews and Currim, 2004). Their overall conclusion is that consumers have different spending behaviours between in-store and online channels. However, opinions are divided regarding which venue (online or offline) encourages or discourages specific behaviours, such as price sensitivity or the average amount spent per basket. For instance, households are found to be more price-sensitive online than offline. This, according

to Degeratu et al. (2000), is due to the online channel's ability to readily accentuate and advertise product discounts and promotions, making consumers more likely to take advantage of those deals to save money.

There is some dated evidence that online prices are lower than in-store prices (Brynjolfsson and Smith, 2000); however, the retailer in this study sources products from local stores for online purchase at the same price. Even if differences occur in time and convenience, it can mean consumers are more willing to pay higher prices (Putrevu and Ratchford, 1997), especially on some product categories (Chintagunta et al., 2003; Donnelly et al., 2021). As a result, under certain conditions, consumers may be less price-sensitive online than offline. This idea is supported by Andrews and Currim (2004), who, using consumers' grocery footprint, conclude that on specific products category, most online consumers spend more than offline shoppers. So, there is no clear consensus regarding which venue funnels more or less price sensitivity.

The above-mentioned papers did not explore differences in consumer behaviour across various channels. Only a limited number of studies have compared online and offline purchase behaviours of the same household (within-subject) (Chu et al., 2008, 2010; Pozzi, 2012). Such studies have found that the same individual is less price-sensitive online than offline. Therefore, individuals are willing to pay more money for products or services online as compared to physical stores. For example, Chu et al. (2008) utilised grocery panel records from a sample of households who purchased through both store channels (online and offline) and found that households were consistently less price-sensitive across all products within a particular product category, frozen pizza. Chu et al. (2010) expanded his previous study by examining the moderating effect of the level of household engagement online versus offline and found that price sensitivity increases

as online engagement grows. Pozzi (2012) attributes the differences in purchasing behaviour to the convenience of online features such as saved baskets and product recommendations. These findings indicate that household decision-makers may demonstrate varying spending habits in online versus offline channels, particularly concerning brands, sizes, and pricing. Such discrepancies underscore the possibility of distinct spending patterns for food and non-food items.

The existing studies have provided some insights into the comparison of online and offline grocery spending behaviour. However, there is a lack of research that explores the impact of retail channels on customer/household purchasing behaviour across different product categories. To address this gap, we aim to conduct a study that uses two years of transactional records with national coverage. Our research questions (RQs) are intended to guide our investigation into the moderating effect of retail channels on purchasing behaviours. The RQs are as follows:

RQ 1a: Do retail channels (online vs. offline) influence consumers' spending behaviour at a basket level?

RQ 1b: Do retail channels (online vs. offline) influence consumers' spending behaviour at a product category level?

These preliminary questions have been formulated to obtain timely verification of the papers discussed earlier with a more advanced and recent dataset. Notably, even comparable research studies like the one conducted by Chu et al. (2010) have become outdated. Furthermore, although their sample size was comparable, it is worth noting that the data they analysed

was limited to a single year, specifically, the period between 2002 and 2003. The subsequent research questions are designed to be exploratory in nature and are intended to delve deeper into spending behaviours.

The moderating role of retail channels in within-subject systematic purchase behaviour

While substantial studies focus on understanding, explaining, or even predicting inertia, such as repeat purchases and brand and size loyalty (e.g. Ehrenberg 1988, McDonald 1993, Klassen and Glynn 1992, Sharp et al. 2012), or on the other hand, variance, such as switching and derived variety-seeking (e.g. Givon 1984, Chang 2011, Punj 2011), little work exists investigating store channels' effects on consumers' systematic behaviour.

Danaher et al. (2003) work was one of the first empirical studies exploring online and offline consumers' purchase behaviour (expressed by brand loyalty) utilising grocery data, concluding that brand loyalty is related to brand share for online consumers but not for offline consumers. Similarly, Andrews and Currim (2004) explored the behavioural preferences between online and offline consumers regarding product sizes (size loyalty) and discovered that online users do less screening regarding product sizes. Hence, they have a more systematic/predictable product size choice. On the other hand, offline consumers tend to vary among different available formats.

A later study by Chu et al. (2010) explored brand and size loyalty for the same household across channels. Using historical grocery panel data, they found that the same household is more brand and size loyal online than in-store and that household characteristics do not affect or have a minimum influence on these behaviours. However, these behaviours are unstable and might vary depending on the consumers' level of engagement and maturity,

particularly in the online venue (Chu et al., 2010).

These disparities in online and offline behaviours (for the same households) could be influenced by several factors, including the consumers' perception of how crowded physical stores are (Aydinli et al., 2021), or what other customers might think about their product choices (Ratner and Kahn, 2002).

According to the research conducted by Smith et al. (2021), when it comes to online shopping, consumers tend to be influenced by various factors that may guide their decision-making. These factors include saved grocery lists, personalised nudges, and push notifications based on past behaviour. The study suggests that these features tend to 'funnel' the consumer's choice, encouraging them to make more systematic and predictable purchases. As a result, this could reduce the likelihood of making spontaneous and entropic choices.

Taken together, these studies suggest a relationship between retail channels and consumers' SPB. However, previous studies rely on traditional measures and not one specifically designed to measure SPB. Therefore, we state the following research questions (see Figure 5.1):

RQ 2a: Do retail channels (online vs. offline) influence consumers' systematic purchase behaviour at a basket level?

RQ 2b: Do retail channels (online vs. offline) influence consumers' systematic purchase behaviour at a product category level?

Systematic purchasing behaviour and its relationship to healthy product purchases across channels

Some explanatory models (e.g. McAlister and Pessemier 1982 and Kahn et al. 1986) that have explored product variety/uncertainty have taken into consideration products' attributes because they influenced consumers' food choice behaviour. The product's brand is one of the crucial factors that can influence a consumer's decision to purchase a product (Givon, 1984). However, measuring the brand's impact is not easy as it is subjective and can only be evaluated as a weighted value based on factors such as brand leadership and brand loyalty. Additionally, it is not accurate to assume that people will switch between products based solely on the market share that the product commands.

The market position is a result of several factors, and it is not necessarily the primary determinant of a consumer's choice (Danaher et al., 2003). There are only a few measures that can be found within transaction data that are directional, quantifiable, or ordinal and that have the potential to impact consumer choice. One such measure is the product's price. However, the relationship between the price and the choice outcome is well-researched, and the potential for a salient contribution is arguably limited. Therefore, marketers need to explore other measures that can influence consumer choice and tailor their marketing strategies accordingly.

As more people choose to change their diet to a healthier one, it's become increasingly important to understand how to encourage consumers to opt for healthier products. Recent studies have revealed that external stimuli, such as sales and promotions, can have a significant impact on consumers' decision-making, causing them to make unhealthy choices (Yan et al., 2017). This is a common occurrence when retailers inundate cus-

tomers with various online and offline discounts and offers. However, the research has also discovered a less apparent factor that affects consumers' purchasing decisions, which is the availability of different payment options across various channels. A study conducted by Thomas et al. (2011) found that consumers who use credit cards as their payment option tend to lean towards unhealthier product options.

More recent research by Yang et al. (2022) has uncovered an interesting link between the time of day and the choices we make when it comes to food. The findings show that as the day wears on, our ability to regulate our own behaviour declines, leading to an increase in unhealthy food choices. In other words, it becomes harder for us to resist the temptation of unhealthy options as the day progresses. This finding has significant implications for individuals looking to maintain a healthy diet, suggesting that being mindful of the time of day when making food choices may support healthy eating habits.

Other research suggests that products' nutritional characteristics, which include sugar, sodium, and saturated fat content, also play a crucial role in consumers' purchasing decisions (Berger et al., 2007; Trivedi et al., 2016). Studies have been conducted to gain a better understanding of the impact of how products' nutritional information is displayed, concluding that clearly highlighting the health benefits of a product is a crucial driver of consumer choice towards healthier options. (Lobstein and Davies, 2009; Nikolova and Inman, 2015). This idea has inspired many public strategies worldwide to reduce unhealthy consumption. Front-of-pack product labelling is one of the main strategies with different implementations depending on the country. Some countries have focused on calorie knowledge, while others have focused on saturated fat or sugar awareness, or a combination of them (e.g. Van Kleef et al. 2008, Santana et al. 2022).

All these regulatory initiatives are part of a larger worldwide campaign to reduce calorie consumption and related illnesses. Research has shown that some of these regulations (e.g. product labelling) generally reduce the consumption of unhealthy products (Downs et al., 2017). Yet, these studies have only examined the impact of these policies on consumers' behaviours in physical stores, not online, where some evidence has shown that it discourages the purchases of unhealthy or vicious products (Huyghe et al., 2017). This online behaviour could be due to many factors, such as the virtual representation of the products, as opposed to a palpable experience, decreasing the desire for immediate gratification (Huyghe et al., 2017).

The work from Zatz et al. (2021) is a significant addition to the literature, providing the first empirical investigation into the correlation between unhealthy products and retail channels for households using transactional data. While the study concludes that households spend less on certain unhealthy products online than offline, it is important to note that the sample size is small, and the purchase inclusion criteria are not very strong. Therefore, we require more comprehensive studies that examine healthy and unhealthy purchase behaviour across retail channels over a longer time period using transactional records with nationwide coverage, as proposed in this study. This can help us gain a better understanding of the factors that influence household purchase decisions and can inform public health interventions aimed at promoting healthier consumption patterns.

The health and welfare of consumers are critical concerns for retailers who seek to understand consumer choice for commercial insight and corporate social responsibility, as highlighted by Grewal and Levy (2007). Retailers can leverage their understanding of consumer behaviour to inform public health interventions to promote healthier consumption patterns, thereby

contributing to societal well-being.

This study asserts that a nutrition-based score is an objective and measurable product attribute. Moreover, it can readily be calculated from the product's nutritional score (as many retailers apply) using a simple five-point scale derived from the retailer score (the rationale is shown below; see Table 5.2). This is likely more consistent with a consumer's appraisal or choice heuristic (very healthy, healthy, neutral/mixed, unhealthy, very unhealthy). The nutrition-based score is applied to the soft drinks market due to several factors:

- The importance of this category with respect to obesity, diabetes and other diseases (Bray et al., 2004; Basu et al., 2013),
- The clear labelling, such as low sugar, light, and diet, signalled conspicuously via brands, product signifiers, and packaging, which ensures customers are likely to be more aware of the differences in the products when they choose.
- The practical ability to obtain nutrition information (from reliable sources) for all products in the category.
- The prevalence of purchase across both channels.

Soft drinks and hydration products have been the focus of significant concern regarding media coverage, public health policy, and campaigns (Basu et al., 2013; Bray et al., 2004). Crucially, they are also often bought in heterogeneous bundles within various shopping mission (basket) forms, particularly in main weekly grocery baskets for a given household (the dataset and analysis confirm this assertion). As such, soft drinks are a good example for investigating consumers' systematic and unsystematic purchase behaviour across retail venues.

It is essential to highlight that the work acknowledges upfront that nutritional value will not necessarily be an overriding driver of choice. Two scenarios that could manifest within one shopping mission (basket) are:

1. A product is bought regardless of nutritional and health implications.
2. A product is bought based on a blend of preference drivers (e.g. brand, flavour, etc.) in which nutritional properties are significant or even salient (e.g. diet drinks or drinks with conspicuous vitamin content).

So, the nutritional score may give insights into antecedents of choice, or it may simply be an arbitrary 'outcome' of choice. Either way, nutritional/health value and the effect on consumer welfare remain crucial concerns. Moreover, the differential effects on the systematic purchase behaviour of a consumer (specifically bundle entropy) have not been investigated. This motivates the exploration of how the overall *health score* of the mix of products purchased varies across channels, as well as the link between the systematic purchase behaviour of consumers and the healthiness of their choices. Hence, the subsequent research question shown in Figure 5.1 are generated:

RQ 3: Do retail channels (online vs. offline) influence consumers' healthy product choices at a product category level?

RQ 4a: What is the relationship between consumers' systematic purchase behaviour and healthy product choices?

RQ 4b: Do retail channels (online vs. offline) influence the relationship between consumers' systematic purchase behaviour and healthy product choices?

This study's dataset and approach allow us to explore the relationship stated in RQ 4a and 4b further. Smith (2019) provides four scenarios of welfare effects from 'choice automation' (the author uses the term exogenous cognition) that can be adapted to structure the possible effect of online purchase on systematic purchase behaviour: a) online purchase decreases entropy and therefore funnels/reinforces behaviour that is *'healthy'*; b) online purchase decreases entropy, and therefore funnels/reinforces behaviour that is *'unhealthy'*. In other words, online purchase reinforces existing behaviour. This is based on some of the findings reported above and the logic that the store is a more stimulating environment and that the effect of a shopping list in-store is less strident than the saved list of regular purchases in a virtual list or regular items. It is also possible that online purchases increase entropy with attendant effects on the *health score* of products. Thus, using Smith (2019) quadrants, four possible segments were defined: systematically healthy (SH), systematically unhealthy (SU), unsystematically healthy (UH), and unsystematically unhealthy (UU) to further explore systematic healthy/unhealthy purchase behaviours across channels.

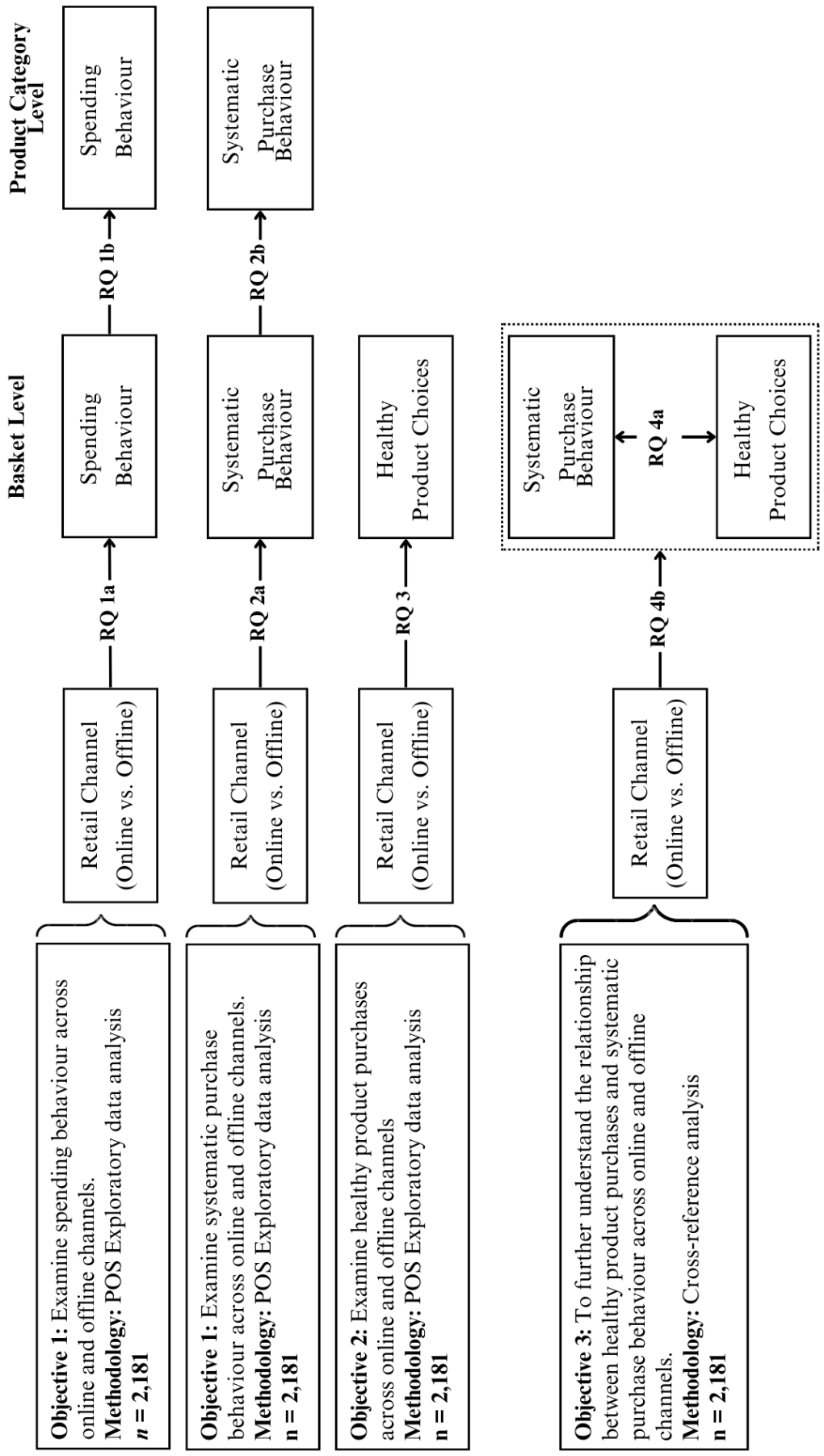


Figure 5.1: Schematic of the objectives and research questions of the study.

5.1.4 Study Design

The study aims to showcase how *bundle entropy* can be used as a valuable measure of systematic purchase behaviour on real-world data to understand differences across retail channels within households. Additionally, the study demonstrates insightful insights that can be obtained by cross-referencing *bundle entropy* with other measures, in this case, a measure of healthy choices across channels (online versus offline). To accomplish this, the study conducts the following steps on the dataset utilised in this study:

- *Frequent customer across channels:* The objective of this study is to gain a comprehensive understanding of the purchasing behaviour of households that purchase food products both online and offline. The study has carefully chosen households that exhibit a consistent pattern of repeat purchases across both channels. To ensure the households are frequent purchases, specific basket and spending criteria were established (refer to section 5.1.6).

- *Data Preprocessing:* To ensure accurate analysis, it is crucial to transform raw data into clean and reliable data. This entails checking for missing data and duplicates, transforming data features to the appropriate type, and eliminating any outliers. By doing so, the data becomes more consistent and accurate, thereby improving the quality of the analysis.

- *Nutritional soft drinks mapping:* As one of the primary aims of this study is to examine the relationship between nutritional choices and systematic purchase behaviour, linking shopping purchases to their nutritional content was an essential element. To accomplish this, we use the retailer's application programming interface (refer to section

5.1.5).

- *Data Engineering:* The study computes three features for each individual: *basket spend*, bundle entropy and *health score* (for soft drinks). It is important to note that the study uses the normalised version of *bundle entropy* (refer to section 5.1.5).

- *Data Analysis:* To compare how *basket spend*, *bundle entropy* and health outcomes vary across channels, the study employs the Wilcoxon signed-ranks test and data distribution analysis. Additionally, the study adapts a consumer behaviour categorisation model from Smith (2019) to classify consumers based on their *bundle entropy* and healthy choices.

The following section provides a comprehensive breakdown of how each step was conducted.

5.1.5 Methods

Measuring spending purchase behaviour

To measure and compare the overall spending purchase behaviours offline versus online (RQ 1a), the total spent per basket for each household (2,181) and channel is computed. Afterwards, the average money spent per basket is defined as *basket spend*; This is done to be consistent with the deployment of *bundle entropy* where a bundle could be a whole basket or a sub-set thereof.

Then, descriptive statistics and visual exploration are performed to compare general differences across channels. Then, the online and offline *basket spend* are tested to examine if there are statistical differences between

their means. Since in this study, the same individual is compared across two different setups (channels) and the *basket spend* data is not normally distributed, the Wilcoxon signed-ranks test was utilised, setting the significance and confidence interval levels at 5% and 95%, respectively. An alpha level of 0.05 was also implemented for all Wilcoxon signed-ranks tests. The same approach was followed to compute the mean money spent per product category to compare spending behaviours across each category (RQ 1b). By conducting this, the study aims to gain insights into the spending behaviour of households offline and online and how they differ across channels and product categories.

Measuring systematic purchase behaviour

This study explores the concept of systematic purchase behaviour across retail venues (online vs offline). This behaviour is defined as the extent to which an individual's basket or bundle composition can be predicted or not predicted. In other words, if a person's purchases are consistent and follow a pattern, then their behaviour is considered systematic. However, if their purchases are random and unpredictable, then their behaviour is considered unsystematic.

The most relevant existing research was examined to determine the best tool to measure this behaviour, specifically studies that examine how to quantify the predictability of consumers' product choices using their transactional records. By doing so, this study aims to better understand systematic purchase behaviour and how it can be measured effectively across different retail setups.

Existing work on brand choice and variety/uncertainty behaviour includes very straightforward methods, such as the distinct count of products in a

single basket where the larger the number, the more variety exists in the purchase occasion (Kahn and Lehmann, 1991). The concept of entropy via uncertainty has been used to measure predictability (Smith et al., 2014; Akaika, 1985) and to measure the variety of items within a set. Alexander (1997) argues that using an entropy-based measure is more reliable and informative than just counting unique items within a single group. Subsequent work further articulates the utility of entropy in this context compared to measures such as the Hirschman–Herfindahl and the Gini coefficient Straathof (2007) with Watson (2009) noting the measure encodes aspects related to the distribution, rareness, and commonness of the products contained in purchase occasion and not only information regarding variety in terms of the number of distinct types of products in a basket. Hence, directly applying entropy will, arguably, be appropriate to capture the uncertainty of predicting a single item within a single set. It does not, however, quantify the predictability of a collection of items.

To illustrate and review this point again, consider a household that, over four transactions, always purchases the same cheese (c) and ham (h). Hence, the sequence of baskets will be [(c,h),(c,h),(c,h),(c,h)]. From a product-level viewpoint, as entropy does, the sequence of baskets will look like a list of all the items [c,h,c,h,c,h,c,h] with a 50/50 probability of predicting either of the items. This is because both products have equal chances to appear next (equal probability of 0.5). As a result, entropy will report maximum uncertainty or unsystematic purchase behaviour. However, looking at the sequence of purchases from a basket-level perspective, it is evident that the same pair of products (c,h) is always bought. Hence, because the probability of (c,h) is 1, we can predict the next basket with 100% confidence.

This shows that applying entropy at different levels (item versus basket)

changes the predictability task. Entropy at the product level, also known as Item Entropy (IE), predicts which item will be added next to the basket at any given time. Here, each distinct item is considered a symbol. Entropy at the basket level or Basket Level Entropy (BLE) considers each distinct basket a symbol, shifting the prediction task to which basket will be purchased next.

This scenario can also be seen from a customer-level perspective. In other words, we can determine whether a consumer's purchasing behaviour is systematic (predictable) or unsystematic (unpredictable) based on his transaction history. Guidotti et al. (2015) propose a measure called Basket Revealed Entropy (BRE) that directly measures how predictable a sequence set of baskets is for a given customer by reducing each customer's baskets to their most frequent purchase patterns (item choice combination). In other words, it creates a new subset of frequent sub-baskets of the initial ones using frequent item-set mining techniques and the Apriori algorithm. Identifying the frequent sub-baskets depends on a minimum support parameter (*minsup*), which determines whether a sub-basket is included or excluded from the frequent sub-baskets. A sub-basket can range from one to all items within the basket. Then, each basket is classified into the largest frequent sub-basket it contains. If a basket does not include any of the frequent sub-baskets found, it will be represented by itself with a frequency of one. After this, each original basket no longer contains the original items. Instead, only contains the assigned frequent sub-basket with a given appearance probability. BRE is then calculated using the traditional formula of entropy.

BRE has certain limitations that need to be considered. One of these limitations is that the *minsup* parameter, which specifies the minimum percentage of baskets that contain a specific frequent sub-basket, needs to

be manually set. The choice of *minsup* can greatly affect the performance of BRE. Generally, higher *minsup* values result in lower BRE scores, while lower *minsup* values produce higher scores. Consequently, selecting the appropriate *minsup* value is not a straightforward task and requires prior knowledge and understanding of the dataset.

The BRE strategy proposes a unique approach to predict the next frequent sub-basket in a set of purchases. According to this strategy, each sub-basket from a new subset of frequent sub-baskets becomes a symbol. Predicting the next frequent sub-basket involves disregarding any infrequent items in the original basket. This is different from precisely predicting the original basket of a given customer without simplifying or removing any infrequent items purchased across a set of baskets.

To measure the exact predictability of the original baskets across a set of purchases, Mansilla et al. (2022) proposed a novel metric called *bundle entropy*. This metric ensures three intuitive properties that make it suitable for real-world applications. These properties ensure that:

- A customer's set of baskets where all baskets are identical will be represented by a *bundle entropy* of *zero*, meaning a completely systematic purchase behaviour (fully predictable).
- A customer's set of baskets where all baskets are unique will be represented by a *bundle entropy* of *one* (in case *bundle entropy* is normalized), meaning a completely unsystematic purchase behaviour (fully unpredictable).
- Baskets with a higher degree of similarity, in terms of sub-baskets, will have lower *bundle entropy* values than baskets with lower similarities.

By incorporating these properties, *bundle entropy* provides a practical and

effective solution to the predictability task.

This study considers customers' original baskets, not subsets, which might exclude valuable information. Considering this and the previous evidence, the study will employ *bundle entropy* to assess the consumers' systematic purchase behaviour. The examples in Table 5.1 illustrate some simple purchase scenarios that support the decision to use *bundle entropy* instead of the other available measures. Therefore, to explore consumers' product purchase uncertainty across channels, the study computes the *bundle entropy* for each individual by applying equation 5.1 (Mansilla et al., 2022) to their whole set of baskets, where each product has its own product identifier (ID) [1]. This way, we get two *bundle entropy* scores per person, one for online purchases and the other for offline purchases (RQ 2a). Afterwards, the *bundle entropy* for each product category online and offline is computed (RQ 2b). The study compares channel-specific statistics using descriptive statistics and visual analysis. Finally, the Wilcoxon signed-ranks test is applied to analyse statistical differences between online and offline *bundle entropy* mean scores at the basket and product category levels.

Normalised Bundle Entropy

When comparing measures like entropy-based measurements, it is essential to standardise them. This is because comparing raw measures can be misleading and can lead to erroneous conclusions. For example, this study will compare the *bundle entropy* of the same households across two different channels. To ensure a fair comparison, the measurements need to

[1] The product IDs are specific to different pack sizes. Hence, a can of Diet Coke will be regarded as a different product from a 1-litre bottle of the same. This is justifiable since the pack size is a function of the purpose and utility derived. Equating products purely on brand and variant is highly questionable given the likely variance in needs and requirements concerning pack size.

be normalised. Normalisation will help to account for differences in choice set sizes. Doing so avoids the risk of conflating uncertainty measures with access to larger choice sets. This will enable accurate and meaningful comparisons between the two channels.

Consumer goods purchasing involves making decisions within a choice set, which can be influenced by multiple external factors, such as household size, income, education, and buying for others, as well as internal factors, such as variety-seeking, self-control, and frugality. To ensure a fair and meaningful comparison of product choices, it's essential to normalise against a measure of choice set size, such as the number of unique baskets. This normalisation helps to account for the impact of the size of choice set on the product choice behaviour of customers. To achieve this normalisation, previous studies (Guidotti et al., 2015; Mansilla et al., 2022) have recommended dividing the measure of *bundle entropy* by the number of unique baskets. By doing so, we obtain a normalised *bundle entropy* score between zero and one, which provides an equitable and consistent measure of product choice behaviour. A score of zero indicates low product purchase uncertainty, which means that the customer has stable and predictable purchase choice behaviour. On the other hand, a score of one represents high product purchase uncertainty, indicating that the customer has highly dynamic and unpredictable purchase choice behaviour. Scores between zero and one reflect a more balanced and neutral behaviour.

It is important to note that practitioners can use other normalisation techniques to achieve invariance to different aspects or definitions of choice group sets as long as the *bundle entropy* properties and associated proofs remain valid. While it's possible to use the non-normalised version of *bundle entropy*, it does not account for the effects of varying choice set sizes, which may lead to misinterpretations when comparing individuals across

channels. Therefore, the normalised version of *bundle entropy* (BE) is recommended for accurate comparisons:

$$BE(\mathcal{B}) = \frac{1}{\log_2 |B|} \times \sum_{b_k \in B} p(b_k) R(b_k) \tag{5.1}$$

Where $\mathcal{B}$ represents the list of all baskets purchased by a given household, b_k denotes each basket containing unique items, and B is the list of unique baskets. Thus, $p(b_k)$ represents the probability of observing basket b_k, whereas $R(b_k)$ is a measure of self-information that quantifies the loss associated with presuming the appearance of basket b_k. To illustrate this concept and how BE is computed, consider a synthetic example (also shown in Table 5.1) where Customer 1 (C1) has three baskets, each containing distinct items:

- Basket 1: (a, b, c)
- Basket 2: (a, b, d)
- Basket 3: (a, b, e)

The BE calculation follows several key steps:

1. **Determine the Probability of Each Basket:** Each unique basket appears once, and there are three baskets in total. Thus, the probability $p(b_k)$ of each basket is:

$$p((a, b, c)) = \frac{1}{3}, \quad p((a, b, d)) = \frac{1}{3}, \quad p((a, b, e)) = \frac{1}{3}$$

2. **Calculate the Regret Measure $R(b_k)$ for Each Basket:** To quantify

the predictability of each basket, the regret measure $R(b_k)$ is computed. This measure considers the overlap of each basket b_k with all other baskets $b_q \in B$, weighted by their respective probabilities. For a given basket b_k, $R(b_k)$ is calculated as follows:

$$R(b_k) = -\log_2\left(\sum_{b_q \in B} \frac{|b_k \cap b_q|}{\max(|b_k|, |b_q|)} \times p(b_q)\right)$$

Example Calculations:

For Basket 1, (a, b, c):

$$R((a,b,c)) = -\log_2\left(\frac{1}{3} \times 1 + \frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{2}{3}\right) = -\log_2\left(\frac{7}{9}\right) \approx 0.3626$$

The calculations for $R((a,b,d))$ and $R((a,b,e))$ are identical due to the symmetry of overlaps, resulting in $R(b_k) \approx 0.3626$ for each basket.

3. **Compute Bundle Entropy (BE):** With the regret measures $R(b_k)$ calculated, we compute BE by averaging these values, weighted by $p(b_k)$, and normalizing by the number of unique baskets. This process is represented by:

$$BE(B) = \frac{1}{\log_2(3)} \sum_{b_k \in B} p(b_k) R(b_k)$$

Substituting the values:

$$BE(B) = \frac{1}{1.585}\left(\frac{1}{3} \times 0.3626 + \frac{1}{3} \times 0.3626 + \frac{1}{3} \times 0.3626\right)$$

Simplifying yields:

$$BE(B) = \frac{1}{1.585} \times 0.3626 = 0.229$$

The resulting BE value of approximately 0.229 reflects the relatively low predictability of the basket sequence, indicating a moderate level of entropy in the consumer's purchasing behaviour.

Table 5.1: Comparison of entropy-based measure for purchase behaviour. Figure abbreviations: Bundle Entropy (BE), Item Entropy (IE), Basket Level Entropy (BLE), Basket Revealed Entropy at 10% *minsup* (BRE10), Basket Revealed Entropy at 240% *minsup* (BRE24), Basket Revealed Entropy at 70% *minsup* (BRE70).

Expected Prediction	**Set of Baskets**	**BE**	**IE**	**BLE**	**BRE (10%)**	**BRE (24%)**	**BRE (70%)**
Very high	$[(abc),(abc),(abc)]$	0.00	1.00	0.00	0.00	0.00	0.00
High	$[(abc),(abd),(abe)]$	0.23	0.91	1.00	1.00	1.00	0.00
Medium	$[(ac),(ad),(xc),(xd)]$	0.50	1.00	1.00	1.00	1.00	1.00
Low	$[(ab),(bc),(cd),(de)]$	0.60	0.97	1.00	1.00	1.00	1.00
Very low	$[(abc),(xyz),(pqr)]$	1.00	1.00	1.00	1.00	1.00	1.00

Measuring the healthiness of products (soft drinks)

In order to address research questions 3, 4a, and 4b, the study will compute the *health score* of each product in the soft drink category. In this study, *health score* is a measure that represents how healthy or unhealthy a product is based on its nutritional components in a single ordinal value. To determine the *health score*[2], it will be used a nutrient profiling model developed by the UK Food Standards Agency (Food Standards Agency, 2006). However, in order to compute this it is necessary to start by linking the purchased products (soft drinks) to their nutritional content.

The dataset used in this study does not contain a text string indicating the

[2]There are several methods to measure or categorise products into healthy or unhealthy using their nutrient profile. For instance, the Swedish Keyhole scheme is a well-known method in Europe (The Swedish National Food Agency, 1980). Similarly, the front-of-pack scheme is another accepted method that labels food using colour-coded signals (red, amber, and green) (Food Standards Agency, 2007). However, both of these methods produce a nominal scale for categorising products. On the other hand, single-score profiling methods are quantitative measures to score healthy and unhealthy products. The most advanced single-score model is 'The UK Ofcom Nutrient Profiling Model', which produces a unique score determined from essential micronutrients (Rayner et al., 2009).

name of a specific product. However, all products are categorised into a six-level hierarchy based on their *division*, *group*, *department*, *class*, *subclass*, and *brand*. This classification system is used to match the products with their corresponding nutritional content. As mentioned previously, the study uses the retailer's APIs[3] as the data source to map the soft drinks to their nutritional content. In this regard, a two-stage method was employed to match the nutritional data to the product data.

1. Firstly, the soft drinks dataset was imported into a new PostGreSQL database, including the unique identifier of each product and all the relevant hierarchy levels. The matching process was performed manually using the *subclass* and *brand* fields of each product in the product table. These fields were typed into a Python code that connected to the Tesco API for *grocery search* to find the closest match. Once a product was matched with one in the API, a unique internal code named *tpnb* was added to the PostGreSQL product table as a new column.

2. Secondly, the *tpnb* of each soft drink was entered into another Tesco API for *macronutrients search*. The *tpnb* allowed access to the exact macronutrients for each soft drink. The API contained information on specific macronutrients, including Energy (kJ), Energy (kcal), Fat (g), Saturates (g), Carbohydrate (g), Sugars (g), Fibre (g), Protein (g), and Salt (g). Each of these macronutrients was added to the PostGreSQL product table as a new column.

Using the method described above, the study matched all soft drinks in the retailer's dataset to their corresponding nutritional data, representing 100% of the total sales quantity and spend on soft drinks.

[3]https://www.tescolabs.com/

The *health score* was computed using the new PostGreSQL product table that contains the macronutrient information of each soft drink. The computation involved the following steps:

1. Calculate the total 'A' points using equation F.1 (See Appendix F.0.1), where each product nutrient attribute (energy, fat, sugar, and salt) can get points between 0 and 10 (see Table F.1 in Appendix F.0.1). However, if the product scores more or equal to 11 points in total 'A', protein points will not count unless the product scores five on fruit, vegetable, and nut attributes.

2. Then calculate the total 'C' points with equation F.2 (See Appendix F.0.1), where nutrient attributes (fibre, protein and fruit, vegetable and nut points) can score points between 0 and 5 (see Table F.2 in Appendix F.0.1).

3. Then, compute the overall score, which can be calculated differently. First, if 'A' points are less than 11, then equation F.3 is applied (See Appendix F.0.1). Also, equation F.3 needs to be applied if 'A' points are equal or greater than 11, and the product gets 5 points on 'fruit, vegetables, and nuts'. Otherwise, we use equation F.4 (See Appendix F.0.1).

4. Finally, the score is adjusted to a 1 to 100 scale with equation F.5 (See Appendix F.0.1).

This direct measure provides then an ordinal score for each product. After computing the *health score* for each product, the values ranged from a minimum score of 67.0 for high-sugar carbonated drinks to a maximum score of 72.0 for different types of water. Then the minimum and maximum values were used to classify each product into five categories (very

unhealthy, unhealthy, neutral, healthy, and very healthy) based on the criteria outlined in Table 5.2. This classification more accurately reflects how consumers evaluate these products based on intuition rather than using a cardinal score. Additionally, this approach provides greater interpretability and allows us to categorise products more effectively. Examples of product categories for each classification can be found in Table 5.2.

Table 5.2: Health score classification with assigned values for analysis purposes.

Health score	**Healthy classification**	**Assigned value**	**Product category example**
67.0 - 68.0	Very unhealthy	-2	High-sugar flavoured carbonated drinks
68.0 - 69.0	Unhealthy	-1	Fruit juices with added sugar/sweetener
69.0 - 70.0	Neutral	0	Diet colas
70.0 - 71.0	Healthy	1	100% Natural fruit juice
71.0 - 72.0	Very healthy	2	Water (still/sparkling)

Afterwards, the healthy categories are assigned to a five scale (shown in Table 5.2), which allows the calculation of a more interpretable average *health score* for the soft drinks bundle across all individuals. The Wilcoxon signed-ranks test is employed to determine statistical differences, with significance levels set at 5% and a confidence interval of 95%. The bundle *health score* is then compared to *bundle entropy* values. This study utilises a distribution analysis and a modified version of Smith (2019) consumer behaviour categorisation model to classify consumers based on their *bundle entropy* and healthy choices.

5.1.6 Empirical Results

The derived household cohort characteristics

The study had access to purchase history data of 1,130,262 customers from a UK-based grocery retailer. To thoroughly investigate the research ques-

tions, a sample of the dataset was extracted to examine differences within the study subjects. For this purpose, the data was filtered to include only the active households that made at least five purchases on each channel and spent a minimum of £5 per basket during the data period. It is important to note that these inclusion criteria are more rigorous compared to the standards utilised in prior studies (e.g. Chu et al. 2010). As a result, the study was left with a sample size of 2,181 customers who exhibited significant online and offline purchases. The raw data reveals that during a 19-month period, 2,822,624 (45%) items were sold online, while 3,429,000 (55%) items were sold offline. Despite being a small subset of the raw data, this information reflects the regular purchasing habits of consumers across both online and offline retail channels. This data is incredibly valuable as it provides deep insights into the consistent consumer choices when making purchases. The study affirms that this balance is reasonable and holds the potential to yield significant and meaningful conclusions.

During the 19 months, the households completed a total of 228,488 transactions (visits/baskets), of which 166,085 (73%) were offline, and 62,403 (27%) were online. Table 5.3 illustrates the comparison between the household characteristics of offline and online transactions. Each household, on average, conducted 106.1 transactions (sd = 77.8), with a mean of 77.1 (sd = 73.3) offline and a mean of 29.0 (sd = 23.8) online. When shopping online, households acquire more items per transaction, with an average of 44.0 (sd = 17.1) items compared to 24.7 (sd = 14.6) items offline. Moreover, they also purchase a greater variety of unique items per transaction online (mean = 18.9; sd = 9.2) than offline (mean = 12.1; sd = 7.2).

The following statistics, presented in Table 5.4, illustrate the percentage of product sales by category for both offline and online channels. It is important to note that the categories that contributed the most to the total

sales for both channels are *Fresh Food* (42.9% offline versus 47.1% online), *Ambient dry grocery* (29.4% offline versus 30.8% online), and *Non-foods grocery* (13.5% offline versus 13.8% online). However, there is a significant disparity in the *Home & Wear* category, which comprises only 4.6% of the total offline sales but less than 1% of the online sales.

In Figure 5.2, it can be seen the distribution of sales in each of the eight product categories across both offline and online channels. The most striking variation is observed in the *Home & Wear* category, where 97.1% of items such as clothing, footwear, and leisure items are sold in-store, and only 2.9% are sold online. In contrast, 88.7% of *Miscellaneous Items* like catering are sold online, and only 11.3% are sold offline. This significant difference may be attributed to various reasons, such as how discounts and offers are promoted in each product category or how customers interact with certain products. It is noteworthy that *Home & Wear* items may require physical evaluation before purchase, unlike *Miscellaneous Items*, which may explain the higher preference for in-store sales in this category.

There are expected differences between offline and online sales in various categories. In the *Wine & Spirits* category, offline sales are much higher, with 68.1%, while online sales are 31.9%. Similarly, in the *Tobacco Kiosks* category, offline sales are 60.7%, while online sales are 39.3%. In the *Bread-/Bakery* category, offline sales are expected to be 60.1% while online sales are 39.9%.

The most interesting finding is in the *Fresh Foods* category. Although most products in this category tend to expire rapidly, offline sales were anticipated to be higher than online sales. However, the offline percentage is only 5% higher than online sales, with offline sales of 52.6% and online sales of 47.4%.

Table 5.3: Transaction characteristics by retail channel (offline versus online)

Transaction Characteristics	**All transactions**	**Offline transactions**	**Online transactions**	*P*
Mean transactions	106.1 ± 77.8	77.1 ± 73.3	29.0 ± 23.8	< 0.001
Mean items per transaction	30.7 ± 13.7	24.7 ± 14.6	44.0 ± 17.1	< 0.001
Mean unique items per transaction	11.2 ± 5.5	12.1 ± 7.2	18.9 ± 9.2	< 0.001

Note: *P* values are determined by the Wilcoxon signed-ranks test.

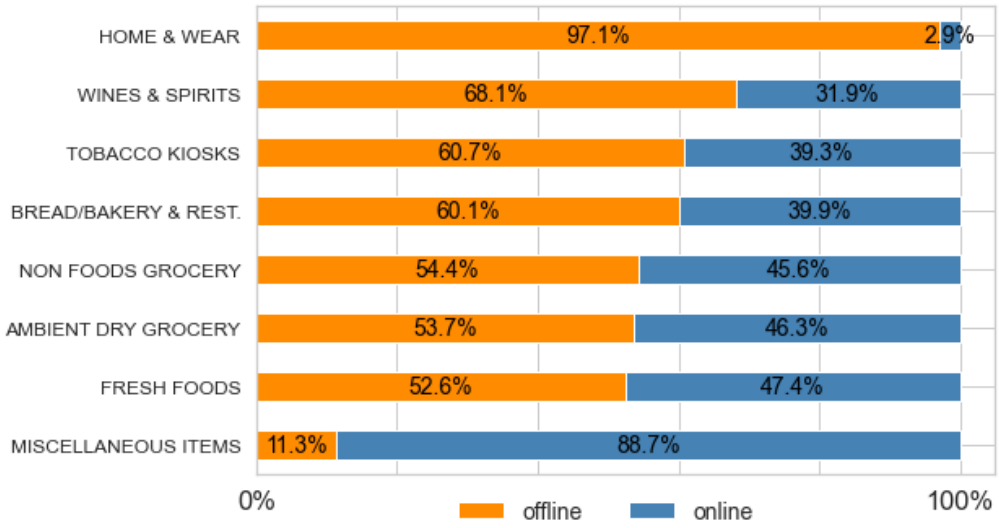


Figure 5.2: Percentage of the relative sales in each product category in offline versus online transactions among 2,181 households.

Table 5.4: Proportion of total items on each product category sold online versus offline.

Product Category	Examples	Percentage of total sales	
		Offline	Online
Fresh Foods	Vegetables, meat & poultry	0.429	0.471
Ambient Dry Grocery	Non-perishable, canned food	0.294	0.308
Non-Foods Grocery	Laundry, cleaning, beauty	0.135	0.138
Bread/Bakery	Bread, cakes, treats	0.069	0.055
Home & Wear	Clothing, footwear, leisure	0.046	0.001
Wines & Spirits	Table wines, beers, spirits	0.022	0.012
Tobacco Kiosks	Cigarettes, lottery, gift cards	0.002	0.001
Miscellaneous Items	Catering	0.001	0.012

Basket spend

As mentioned in the previous section, this study first compares the overall *basket spend* of the same households across the online and offline channels, considering first all product categories. The Wilcoxon signed-ranks test was used to determine if there was a significant difference between the two groups. The results in Table 5.5 shows that there was indeed a noticeable difference between the means, with households spending on average £97.7 online compared to £50.7 offline. This means that households spend, on average, 63.3% more per basket online than they do offline.

To further investigate the differences between online and offline spending, the distribution of *basket spend* for both modes of shopping (see Figure 5.3) was analysed. It can be seen that both distributions are right-skewed due to some households spending more than the average per basket. However, these outliers are not significant enough to be considered outliers. Furthermore, it can be observed that the distribution (density) of *basket spend* in the offline channel is taller and more compressed, while the online track is wider and has a higher mean.

Finally, the results of the Wilcoxon test confirm a significant difference between the online (mean = 97.7; sd = 40.3) and offline (mean = 50.7; sd

= 32.2) *basket spend* means, with a ***P***-value ≤ 0.000. These findings suggest that there are significant variations in household spending behaviour across online and offline channels.

Table 5.6 breaks down the *basket spend* for each product category. Overall, the average spending per basket is higher online than offline in five of the eight product categories. The highest difference is observed in the *Ambient Dry Grocery* category, where the average online spending (mean = 28.0; sd = 14.1) is 66% more than the offline spending (mean = 14.1; sd = 9.7). The second-highest difference is observed in the *Fresh Foods* category, where the average online spending (mean = 42.1; sd = 20.3) is 57% more than in-store spending (mean = 23.2; sd = 15.5).

On the other hand, households spend more in-store than online for the *Miscellaneous Items* category, with a difference of 83.3%. Specifically, the average spending for this category online is 2.8±1.2, while in-store spending is 6.8 ± 21.4. In summary, the differences in the online and offline *basket spend* mean values for all categories are statistically significant (***P***-values ≤ 0.000).

Table 5.5: Basket spend (average spend per basket) by channel

Variable	**N**	**Offline Mean (£)**	**Online Mean (£)**	**Percentage Difference**	***P*-value**
Basket spend	2,181	50.7 ± 32.2	97.7 ± 40.3	63.3%	0.000***

***$P \leq 0.001$ Wilcoxon signed-ranks test.

Note 1: percentage difference is calculated based on $(|V1 - V2|/((V1 + V2)/2)) * 100$.

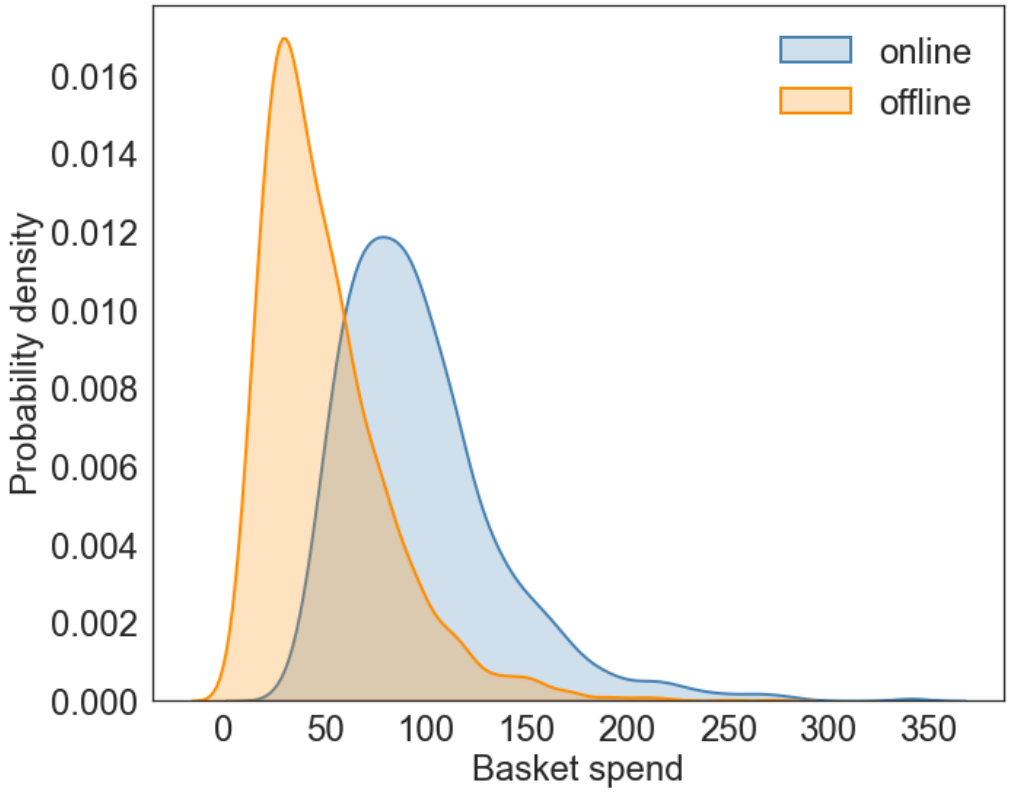


Figure 5.3: Online versus offline basket spend distribution density.

Table 5.6: Basket spend (average spend per basket) by product category

Product Categories	N	Basket spend: Offline Mean (£)	Basket spend: Online Mean (£)	Percentage Difference	*P*-value
Fresh Foods	2,177	23.2 ± 15.5	42.1 ± 20.3	57.9%	0.000***
Ambient Dry Grocery	2,181	14.1 ± 9.7	28.0 ± 14.1	66.0%	0.000***
Wines & Spirits	1,596	15.9 ± 15.9	18.7 ± 18.3	16.2%	0.000***
Non Foods Grocery	2,181	1.9 ± 7.0	16.3 ± 11.8	39.7%	0.000***
Tobacco Kiosks	186	19.1 ± 23.2	13.8 ± 24.2	32.2%	0.000***
Home & Wear	1,176	11.0 ± 9.1	4.8 ± 5.8	78.5%	0.000***
Bread/Bakery	2,113	3.3 ± 1.7	3.7 ± 2.4	11.4%	0.000***
Miscellaneous Items	441	6.8 ± 21.4	2.8 ± 1.2	83.3%	0.000***

$^{***}P \leq 0.001$ Wilcoxon signed-ranks test.
Note 1: percentage difference is calculated based on $(|V1 - V2|/((V1 + V2)/2)) * 100$.

Bundle Entropy

The data presented in Table 5.7 provides a statistical summary of *bundle entropy* for both online and offline channels, along with the results of the Wilcoxon signed-ranks test. The findings demonstrate that if all households are aggregated, they generally experience lower *bundle entropy* online than

offline. The results of the Wilcoxon test establish that the difference in the means of online *bundle entropy* (mean = 0.51; sd = 0.13) and offline *bundle entropy* (mean = 0.75; sd = 0.14) is statistically significant, with a ***P***-value ≤ 0.000 and a percentage difference of 38%. Furthermore, Figure 5.4 illustrates a substantial difference in distribution densities between online and offline *bundle entropy*. Offline *bundle entropy* has a skewed distribution to the left, with a higher peak than online *bundle entropy*, which shows a more symmetrical distribution.

Moreover, the data suggests that when customers make their bundle choices online, the maximum level of *bundle entropy* does not exceed 0.89. This means that, on average, customers are more confident and certain about their bundle choices in the online channel. However, the same cannot be said for the offline channel, where the *bundle entropy* values can go as high as 0.99. This indicates that customers experience higher levels of uncertainty and unpredictability in their bundle choices when shopping offline. Essentially, the offline *bundle entropy* distribution implies that customers tend to make more impulsive and unpredictable purchases when shopping offline than online. These findings provide valuable initial insights into the differences in customer behaviour when making bundle choices in both online and offline channels.

The findings presented in Table 5.8 highlight the *bundle entropy* at the product category level, which is consistent with the results shown in Table 5.7, analysed at the channel level. The *bundle entropy* tends to vary in different product categories and between online and offline channels. The results suggest that the *bundle entropy* is lower online than offline for all categories except for *Miscellaneous Items*, which also accounts for the most significant difference in *bundle entropy*, with a 52.65% variation between the online (mean = 0.81; sd = 0.21) and offline (mean = 0.47; sd = 0.43)

channels.

The category *Home & Wear* exhibits the second-highest difference in *bundle entropy*, with a 47.85% fluctuation between online (mean = 0.55; sd = 0.47) and offline channels (mean = 0.90; sd = 0.18). This means that when shopping online, households tend to choose more consistent product bundles than in-store shopping. This is followed by *Fresh Foods*, which displays a 36.36% difference in *bundle entropy* between online and in-store shopping.

Further analysis reveals that *Ambient Dry Grocery, Bread/Bakery, Non-Foods Grocery, Tobacco Kiosks, and Wines & Spirits* also display a significant difference in *bundle entropy* between online and offline channels. The differences range from 29.83% to 19.04%, respectively. The mean values of *bundle entropy* for each category, both online and offline, are statistically significant, with P-values ≤ 0.000.

Table 5.7: Bundle Entropy by channel

Variable	**N**	**Offline Mean**	**Online Mean**	**Percentage Difference**	***P*-value**
Bundle entropy	2,181	0.75 ± 0.14	$0.51 \pm .13$	65.0%	0.000***

***$P \leq 0.001$ Wilcoxon signed-ranks test.
Note 1: percentage difference is calculated based on $(|V1 - V2|/((V1 + V2)/2)) * 100$.

Table 5.8: Bundle Entropy by product category

		Bundle Entropy			
Product Categories	**N**	**Offline Mean**	**Online Mean**	**Percentage Difference**	***P*-value**
Miscellaneous Items	441	$.47 \pm .43$	$.81 \pm .21$	52.65%	0.000***
Home & Wear	1,176	$.90 \pm .18$	$.55 \pm .47$	47.85%	0.000***
Fresh Foods	2,177	$.72 \pm .15$	$.50 \pm .16$	36.36%	0.000***
Ambient Dry Grocery	2,181	$.80 \pm .13$	$.59 \pm .15$	29.83%	0.000***
Bread/Bakery	2,113	$.76 \pm .16$	$.59 \pm .21$	24.08%	0.000***
Non Foods Grocery	2,181	$.80 \pm .15$	$.64 \pm .15$	22.83%	0.000***
Tobacco Kiosks	186	$.41 \pm .45$	$.34 \pm .38$	19.76%	0.000***
Wines & Spirits	1,596	$.80 \pm .28$	$.66 \pm .35$	19.04%	0.000***

***$P \leq 0.001$ Wilcoxon signed-ranks test.
Note 1: percentage difference is calculated based on $(|V1 - V2|/((V1 + V2)/2)) * 100$.

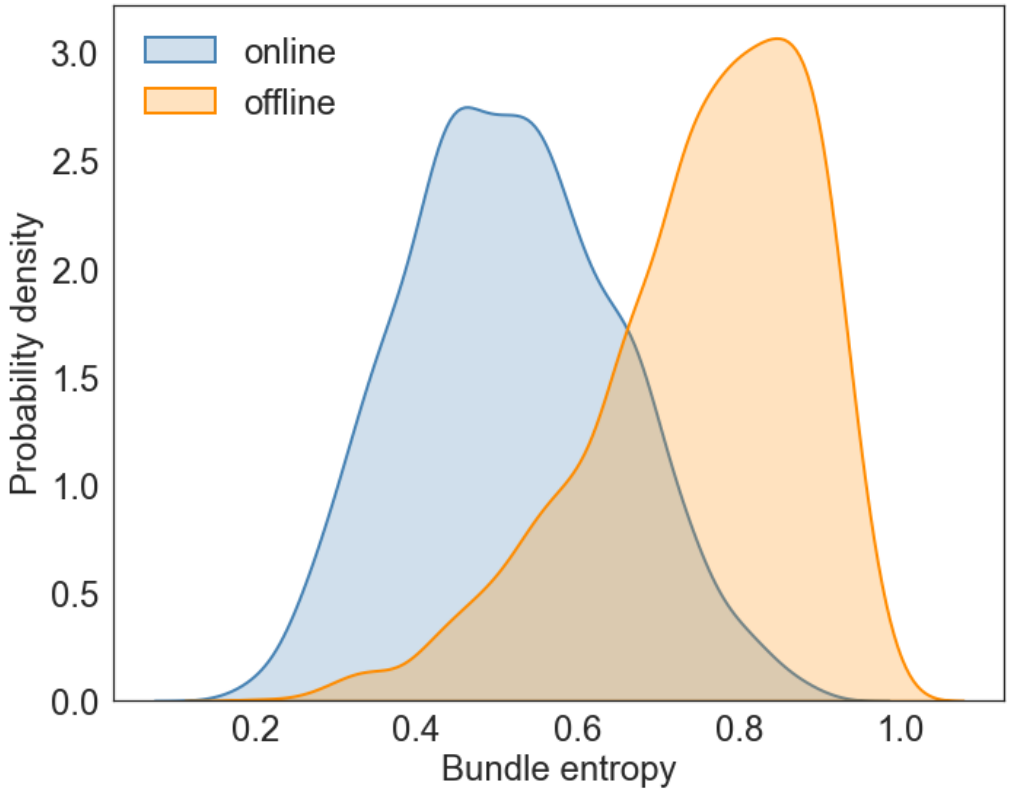


Figure 5.4: Online versus offline bundle entropy distribution density (the interpolation lines extend beyond the actual minimum value).

Product health attribute outcomes

The descriptive statistics and Wilcoxon signed-ranks test results of the overall ranked score of product health attributes per channel in the soft drink category are presented in Table 5.9. The results indicate that there is only a minimal difference (1.9%) between the means of the online and offline bundle *health scores*. The Wilcoxon test further confirms that there is no statistical difference (P-value = 0.681) between the means of the online (mean = 0.52; sd = 1.00) and offline (mean = 0.53; sd = 0.86) *health score*.

However, a closer examination of Figure 5.5 reveals that the distribution densities of the online and offline channels are remarkably different. Despite having similar means between each channel, the actual health outcomes and choices of customers are significantly different. The distribution density

of the offline channel appears to follow a distribution pattern closer to normal, with a mode around 0.5. This suggests that most customers tend to purchase either neutral (e.g. diet cola) or healthy drinks (e.g. natural fruit juice).

On the other hand, the online channel displays a more bi-modal distribution pattern with peaks on the values -1 and 2. This indicates that customers tend to choose either unhealthy (e.g. full sugar cola) or very healthy drinks (e.g. water) rather than more neutral products, which are still well represented. It is also noteworthy that the range of products bought online is wider than those purchased offline, which explains the higher standard deviation of 1.0 in the online channel compared to 0.8 in the offline channel.

Overall, the results suggest that while the means of the online and offline channels for the soft drink category are similar, the actual health outcomes and choices of customers differ significantly between the two channels. This highlights the importance of understanding and catering to customers' different preferences and purchasing patterns across various channels.

Table 5.9: Bundle Health Score by channel

Variable	**N**	**Offline Mean**	**Online Mean**	**Percentage Difference**	***P*-value**
Bundle health score	2,181	0.53 ± 0.8	0.52 ± 1.0	1.9%	0.681

Note 1: the P-value is determined from the Wilcoxon signed-ranks test.
Note 2: percentage difference is calculated based on $(|V1 - V2|/((V1 + V2)/2)) * 100$.

Bundle entropy versus health attribute outcomes

Figure 5.6 provides a visual representation of the relationship between *bundle entropy* and bundle *health score* attribute on each channel. This density plot helps identify patterns and trends in household purchasing behaviour. The plot also displays the conceptual categorisation of households

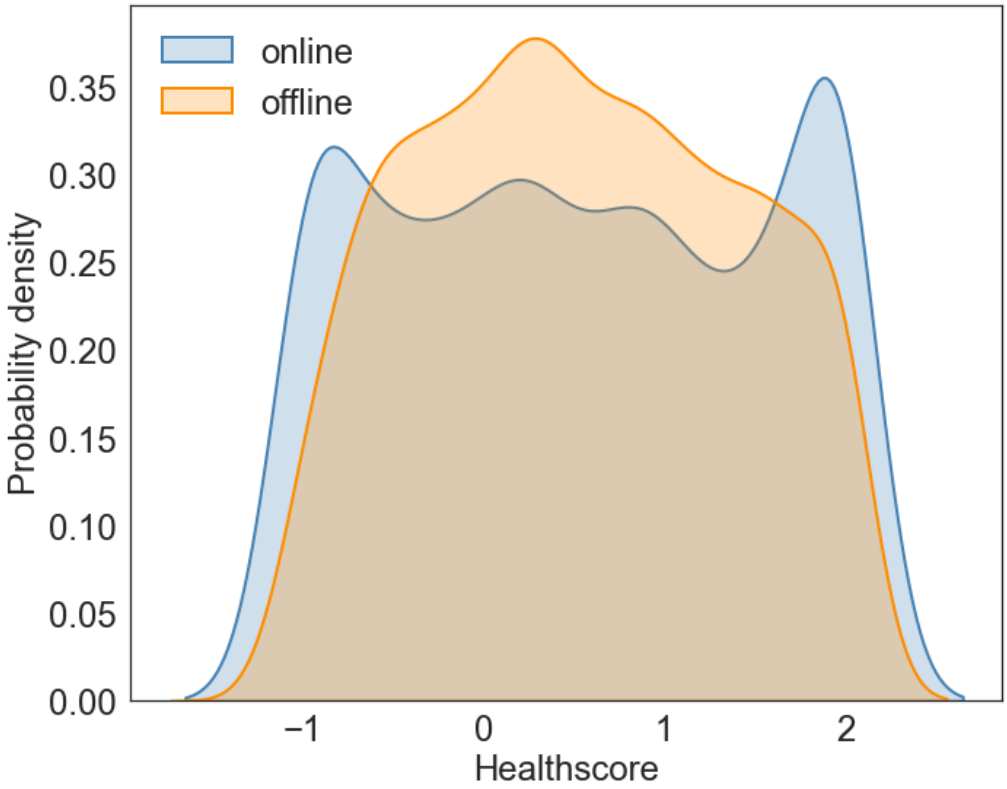


Figure 5.5: Online versus offline health score distribution density.

into four groups based on their *health score* attribute: *unsystematically unhealthy*, *unsystematically healthy*, *systematically healthy*, and *systematically unhealthy*. This categorisation has been adapted from Smith (2019).

Upon analysing the online channel graph, it is observed that around 31.91% of households can be classified as *systematically healthy*. These households are highly predictable in terms of their purchasing choice behaviour of healthy drinks, expressing low *bundle entropy*. They tend to buy a single product or a combination of healthy drinks in a systematic manner. For instance, some households consistently buy a single product (e.g. still water), while others opt for different bundles of healthy drinks (e.g. sparkling water and 100% natural orange juice).

The second largest group in the online channel, representing roughly 30.08% of households, is classified as *systematically unhealthy*. These households

tend to purchase unhealthy soft drinks with poor nutritional value, such as regular full-sugar cola drinks. They exhibit a neutral to low *bundle entropy* regarding unhealthy soft drinks, meaning that their purchasing behaviour is highly predictable.

Overall, the online channel data reveals that about 62% of consumers exhibit highly predictable purchasing behaviour, while only 38% of them are less predictable.

The offline channel shows a different story. In this channel, the largest groups are those that are *unsystematically unhealthy* (36.27%) and *unsystematically healthy* (31.04%), constituting approximately 67.1% of the total households. The *unsystematically unhealthy* group represents households with an unpredictable and random mix of unhealthy soft drink products, with no repetition or a meagre rate of bundle repetition. In contrast, the *unsystematically healthy* group describes households that purchase an unsystematic, thus highly unpredictable, mix of healthy soft drink products.

The next group, *systematically healthy*, represents households that regularly purchase the same items but in the healthy range. This group accounts for 17.1% of the total sample. Finally, the consumers classified as *systematically unhealthy* represent households that consistently purchase unhealthy soft drink products. This group accounts for 14.67% of households.

There is a marked and striking difference in the morphology of choice between channels. The channel and decision-making are clearly associated with pronounced and observably variant outcomes regarding variety and health attributes. Initial topic analysis (in section 5.1.6) of the whole basket does not suggest entirely different shopping missions; the differences do not appear to be variant functions of the basket (most are primary weekly baskets or substantive top-ups), and baskets are distributed throughout the

19 months as per the sampling protocol. Possible explanations are reviewed below. Please find the detailed breakdown of the percentage representation of each soft drink out of the total items sold per channel in Appendix G. It also includes the percentage of soft drinks sold online and offline. Both support the marked difference in the morphology of choice online vs offline.

5.2 Discussion and Conclusion

This study's first research question (**RQ 1a**) was to investigate the difference in consumer spending behaviour between online and offline channels. To achieve this, the study computed the *basket spend* (average spend per basket) for both channels (online and offline) and analysed the results.

The findings indicated that the mean *basket spend* was significantly higher in the online channel, with consumers spending almost twice the amount on each visit compared to offline purchases (online mean = 97.7 vs offline mean = 50.7). This difference was attributed to the unique features of online shopping, such as lower search costs, convenience, and delivery functions, which make it easier for consumers to make purchases. This finding aligns with previous research, such as the work of Smith (2015), who highlighted the growing trend of increased consumer spending in online channels.

Additionally, the study analysed the spending patterns of households across various product categories. The results showed that, in most cases, households spend more per basket online than offline. However, there were a few exceptions, such as *Tobacco Kiosks, Home & Wear*, and *Miscellaneous Items*, which had higher *basket spend* offline. These categories also had fewer households, 186, 1,176, and 441, respectively.

Overall, the study found that aside from a few exceptions, the general finding is that, from a channel (**RQ 1a**) and a product category (**RQ 1b**) perspective, households tend to spend more per basket online than offline. This is broadly in line with previous studies (e.g. Andrews and Currim 2004 and Chu et al. 2010) that state that consumers are, in general, less price-sensitive when they shop online than in traditional stores. These higher *basket spend* patterns might be influenced by unique online features like lower search costs, convenience, delivery functions, and the shopping mission's nature. Importantly, our study unveils a salient finding that contradicts conventional expectations. Despite the larger potential pool of products and greater spending online, we found that the spending pattern does not lead to greater entropy. These findings challenge assumptions and highlight that shopping purchase behaviour is much greater online than offline.

In order to address **RQ 2a**, an analysis of the *bundle entropy* of both the online and offline channels was conducted. The findings suggest that households experience less *bundle entropy* (the level of uncertainty involved in selecting a bundle of products) in the online channel, with a mean score of 0.51, compared to the offline channel, with a mean score of 0.75.

To investigate this phenomenon further (**RQ 2a**), *bundle entropy* was examined across all product categories. Interestingly, it was found that for all categories except *Miscellaneous Items*, the *bundle entropy* score was lower online than offline, with a reduction of at least 20%.

These results indicate that the decision-making process for online shopping differs significantly from that of offline shopping. In a physical store, consumers are in a more engaging environment where they can easily come across products they may wish to buy and are often influenced by the pub-

lic nature of the context (Ratner and Kahn, 2002). On the other hand, online shopping is characterised by individually targeted offers, web page configurations (such as personalised recommendations), and a list of previously purchased products that are updated and edited based on past transactions.

Assertions made about analytics informing and influencing purchase (exogenous cognition) by Smith et al. (2021) provides a comprehensive and credible explanation for the differences online and offline. According to Xu et al. (2022), the channel itself can impact the decision-making process by shaping preferences and morphology. In particular, online channels can be biased due to analytics-driven factors such as personalised offers, push notifications, and page configurations. Although the channel can affect the choices made, more research is needed to understand the differential decision process, as highlighted by Ratchford et al. (2022). Therefore, it is crucial to conduct empirical tests using transactional data or experimental designs to validate these ideas. Through such studies, we can gain a deeper understanding of how customers make decisions in different channels and, in turn, help businesses improve their strategies accordingly.

Differential tendencies for variety-seeking according to channel is another explanation of variance in *bundle entropy*, but why is it more marked in-store? Again, the work of Ratner and Kahn (2002) might help to explain this to some extent, but it cannot account for the differential in health attribute choice.

The literature on variety-seeking promotes the idea that this is a decisive factor. However, from the registration information for the retail dataset used here, we know that only 5% of the households are single-person units. Variety-seeking is often posited as an individual characteristic or propen-

sity, and how this affects purchases on behalf of others (by the primary shopper) is unclear. We, therefore, contend that further research is required to cross-reference purchase data with data on the differential effect of the decision process for discrete channels and research into the effect of variety-seeking propensities for multi-person households.

Consumer research has traditionally focused on individual processes, but many high and low-involvement products are bought through household decision-making that involves multiple people (Kirchler, 1995). Therefore, to fully understand variety-seeking tendencies, it is necessary to cross-reference psychographic data with purchase data for various household configurations. Such research will provide insight into how variety-seeking tendencies influence purchasing behaviour in different channels and how they impact households with multiple people.

Concerning healthy/unhealthy product choice outcomes (**RQ 3**), we found that the same individual/household does not have a significant statistical difference between the mean online (mean = 0.52) and offline (mean = 0.53) bundle *health score*; however this masks the true anatomy of health choice. We found a marked difference between the online and offline bundle *health score* distributions or morphology. The offline channel showed that individuals tended to opt for either neutral or healthy soft drink products, within the range of 0.0 and 1.0, and had a unimodal distribution (see Table 5.9).

On the other hand, online individuals selected soft drink products that spanned a wider range, from unhealthy to very healthy and had a bi-modal distribution. This difference in product choice might be because product search and comparability cost less in terms of time online than offline (Danaher et al., 2003; Ratchford et al., 2022) or a function of how

nutritional information or other cues and signifiers are presented. This is certainly in line with Nikolova and Inman (2015) study (mentioned in section 5.1.3), which states that a clear display of products' nutritional information could lead to healthier behaviours. Once again, it highlights that the online decision-making process remains somewhat opaque, and while it affects outcomes, more research is needed to understand how and why it happens.

Lastly, this study examined the relationship between *bundle entropy* and *health score* (Figure 5.6). The findings suggest that when it comes to the offline channel, there is a single-mode relationship, whereas the online channel has a bimodal relationship with a different morphology. When *bundle entropy* is compared in the two channels, it was found that there was a significant difference (**RQ 4a**). Specifically, 61.99% of households had less than 0.64 *bundle entropy* in the online channel, while only 31.77% of households had less than 0.64 *bundle entropy* in the offline channel. This significant difference indicates that online consumers are more predictable than offline consumers.

Furthermore, households were classified into four groups and found that the most prominent group in the online channel (32% of consumers) was *systematically healthy*, whereas the *unsystematically unhealthy* group was the most prominent in the offline channel (31% of households). This finding reveals that household decision-makers are more predictable when shopping for soft drinks online than in-store.

Regarding **RQ 4b**, results show that the percentage of systematic households purchasing healthy soft beverages (*systematically healthy*) is higher online (31.91%) than offline (17.1%). This means that approximately 15% of the systematic households purchasing healthy soft drinks online change

their purchase behaviour when shopping in-store. On the other hand, the findings indicate that the proportion of unsystematic households acquiring healthy soft drinks was higher offline (31.04%) than online (15.41%). This suggests that approximately 16% of the unsystematic households are less predictable when shopping in-store than online when it comes to making healthy choices.

5.2.1 Practical implications

The results have identified notable differences in shopping purchasing behaviour that could have significant practical implications for the retail industry's role in promoting public nutrition, health, welfare, and corporate social responsibility. Moreover, the outcomes of this research could have policy implications for public health if they hold true in other contexts.

Bundle entropy is a highly valuable, straightforward, and computationally efficient measure, which makes it easy to apply in various fields of consumer behaviour research, especially where large datasets are available. For instance, retailers can leverage the *bundle entropy* measure to identify which customers are more systematic or predictable and use that information to provide personalised bundles, which could lead to increased revenues. Similarly, this measure can be used as a preliminary step to develop more sophisticated bundle recommendation or next-basket prediction models. Additionally, when combined with other measures, *bundle entropy* can provide new and actionable insights. Further research can be done to examine *bundle entropy* at a more granular level, such as sub-class product categories, as some retailers have hundreds of product sub-classes.

Finally, it is crucial to emphasise the immense potential of the *bundle*

entropy measure to provide valuable insights into probability problems involving multiple combinations. This measure is highly advantageous due to its ability to efficiently and accurately measure the degree of propensity towards specific choice combinations. Its parsimony ensures that it can be easily incorporated into various applications, making it a valuable tool within and outside of consumer research and marketing. The accessibility of the *bundle entropy* measure encourages researchers and practitioners to explore its potential in addressing multifaceted challenges in the retail landscape, fostering innovation and strategic decision-making.

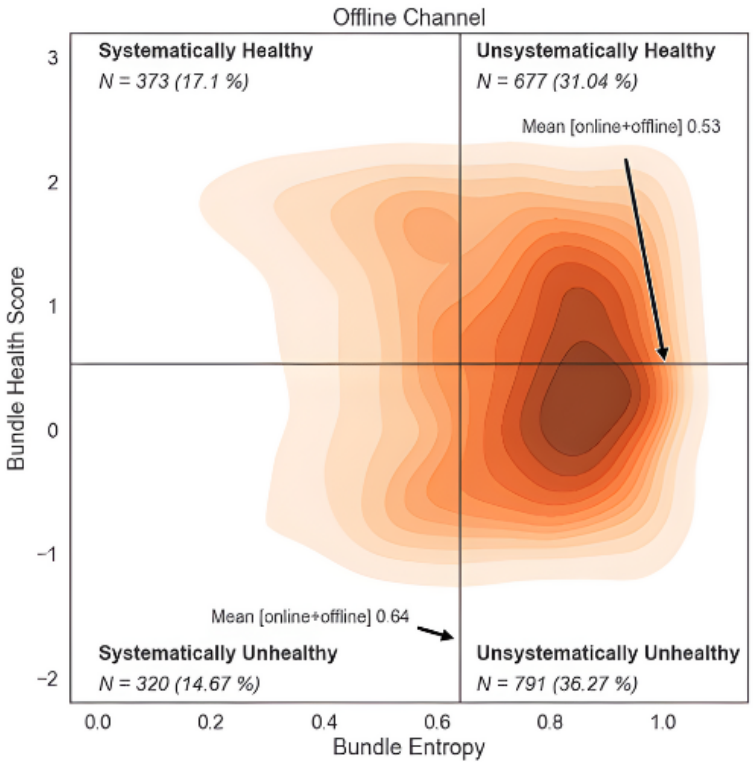
5.3 Subsequent Studies

The research set out to investigate the practical applications of *bundle entropy* and its real-world effectiveness. Specifically, the study aimed to uncover valuable insights for decision-making within the retail industry, with a focus on identifying the consistency of individuals when making online and offline purchases. Furthermore, it sought to determine which types of soft drink products are more commonly purchased online or offline, as well as consumers' preferences for healthier or less healthy soft drinks. The research also highlighted the accessibility and simplicity of using *bundle entropy* as a measure. While the study does have limitations, these will be thoroughly addressed in section 7.3.

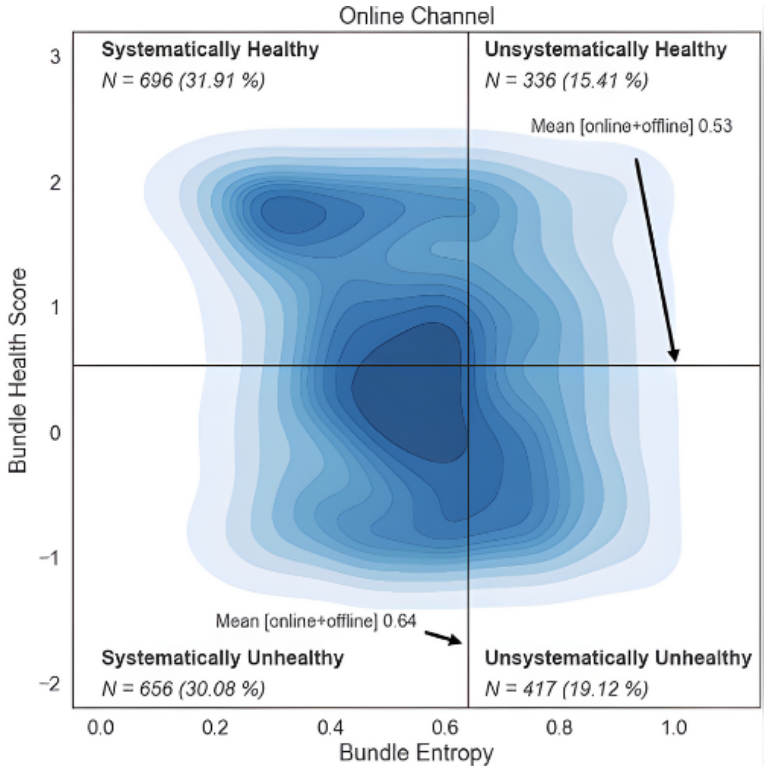
The findings from Studies 1 and 2 highlight the significance of *bundle entropy* in shaping performance and its diverse applications. These applications expand beyond the scope of the studies and could potentially include using *bundle entropy* as an additional variable in predictive models, for example. This approach could shed light on the intricate interactions and

predictive power of *bundle entropy* when combined with other variables.

However, before delving into additional applications of *bundle entropy*, it is essential to thoroughly explore and comprehend its intricacies and identify the primary drivers of its behaviour. This will be the key focus of *study 3*, aimed at gaining a comprehensive understanding of the factors influencing systematic purchase behaviours.



(a) Bundle entropy and bundle healthy score categorisation - offline channel



(b) Bundle entropy and bundle healthy score categorisation - online channel

Figure 5.6: bundle entropy and bundle healthy score categorisation per channel. Mean bundle entropy of 0.64 represents the average bundle entropy across all individuals considering both purchasing channels (online and in-store). Similarly, the mean bundle health score of 0.53 represents the average health score across all individuals in both purchasing settings.

Chapter 6

The Anatomy of Bundle Entropy

This chapter is based on work accepted at the 2024 British Academy of Management conference:

Mansilla, R., Smith, A., Smith, G. and Goulding, J. (2024). The relative power of behavioural, demographic, and psychographic variables as predictors of systematic purchase behaviour. *British Academy of Management.*

Contents

6.1 Study 3: The relative power of behavioural, demographic, and psychographic variables as predictors of bundle entropy

6.1.1 Introduction

For many years, consumer purchasing behaviours have been analysed through the use of demographic and psychographic variables. This practice has been extensively studied, and numerous works in the field have confirmed its importance (e.g. Rich and Jain, 1968, Slocum and Mathews, 1970, Dubois and Duquesne, 1993, Bellman et al., 1999, Baumeister, 2002, Silvera et al., 2008, Islam et al., 2017). However, there is a notable gap in the literature regarding the significance of behavioural variables derived from large transactional records in modelling and understanding purchase behaviours.

The Retail industry has seen a surge in transactional data with the introduction of information technologies like bar codes, point-of-sale systems, and sensors (Rivera et al., 2021; Larson et al., 2005). This exponential growth in data provides an excellent opportunity for researchers and practitioners to delve into consumers' buying behaviours. However, despite the valuable insights that large transactional records hold, there is a noticeable gap in the literature. This gap is primarily due to the numerous challenges associated with managing such extensive volumes of data, including data storage, quality, security, integration, and analysis (Dekimpe, 2020).

While valuable information on various buying behaviours exists within these records, it often remains obscured by the sheer size of the data. Traditional statistical methods, while popular for studying consumer behaviours (e.g. Robertson and Kennedy 1968, Guadagni and Little 2008),

fall short in addressing current requirements, particularly when dealing with large multi-dimensional datasets and intricate relationships (Faraway and Augustin, 2018).

Recent advancements in big data analytics and machine learning have been instrumental in addressing these challenges and limitations of traditional statistical approaches. Machine learning methods, in particular, excel in analysing large multi-dimensional datasets, uncovering both linear and non-linear patterns within the data (Faraway and Augustin, 2018), and accurately predicting complex behaviours across various domains (e.g. environmental behaviours (Lavelle-Hill et al., 2020), child obesity (Long et al., 2023), human mobility (Smith et al., 2014), clinical diagnoses (Dolan et al., 2023a)). Given their versatility, machine learning methods are better aligned with the current needs of both researchers and practitioners (Asniar and Surendro, 2019).

This is because machine learning adopts an algorithmic modelling strategy, unlike traditional statistics, avoiding any assumptions regarding the data-generating process. Its objective is to identify the best function for predicting outcomes from potential predictors (Breiman, 2001). The evaluation of the models focuses on out-of-sample (also known as validation data) prediction performance to guard against overfitting. Features and models are selected based on prediction performance rather than prior data assumptions (Lavelle-Hill et al., 2023). Thus, as mentioned before, these methods have the potential to accurately predict various forms of human behaviour by analysing large multi-dimensional datasets.

However, despite this potential, they have not been widely utilised for understanding consumer behaviour in a retail context (Rivera et al., 2021). There are only a limited number of studies available (e.g. Zuo 2016, Zuo

et al., 2016, Javed Awan et al., 2021) due to several practical obstacles. For instance, accessing comprehensive datasets from retailers is challenging from an operational and data agreement perspective.

Additionally, while machine learning methods excel at prediction, they are not particularly good at providing explanations (Fisher et al., 2019). This is pertinent as consumer research often seeks to comprehend the determinants of the behaviour being studied. As a result, traditional statistical methods, which prioritise explanation and are driven by hypotheses, continue to be favoured (Lavelle-Hill et al., 2023).

Despite the prediction-oriented nature of machine learning methods, there are techniques like variable importance that can provide insights into the most significant factors influencing the behaviour under study. It is important to note that many methods for assessing variable importance prioritise improving prediction accuracy over enhancing the interpretability of the model (see Grömping (2015) for a comprehensive overview of variable importance for regression tasks). However, more recent methods such as SHapley Additive exPlanations (SHAP) values (Lundberg et al., 2017) and Model Class Reliance (MCR) (Fisher et al., 2019; Smith et al., 2020) have shown promise in revealing complex relationships between predictive variables and offering more detailed explanations.

These variable importance methods have demonstrated strong explanatory capabilities in studies of other behaviours, such as child obesity (Long et al., 2023), adherence to asthma (Ljevar et al., 2021), and ovarian cancer (Dolan et al., 2023a). However, there have been limited studies that have explored the use of these variable importance methods to gain insights into the drivers influencing consumer purchase behaviour (e.g. Arboleda-Florez and Castro-Zuluaga 2023, Chen et al., 2023).

This study aims to address a gap in research by leveraging datasets from a prominent UK retail and pharmacy chain. The first dataset, referred to as the *survey dataset* in this study, encompasses demographic and psychographic information of consumers from selected households and loyalty card holders who participated in an incentivised consumer panel. The second dataset, referred to as the *transactional dataset* in this study, comprises transactional information from individuals who took part in the survey over a span of 4 years.

By analysing these combined datasets, the study seeks to investigate the impact of behavioural, demographic, and psychographic variables on *bundle entropy* using traditional statistics and machine learning approaches.

In this study, *bundle entropy* is utilised as a measure of systematic purchase behaviour (SPB), as justified in Study 4 of this thesis. SPB was first introduced by Guidotti et al. (2015) in 2015 to characterise an individual's regular purchasing patterns.

To revisit this concept explained in previous chapters, let's consider a hypothetical purchase sequence from Table 6.1. For instance, *Person 1* has a highly systematic buying behaviour, always buying products w, x, and y together in every purchase. In contrast, *Person 2, 3 and 4* follow different patterns of buying behaviour, changing one or more items on every visit to the store (See Table 6.1), raising the question of who is more systematic compared to *Person 1*?

While it may seem simple to evaluate the probability of each product or the repeat purchase rate for v, w, x, y, or z, assessing the evolution of combinations is more challenging. To address this, a measure called *bundle entropy* was introduced in Chapter 4, which can directly measure SPB from consumer's historical purchases.

Table 6.1: The predictability of different purchase scenarios assessed by bundle entropy.

Expected Predictability	Person	Set of Baskets	Bundle Entropy
Extremely high	Person 1	$[(wxy), (wxy), (wxy)]$	0.00
High	Person 2	$[(wxy), (wxz), (wxv)]$	0.23
Medium	Person 3	$[(wxy), (wxv), (xyz), (xyv)]$	0.50
Low	Person 4	$[(wx), (xy), (yz), (zv)]$	0.60
Extremely low	Person 5	$[(wxy), (abc), (stu)]$	1.00

6.1.2 Background

Demographic variables have been used since the 1960s (e.g. Mathews and Slocum Jr, 1969) to predict consumer behaviour, and while they still offer valuable insights into behaviour, they are not a universal solution. Psychographics can also be useful in predicting consumer purchases as they provide a more in-depth understanding of consumers' attitudes, values, beliefs, interests, and other psychological traits that shape their buying behaviour. However, the efficacy of psychographics (and demographics) in predicting consumer purchase outcomes can vary depending on several factors such as the industry, product category, and target audience (Van Trijp et al., 1996). Many academic studies have been conducted to explore the relationship between psychographics and different consumer purchase outcomes, and these efforts date back to the 1970s when scholars such as Green (1977) delved into the subject.

Efforts in the commercial realm to understand consumer behaviour have an equal longstanding history, dating back to the Values and Lifestyles (VALS) (Mitchell, 1984) or List of Values (LOV) (Kahle et al., 1986) approaches. These methods garnered attention for their innovative attempt to categorise individuals into psychographic segments based on their values and lifestyles rather than just demographics.

More recently, personality traits, specifically the Big Five framework (Costa and McCrae, 1992), have emerged as a popular means of analysing buying behaviour. Initial studies, such as Verplanken and Herabadi (2001), Pirog and Roberts (2007) and Sun and Wu (2014) suggested that conscientiousness and extroversion were negatively correlated with impulsive buying tendencies at the individual level. However, more recent research has challenged this view, indicating that while conscientiousness and extroversion may not have a significant relationship with impulsivity, they are positively linked to neuroticism, openness, and agreeableness (Ali et al., 2022).

Personality traits have also been studied in relation to customer loyalty. While some studies have found no significant correlation between personality and loyalty (Bove and Mitzifiris, 2007), more recent research has identified important relationships. For example, extroversion and openness are strong predictors of brand and product loyalty in the context of hedonic value (Matzler et al., 2006). Agreeableness, meanwhile, is positively correlated with loyalty in the toys and video games product category (Lin, 2010). Finally, conscientiousness, extroversion, and agreeableness all show positive relationships with brand loyalty among university students (Smith, 2012).

These efforts persist today, with passive trait measurement techniques aiming to assign people to psychographic groups in order to correlate these traits with preferences and purchase behaviours. However, the effectiveness of this method remains uncertain since individual behaviour often defies categorisation based on psychographic measures alone (e.g. Barber et al., 2012).

Despite the limitations, psychographics offers a promising alternative or

complement to traditional demographic metrics in understanding consumer behaviour (Myers et al., 1971). A comparative study conducted by Sandy et al. (2013) on the effectiveness of demographics versus psychographics yielded mixed results. While demographics were better at predicting electronic device purchases, psychographics demonstrated superiority in predicting TV show choices. This underscores the potential impact of both psychographic and demographic predictors under different contexts.

Further studies have delved into the relative impact of demographic and attitudinal variables on consumer behaviour. The evidence shows that incorporating both types of variables can greatly improve the accuracy of predictive models. For example, a study conducted by Sorce et al. (2006) demonstrated that the integration of these variables raised the proportion of variance explained by almost 50% in the realm of online shopping intentions. Additionally, Rahim et al. (2014) noted that the interaction between demographic factors and psychographic measures can be quite nuanced, highlighting the complexity of consumer decision-making processes.

Notably, both demographics and psychographics can be influenced by autonomic or syncretic decisions within a household (Kirchler, 1995, 1988). Demographic information such as *age*, *gender*, *income*, and *education level* can provide insights into consumer behaviour but may not capture the full complexity of decision-making within a household. For instance, while *age* can indicate certain purchasing behaviours like having children, it does not necessarily explain how decisions are made within a household.

On the other hand, psychographics delve deeper into the psychological and behavioural traits of individuals, offering insights into their attitudes, values, interests, and lifestyle preferences. These factors can greatly influence consumer choices, particularly in syncretic decision-making scenarios where

multiple household members contribute to the decision-making process. However, even in cases where a lead decision-maker is apparent, psychographics may not always emerge as the dominant predictor of consumer buying behaviour (Sandy et al., 2013).

The issue of household versus individual dynamics further complicates the relationship between demographics and psychographics. While psychographics are inherently focused on individual decision-makers, demographics often exhibit a stronger association with household-level features. For example, demographic variables like household composition and marital status explicitly reflect characteristics of the household unit and can significantly impact purchasing decisions made for the entire household.

Taking into account both demographic and psychographic variables when studying buying behaviour can provide a more comprehensive understanding (Koll and Plank, 2022; Sorce et al., 2006). Demographics offer insights into the structural aspects of consumer populations, while psychographics reveal the motivations and preferences that drive consumer choices. However, a natural question arises: what is the role of past behaviours in studying buying behaviour?

As mentioned before, multiple studies have indicated that demographic and psychographic variables can play a significant role in explaining different purchase behaviour outcomes. However, it is also widely agreed upon that these variables alone can only account for a relatively small portion of the overall variance of buying behaviours (Li and Russell, 1999). For instance, back in the '60s, Rich and Jain (1968) study on choice behaviour already recognised that demographics related to social class (e.g. *income, education, house type*) and lifestyle variables only explained a small proportion of an individual's choices. In a later study, Bellman et al. (1999) explicitly

compared the effectiveness of demographics against past behaviour measures in the online setting and found drastic results.

Demographics explained less than 1% of the online purchase behaviour variance, while past behaviours explain a significant proportion. The contribution of past behaviour has also been compared against psychological traits. Bosnjak et al. (2007) found that in predicting online shopping intention, psychological variables were able to explain 35% of the variance, while in combination with past purchase behaviour, the percentage of variance explained increased to 73%. A more recent study also found that although intentions are generally good predictors of behaviour, some people fail to carry out their intentions and instead revert to past patterns of behaviour unless external factors alter the buying context (Asniar and Surendro, 2019; Koll and Plank, 2022).

6.1.3 Current Work

The study of consumer purchasing behaviours based on transactional records is a relatively new area of research. Some attempts have been made to extract habit tendencies and propensity scores from online purchase and loyalty card data (Smith et al., 2004; Chu et al., 2010; Mansilla et al., 2022). However, there is still a significant research gap in understanding the importance of behavioural variables derived from transactional records in predicting buying behaviours.

One of the primary reasons for this gap is the challenges associated with data accessibility. Obtaining comprehensive datasets that cover diverse behavioural variables on a national scale is more complex than acquiring individual datasets. Retailers have a unique opportunity to collect transac-

tional data nationally, providing a valuable source of behavioural metrics. However, these datasets are often filled with noise and inconsistencies, making analysis complex.

Recent advancements in data analytics have opened up new avenues for consumer behaviour research. Behavioural variables extracted from transactional records (Asniar and Surendro, 2019) and the adoption of advanced approaches like big data analytics and machine learning, instead of traditional statistical approaches, offer promising opportunities (Noori Hussain et al., 2023).

Despite their potential, these approaches have not been widely explored in marketing literature due to the challenges of obtaining comprehensive datasets and creating interpretable machine-learning models with high dimensions.

This study addresses this issue by accessing diverse datasets from a leading UK retailer and pharmacy chain. These datasets connect consumer demographics, psychographics, and purchase data for a group of households and loyalty cardholders who participated in an incentives consumer panel. By accessing these datasets, the study aims to explore the effectiveness of both traditional statistical and machine learning methods in understanding complex buying behaviour from large datasets. Therefore, the study establishes the following research imperative:

- To explore the relative power of demographic, psychographic and behavioural variables in predicting and explaining *bundle entropy* (at the household level) as a measure of systematic purchase behaviour.

6.1.4 Study Design

To explore the impact of transactional, demographic, and psychographic variables on predicting *bundle entropy*, this study combines two real-world datasets with a wide range of predictive variables. As mentioned in the Introduction section, these two datasets are referred to as *survey dataset* and *transactional dataset*, both described later on. Initially, the predictive power of these variables is assessed using Ordinary Least Squares (OLS), a traditional statistical approach. Different group sets of variables are tested to identify the most effective model (See Table 6.4).

Recognising the limitations of traditional statistical methods when dealing with large multi-dimensional datasets and complex relationships (Faraway and Augustin, 2018), the study also investigates machine learning algorithms such as XGBoost (XGB) (Chen et al., 2015) and Random Forest (RF) (Breiman, 2001). These methods are better equipped to analyse large multi-dimensional datasets and can uncover both linear and non-linear patterns within the data (Faraway and Augustin, 2018), enabling accurate predictions. This adaptability makes them well-suited for predicting *bundle entropy*.

In line with standard machine learning practices, the study compares all models against a dummy model to establish a performance benchmark. Here, the dummy model is defined by calculating the mean BE for all individuals within the dataset. This approach allows for an assessment of whether the models exceed random chance. As previously mentioned, both SHAP (Lundberg et al., 2017) and MCR (Fisher et al., 2019; Smith et al., 2020) methods have demonstrated noteworthy explanatory capabilities across various domains. Therefore, this study utilises both SHAP and MCR (both methods further explained in section 6.1.5) methods to evalu-

ate the extent to which each set of variables contributes to understanding the nature of *bundle entropy*.

The study employs various steps to analyse the dataset and develop a comprehensive understanding of the factors influencing *bundle entropy* and, ultimately, SPB. An overview of these methods is as follows (See Figure 6.1 for a general overview of the study design):

- ***Data collection of the transactional and survey datasets:*** This study is based on a comprehensive analysis of two primary datasets from a single retailer and pharmacy chain. The *survey dataset* is derived from an online survey conducted among members of the retailer's loyalty program who consented to content to participate in the survey and share their loyalty program ID so their data can be linked with other datasets. The survey collected extensive information on demographics, psychological factors, and shopping motivations. The *transactional dataset* contains transactional purchase history data of 1 million loyalty program members covering a period of 4 years.

- ***Mapping participants between the survey and the transactional datasets:*** In order to investigate the impact of transactional, demographic, and psychographic attributes on consumers' *bundle entropy*, the study merged the *survey dataset* with the *transactional dataset*. This integration provided a broad spectrum of input variables for analysis using traditional statistical and machine learning methods, as further detailed later on. The retailer had previously linked the *survey* and *transactional* data with the participants' consent, allowing for the matching of customer IDs across both datasets and enabling the exclusion of customers not included in the *survey*

dataset from the *transactional* one. For a more in-depth understanding of the datasets, please see section 3.2.

- ***Data Preprocessing and sample selection:*** Refining the raw datasets is crucial for accurate analysis. This involves removing missing or duplicate data, converting features into their appropriate format, and eliminating irregularities. This makes the data more reliable and easy to analyse. This process utilized a combination of PostgreSQL and Python to ensure data integrity and prepare for analysis. Initially, PostgreSQL facilitated the structuring of data, allowing for efficient management of large datasets. SQL queries were employed to systematically identify and eliminate inconsistencies and duplicate entries, thereby establishing a reliable foundation for subsequent analysis. Following this, Python's data processing libraries, particularly Pandas and NumPy, were used for advanced data cleaning, managing missing values, and transforming variables.

  When refining the data sample, criteria were applied for both upper and lower spending limits per basket to ensure that the analysis only included regular customers with reasonable spending habits. Furthermore, all product categories were included in the analysis to examine *bundle entropy* across the board. More details provided in section 6.1.5

- ***Feature engineering:*** The study uses the *transactional dataset* to compute several metrics related to consumer spending and frequency patterns: *total distinct items*, *average distinct items per basket*, *number of visits*, *days between first and last purchase*, *average gap between visits*, *time of the day (morning, evening, afternoon)*, *average spend per basket*, *median spend per basket*, *total spend*. The output variable is *bundle entropy*, which will be assessed by *bundle entropy*. Details of

how each variable was computed are provided in the following section (6.1.5).

- ***Experimental procedure:*** The experimental procedure comprises two stages. Firstly, the study conducts an OLS regression using all the input variables from the linked *transactional* and *survey datasets*. The predictive power of the OLS model is evaluated using R^2 and adjusted R^2 (standard metrics when assessing traditional linear models). If the OLS model demonstrates strong predictive power, it will be integrated into the second stage as a Linear Regression (LR) model, following the traditional Cross-industry standard process for data mining (CRISP-DM) applied to the other machine learning models under exploration. Section 6.1.5 provides a description of the models being explored.

  In the second stage, the study compares the performance of different machine learning models in predicting consumers' *bundle entropy*. The study adheres to the iterative steps from CRISP-DM (Shearer, 2000), involving model training, model evaluation, and hyper-parameter tuning. Additionally, to assess the relative importance of each variable in predicting *bundle entropy* (See Figure 6.1), the study employs two machine learning techniques: SHAP values and MCR (further explanation provided in section 6.1.5). These methods will aid in enhancing the interpretability of the model.

6.1.5 Methods

This section provides a thorough explanation of the processes and motivations that were applied to the models, such as the inclusion criteria,

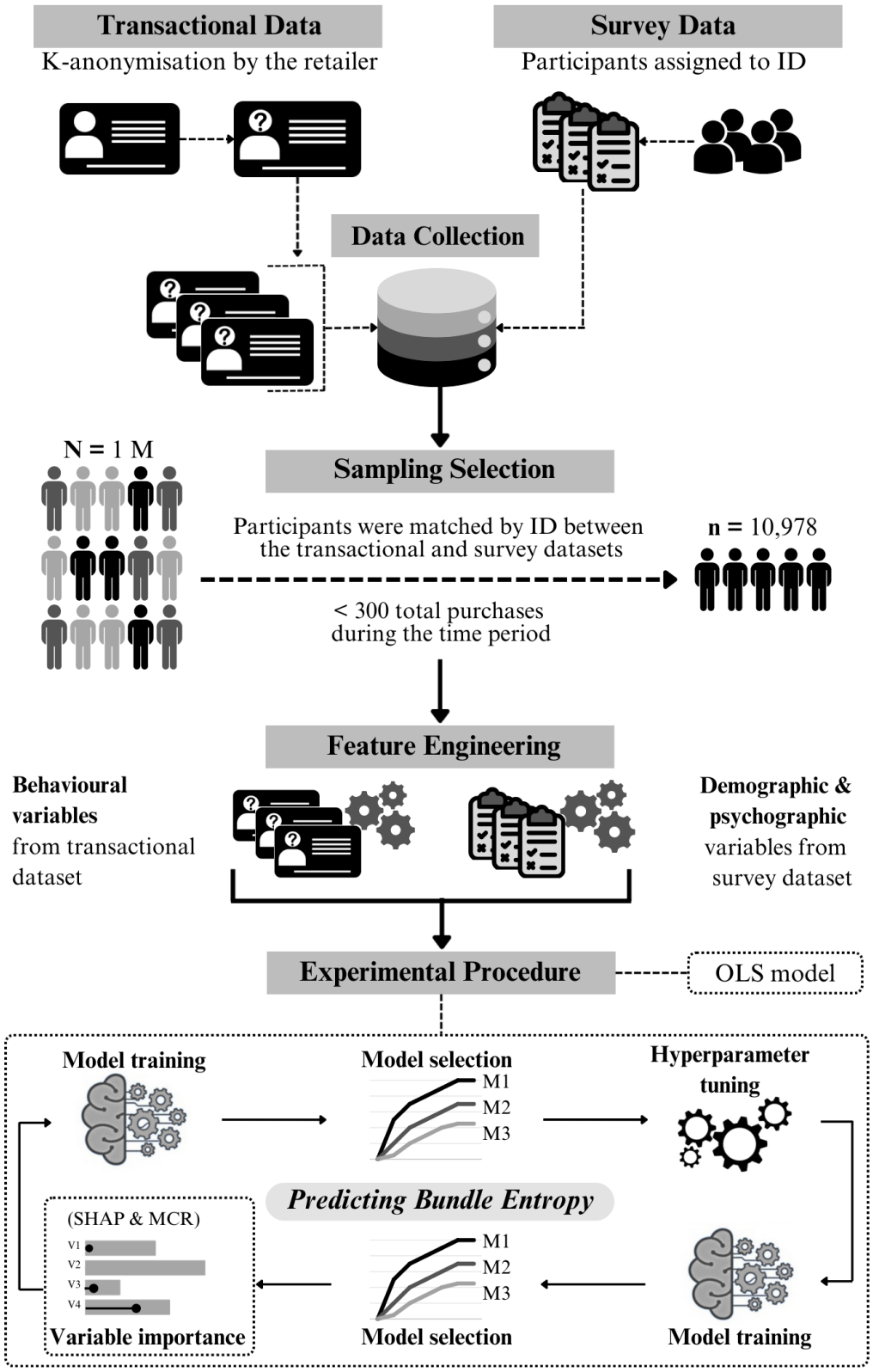


Figure 6.1: Workflow describing the general study design. Figure abbreviations: Ordinary Least Squares (OLS), Model Class Reliance (MCR), SHapley Addictive ex-Planations (SHAP).

cleaning, and the input variables engineered, are also included in this section. Furthermore, this section introduces the machine learning algorithms utilised to model *bundle entropy* and how the best-performing model is selected. Finally, the section concludes by explaining and discussing the feature importance tools employed to determine the relative predictive capabilities of the input variables.

It is worth mentioning that in this study, 'variable' and 'feature,' as well as 'dependent variable' and 'output variable,' and 'independent variable' and 'input variable' are used interchangeably.

Measuring systematic purchase behaviour

This study utilises the definition of *bundle entropy*, as provided by Mansilla et al. (2022). This definition characterises an individual's buying behaviour as either systematic or unsystematic based on their shopping basket over time. To assess consumers' SPB, the study will be utilising *bundle entropy* since its accuracy has been justified in the previous study of this thesis, supporting the decision to use the measure. Please refer to *Study 1a* in section 4.3 for a comprehensive explanation of how *bundle entropy* functions.

Machine learning model selection

Our study aims to predict *bundle entropy* (as a measure of SPB) by leveraging transactional and survey using transactional and survey data. This will be done through a regression task. Given the national coverage of the dataset, it is likely that there will be outliers, noise, and missing data for some of the input variables. We have incorporated 27 input variables from the merge data set, comprising both pre-existing and engineered variables

(e.g. *Average gap between purchases*, *Average distinct items per basket*, *Period covered (in days)*). While a diverse range of input variables can enhance the model's performance, it can also increase the model's complexity, making it difficult to interpret due to its high dimensionality. Nonetheless, this presents an opportunity to investigate a wider array of models that can effectively capture both linear and non-linear connections between variables, which may have gone unnoticed in lower-dimensional spaces.

Consequently, the study compares several machine learning regressors: Linear Regression (LR) (Stigler, 1981), Decision Tree (DT) (Breimann et al., 1984), XGBoost (XGB) (Chen et al., 2015), Random Forest (RF) (Breiman, 2001), K-Nearest Neighbours (KNN) (Cover and Hart, 1967), and Support Vector Machines (SVM) (Hearst et al., 1998). These algorithms are widely used due to their robust predictive power and performance on real-world data. Model evaluation will be conducted using the *R-squared*, the *mean absolute percentage error* (MAPE), and the *mean absolute error* (MAE) (Scheinost et al., 2019).

Model Interpretation

All machine learning algorithms tested in this study have their own unique approach to determining the significance of individual features to the model's performance. For instance, LR quantifies feature importance by considering the absolute values of the coefficients (Tibshirani, 1996). RF measures importance by assessing the mean decrease in impurity per input variable (Leo, 2001). Lastly, XGB calculates importance by analysing the average gain of each variable across all trees and their respective boosting cycles (Chen et al., 2015). They all provide some insight into the variables' relative power prediction. However, they all consider just the single model and

do not account for all the other models that perform equally well or even slightly better (Altmann et al., 2010). Hence, It is possible for results and interpretations to be biased towards a particular model when only one of many equally effective relationships between input features and output is learned. This can result in variables being wrongly considered unimportant, model audits being vulnerable to model retraining, and the interpretation of potential causal features being incomplete and misleading. This same effect might happen using other well-known variable importance methods (e.g. Permutation Importance, Information Gain, and Mutual Information) on single models.

As none of the previously mentioned methods offer insights into the direction of the relationship between the dependent and independent variables, SHAP values (Lundberg et al., 2017) method is employed. Although computing SHAP values on extensive datasets may pose computational challenges, they serve as a practical and interpretable tool for understanding the connection between predictors and the predicted outcome. Through the assessment of all potential feature combinations and the highlighting of their relative importance in influencing the model's predictions, SHAP values assign a weight to each predictor. The way in which SHAP values are interpreted is the following:

- A predictor with a negative SHAP value indicates a decrease in the predicted response compared to the average prediction.
- A predictor with a positive SHAP value denotes an increase in the predicted response in comparison to the average prediction
- A predictors with a SHAP value of zero have no influence on the overall prediction.

Consequently, SHAP values provide a comprehensive and easily interpretable insight into how each feature impacts the model's output.

To improve interpretability and avoid biased interpretation towards a single best-performing model (Fisher et al., 2019), MCR is employed (Smith et al., 2020). MCR is a novel measure that provides insights into the extent to which a given input variable is integral to a set of models achieving comparable predictive performance while utilising distinct predictive factors (See Figure 6.2). In contrast to approaches that assess a variable's significance solely within a single predictive model, MCR delivers a range of values that reflect the variable's reliance across all the sets of well-performing models, also known as Rashomon Set (Fisher et al., 2019). MCR provides two essential metrics. The first one called the Minimum Model Class Reliance (MCR-), denotes the minimum variation in predictive accuracy that can be ascribed to a variable in the Rashomon set. The second metric, Maximum Model Class Reliance (MCR+), gauges the maximum potential shift in predictive accuracy that can be associated with a specific variable in the Rashomon set.

Smith et al. (2020) has broadened the application of MCR to encompass general-purpose non-linear algorithms, facilitating the computation of MCR for both regression and classification RF. This expansion is particularly advantageous as it addresses interdependent features and distinguishes between relevant and irrelevant variables in alternative explanations. The aim is to attain a more thorough, interpretable, and transportable understanding of the meaningfulness of the resulting features in this investigation when predicting *bundle entropy*, by utilising both SHAP and MCR approaches.

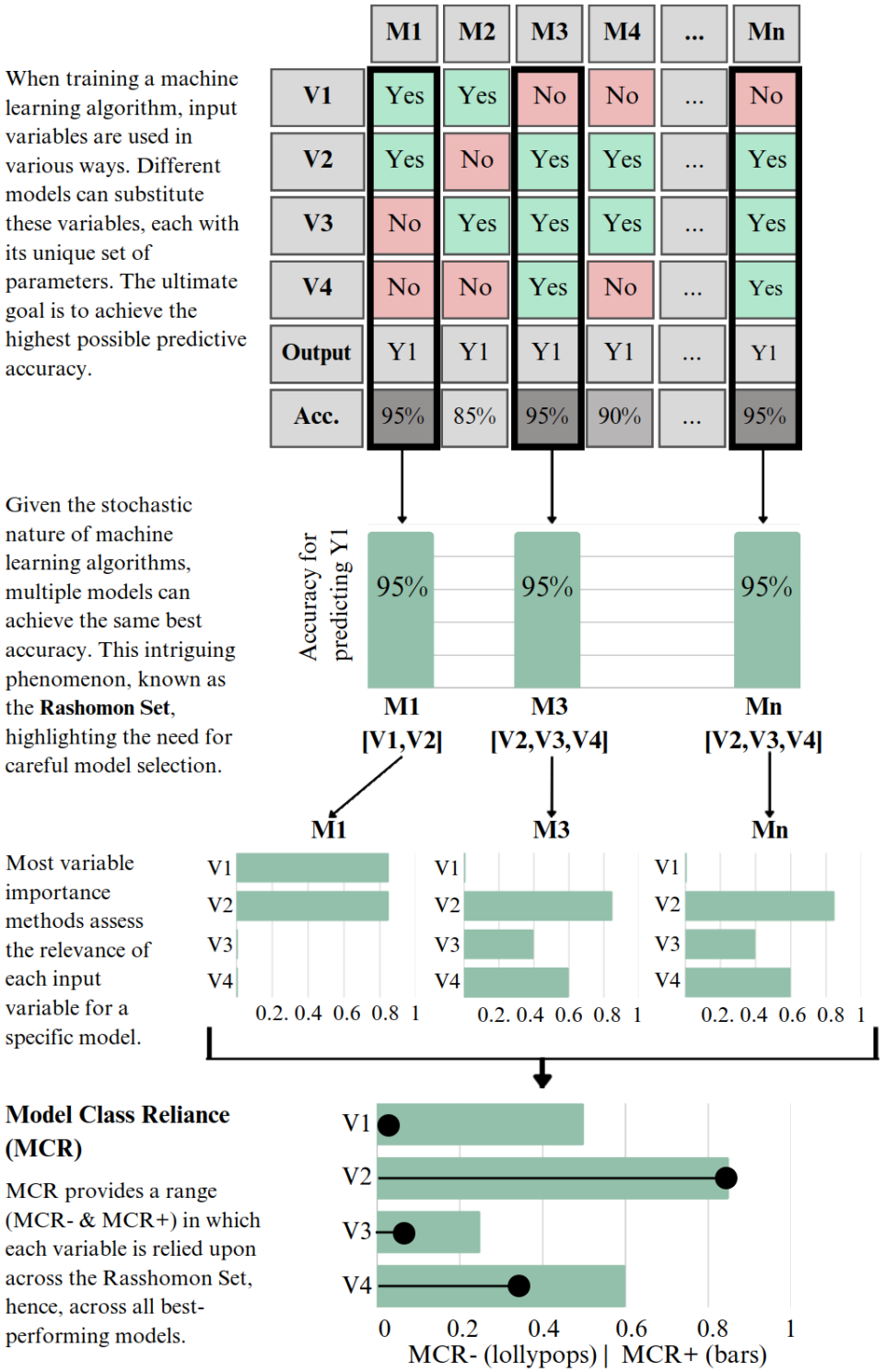


Figure 6.2: Diagram describing the general approach of MCR compared to other methods. Figure abbreviations: Variable (V), Model (M), Output variable (Y1), Model Class Reliance (MCR), SHapley Addictive exPlanations (SHAP).

Data preprocessing and sample selection

The dataset initially comprised transactional records for 1 million customers. Upon integrating the survey data with the customers' purchase history, we found matches for 12,835 individuals. However, 698 participants displayed negative spending values without a clear explanation. We considered the possibility of a storage error, but this could not be confirmed. Consequently, we removed these 698 participants from the dataset, resulting in 12,137 individuals for further analysis.

In general, to thoroughly investigate a specific behaviour over time, a substantial amount of data is required. Therefore, this study focuses on regular customers, which are defined as individuals who frequently make purchases. The definition of a regular customer varies based on the study or business's parameters. In this study, a regular customer is defined as an individual who has made at least five but no more than 300 visits during the entire period. The study set these inclusion criteria after analysing the distribution of the variable *total visits*, which is shown in Figure 6.3. The upper inclusion criteria of 300 visits was set to prevent outliers from skewing the data. As a result, it was found that 1,159 participants did not meet the inclusion criteria. Consequently, the final sample for our analysis consists of 10,978 participants.

Before proceeding with any modelling, all input variables were log-transformed. This is because many of them showed skewness and high variance (refer to Table 6.3). Log transformation can also improve the performance of machine-learning algorithms, particularly linear regression. Although tree-based algorithms are more resilient to the scale of features, they can still benefit from log transformation when input variables are skewed.

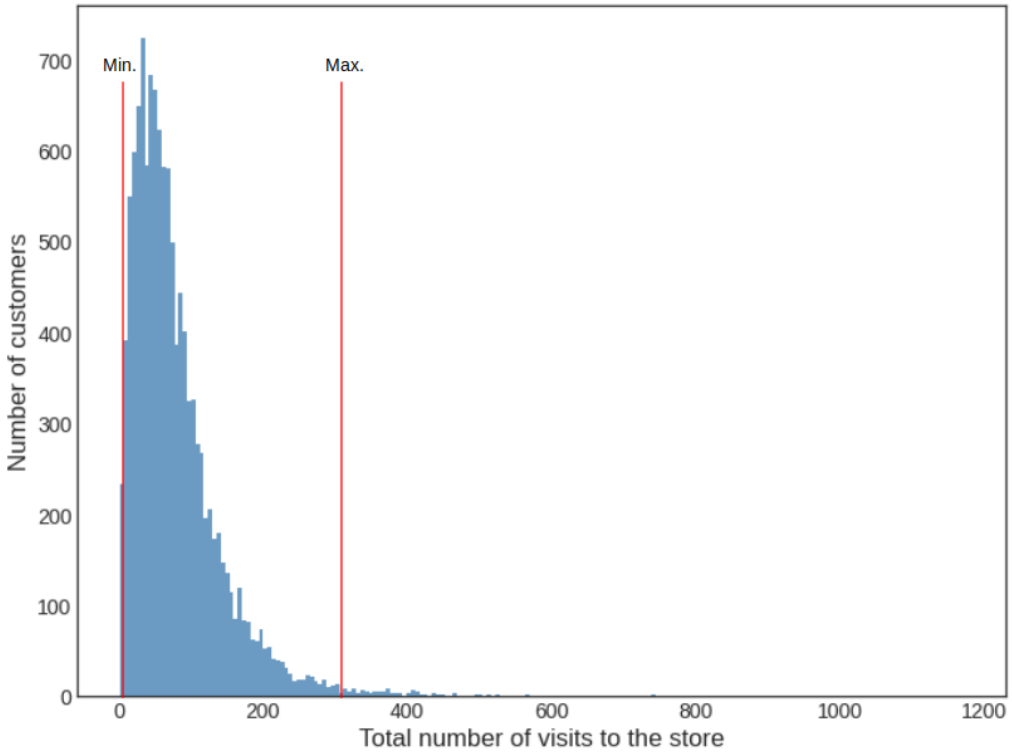


Figure 6.3: Distribution of the total number of visits per customer.

Independent variables

To optimise the machine learning algorithms' performance, some categorical variables from the *survey dataset* required encoding and transformations. This process increases the number of variables input to the models from a total of 27 to 45.

In cases where categorical variables contain multiple levels, encoding these variables into dummy variables increases the overall variable count. Specifically, each unique category within a variable is represented as a separate binary variable, where each dummy variable takes a value of 1 if the observation belongs to that category and 0 otherwise. For example, the variable *gender* was encoded into a dummy variable, where 1 represented male and 0 represented non-male respondents. Variables with more levels, like *day time class*, were encoded into multiple dummy variables, each representing a distinct class such as 'morning,' 'afternoon,' or 'evening'. Similarly,

the variables *qualification* and *marital status*, which contain multiple categories, were transformed into separate binary variables for each category, allowing each unique class to be represented in the model. This process increases the number of variables as these multi-category variables were decomposed into several columns.

Similarly, the *Household composition* categorical variable was transformed into an ordinal variable with three values to avoid redundancy. Respondents who indicated living 'solo' or 'adult with no family' were given a value of one, indicating they likely only purchased for themselves. Those who indicated living with a 'partner' or 'solo with children' were assigned a value of two, indicating they likely purchased for themselves and one other person. Respondents who indicated living with a 'partner and children' or 'family adults' were assigned a value of three, indicating they likely purchased for at least two other people. The *income* variable was also transformed into an ordinal variable due to a clear order in the possible answers. Lastly, the *variety-seeking* variable was created by calculating an average value from the first three questions[1] (7-point Likert scale) in the shopping motivation section of the survey. Question two was re-coded to match the same scale as questions one and three, as it was inversely similar to question one.

The rest of the independent variables are ordinal variables either on a 5 or 7-point Likert scale. The full copy of the survey can be found in Lavelle-Hill et al. 2020 study.

Eight independent variables passed into the model were engineered from the behavioural data set. Let $D = [d_0, d_1, \ldots, d_n]$ represent the list of

[1] 1) I would rather stick to a brand I usually buy than try something I am not very sure of. 2) I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchases. 3) If I like a brand, I rarely switch from it just to try something new.

each individual's unique purchase dates and $B = [b_0, b_1, \ldots, b_n]$ denote the list of unique baskets purchased by an individual. Additionally, let $S = [s_0, s_1, \ldots, s_n]$ convey the list of the total monetary value (£) spent per basket for an individual.

It is worth noting that each individual, basket, and item in the data has a unique identifier that allows the calculation of the following variables:

1. **Total Visits**, it refers to the count of the total number of visits that each individual made to any store. It is calculated using the following equation:

$$TotalVisits(C) = |B| \tag{6.1}$$

   Where C represents each individual in the data set. b_i represents a unique basket purchased by an individual. For this equation, it is assumed that each basket represents one purchase visit.

2. **Inter-basket Mean Gap (IMG)**, it represents the average number of days between purchases, and it was calculated using the following equation:

$$IMG(C) = \frac{\sum_{d_i \in D}(d_i) - (d_{i-1})}{|B| - 1} \tag{6.2}$$

3. **Total Distinct Items (TDI)**, refers to the count of individual items, without repetition, that are present in all the baskets purchased by an individual. To calculate it, the following equation was used:

$$TDI(C) = \sum_{b_i \in B} I(b_i) \tag{6.3}$$

Where $I(b_i)$ represents the unique number of items on a specific basket (b_i). This is possible since each item in the data also has a unique identifier.

4. **Average Items per Basket (AIB)**, refers to the average number of distinct items purchased per individual's shopping basket. It is calculated by dividing the 'TDI' by the total number of baskets purchased by an individual.

$$AIB(C) = \frac{TDI}{|B|} \tag{6.4}$$

5. **Period Covered in Days (PCD)**, is the duration between an individual's first and last purchase, calculated as the total number of days utilising the equation below:

$$PCD(C) = \sum_{d_i \in D} (d_i) - (d_0) \tag{6.5}$$

Where d_0 represents the date of an individual's first purchase and d_i represents the last date.

6. **Total Spend (TS)**, refers to the total monetary value (£) that an individual spent across all their purchases. Computed by the following equation:

$$TS(C) = \sum_{s_i \in S} (s_i) \tag{6.6}$$

7. **Average Basket Spend (ABS)**, refers to the average amount of money (£) spent across all the baskets purchased by an individual. It is calculated by dividing the 'TS' by the total number of baskets

purchased by an individual.

$$ABS(C) = \frac{TS}{|B|} \tag{6.7}$$

8. **Median Basket Spend (MBS)**, represents the monetary value (£) lying in the midpoint of an individual's frequency distribution of purchases (baskets). It was computed by using the following equation: If $|B|$ is odd:

$$MBS(C) = (\frac{|B|+1}{2})^{th} \tag{6.8}$$

If $|B|$ is even:

$$MBS(C) = \frac{(\frac{|B|}{2})^{th} + (\frac{|B|}{2}+1)^{th}}{2} \tag{6.9}$$

A summary of all the variables, including the data source and a brief description, can be found in Table 6.2.

Table 6.2: Description of the independent variables from the transaction and survey data.

	Independent Variable	Data Source	Description
1	Total visits	Transactional	Number of visits during the period covered
2	Inter-basket mean gap	Transactional	Average days between purchases
3	Total distinct items	Transactional	Total number of unique items across all baskets
4	Average items per basket	Transactional	Average count of items per basket
5	Period covered days	Transactional	Total number of days between the first and last basket
6	Total spend	Transactional	Total monetary value (£) spent across all baskets
7	Average basket spend	Transactional	Average monetary value (£) per baskets
8	Median basket spend	Transactional	Median monetary value (£) per basket
9	Day time classification	Transactional	Time of the day when most purchases occurred (morning, evening, afternoon)
10	Age	Survey	Age of the participant
11	Income	Survey	Annual income in the household before taxes
12	Gender	Survey	Type of gender (0: man, 1: woman)
13	Household composition	Survey	Number of people leaving in the same place
14	Qualification	Survey	The highest level of education of the participant
15	Marital status	Survey	The marital status of the participant
16	Openness	Survey	Measures the openness to new experiences
17	Conscientiousness	Survey	Measures of disciplined and carefulness
18	Extroversion	Survey	Measures the level of sociability
19	Agreeableness	Survey	Measures the level of trusting
20	Emotional stability	Survey	Measures of anxiety/pessimism
21	Variety seeking	Survey	Measures the seek of variety
22	Self-control	Survey	Measures the level of self-control
23	Frugality	Survey	Measures general attitudes towards spending
24	Shopping impulsivity	Survey	Measures the level of impulsivity
25	Shopping mission	Survey	Number of product categories bought across all baskets
26	Happiness	Survey	Measures the level of general happiness
27	BAS	Survey	Measures the level of impulsiveness

Dependent variable

In order to determine the *bundle entropy* (BE) for each individual, we applied equation 6.10 to their complete basket of items. In this approach, each item in the basket is assigned a unique product identifier (ID), and distinct product formats are treated as separate entities. For instance, a 250ml bottle of water and a 1-litre bottle of the same product are considered distinct items because their package size impacts their function and value. Evaluating products based solely on their brand and variation is highly questionable, as individual requirements and preferences may necessitate different package sizes.

$$BE(\mathcal{B}) = \frac{1}{\log_2 |B|} \times \sum_{b_k \in B} p(b_k) R(b_k) \tag{6.10}$$

Unique baskets purchased by an individual are denoted by $\boldsymbol{B}$, while the probability of observing a basket b_k is $p(b_k)$ and $\boldsymbol{R}(b_k)$ represents the self-information measure that quantifies the loss associated with presuming the appearance of basket b_k. A more detailed explanation of this equation was described in *Study 1a*, section 4.3.

Descriptive statistics of all dependent and independent continuous and ordinal variables explored during the modelling are shown in Table 6.3. Additionally, the strengths of the relationship among all of them are illustrated in Figure 6.4.

Table 6.3: Descriptive statistics for the dependent and independent variables (N=10,978)

Independent Variable	Min.	Max.	Mean	SD
Bundle entropy	0.13	1.0	0.9	0.1
Interbasket mean gap	1.2	91.0	22.1	15.0
Total distinct items	2.0	438.0	50.5	39.5
Average items per basket	0.1	9.3	2.2	1.0
Total visits	5	298	24.1	20.0
Period covered days	7	365	318.3	48.6
Average basket spend	1.8	507.0	22.3	17.7
Median basket spend	0.4	110.7	4.7	3.7
Total spend	8.8	16,223	500.2	528.9
Age	18	115	50.6	14.6
Income	1.5	7	2.8	1.5
Household composition	1	3	2.1	0.7
Shopping mission	1	19	5.5	4.3
Happiness	1	10	7.1	2.1
Variety seeking	1	7	3.3	1.3
Openness	1	7	4.6	1.2
Conscientiousness	1	7	5.6	1.1
Extroversion	1	7	3.8	1.6
Agreeableness	1	7	5.3	1.2
Emotional stability	1	7	4.4	1.5
Self-control	1	5	2.8	0.7
Frugality	1	6	4.1	0.9
Shopping impulsivity	1	5	2.3	0.9
BAS	1	4	2.3	0.5

Note: SD refers to Standard Deviation.

Experimental procedure

We started by training all the machine learning algorithms (DT, KNN, SPV, LR, XGB, and RF) to predict *bundle entropy* from the behavioural and survey data (the merged data set). As mentioned previously, the task was treated as a regression task, where the outcome variable, *bundle entropy*, ranges from 0 to 1, with 0 indicating low *bundle entropy* and 1 indicating high *bundle entropy*. Scores between 0 and 1 suggest a more balanced or neutral *bundle entropy*.

To thoroughly evaluate the effectiveness of all models, the data is split

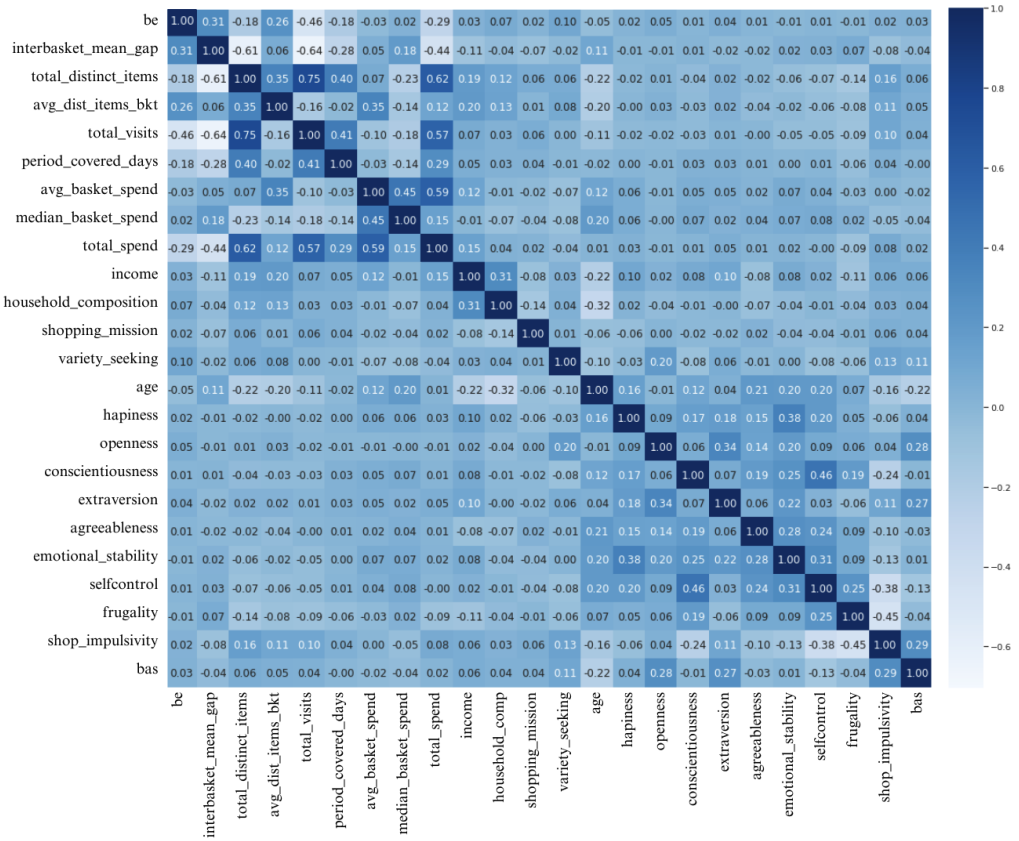


Figure 6.4: Correlation matrix showing relationships between the *bundle entropy* and all the independent variables. We find that multiple features extracted from the grocery shopping data show a significant correlation with *bundle entropy* when performing Pearson's correlation.

into 80% for training and 20% for testing using stratified random sampling (Dahl et al., 2008). To fine-tune each model, the Grid search technique is employed to search through the hyper-parameters. Once the optimal model is identified, it's assessed on the remaining 20% of data. This is a critical step in any prediction task since testing the model on data unseen by the algorithm gives a clear idea of how the model will generalise on *bundle entropy* from other individuals.

Additionally, the models will be trained with a high number of input variables, this is one of the many causes that can contribute to overfitting. Overfitting is a phenomenon where the model exhibits high accuracy in predicting the training dataset yet performs poorly in predicting new data.

Consequently, the model cannot be utilised for generalisation purposes.

To reduce the risk of overfitting, the K-fold cross-validation method is applied and set to K=5 on every iteration (Anguita et al., 2012). This method partitions the dataset into K equal subsets (also known as folds) in which, on each of the iterations (K number of times), one of the folds becomes a temporary test set while the remaining becomes the training set.

On every iteration, the model is trained and the performance is calculated on the hold-out fold until each fold has been used as a validation set exactly once. Hence, there will be K different performance measures. The average of them is calculated to obtain a more robust estimate of the model's performance. Figure 6.5 illustrates a comprehensive explanation of how each fold is used to train the model as well as how the original dataset is split into 80% for training and 20% for the final testing.

The models are evaluated based on three standard criteria: *R-squared*, *MAPE*, and *MAE*. In addition, a baseline metric is required to ensure that the models outperform random chance. To accomplish this, a dummy regressor is implemented to identify the best-performing model.

While the three machine learning algorithms can rank the importance of the variables on the regression task, this study also seeks to interpret the strength and direction of the relationship. Hence, instead of using their associated variable importance approaches, this study applies SHAP values to explore the strength and direction of the associations and MCR to ensure that our findings are consistent and not the cause of random fluctuations in a given model.

After identifying the best-performing model and the variables that impact

the most in predicting *bundle entropy*, the model is trained and tested again, but this time only using the top variables from the SHAP and MCR evaluation. This step evaluates whether removing the variables that seem irrelevant does not decrease the best model's performance.

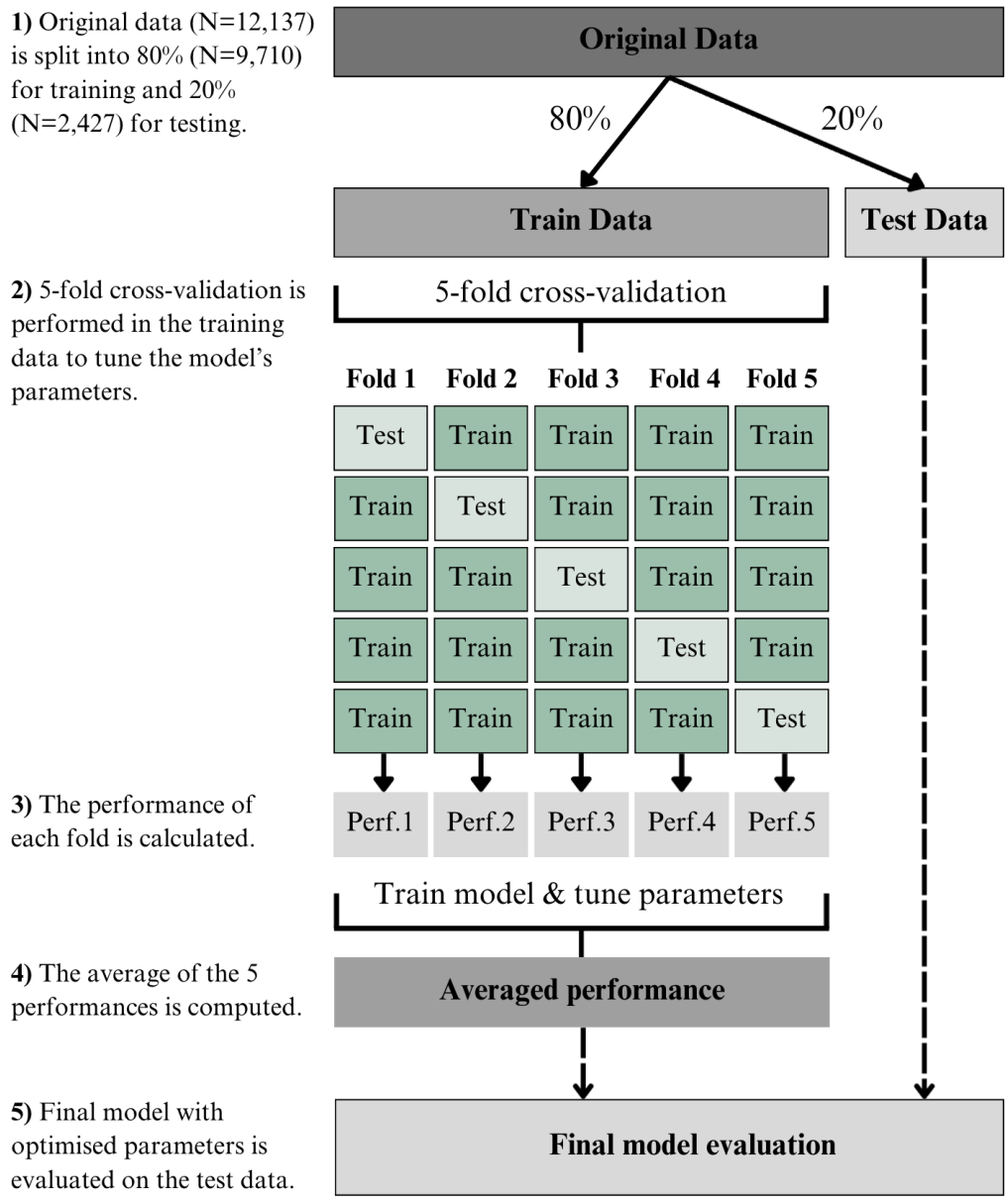


Figure 6.5: Diagram explaining the k-fold cross-validation method used on the training data set.

6.1.6 Results

Systematic Purchase Behaviour regression results

As mentioned in the previous section, the study first explored the performance of a traditional OLS regression. As shown in Table 6.4, different combinations of groups of variables were examined, clearly showing that demographic and psychographic variables by themselves do not perform well in comparison to just using transactional variables. The difference between just using the transactional variables and all the variables is just 0.01. This suggests that demographic and psychographic variables might not be contributing much to predicting *bundle entropy* or that OLS might not be capturing complex relationships between the group of variables.

Table 6.4: Ordinary Least Squares Regression results from different combinations of variables.

Ordinary Least Squares Regression				
Transactional Variables	**Demographic Variable**	**Psychographic Variables**	R^2	**Adjust** R^2
No	Yes	No	0.076	0.074
No	No	Yes	0.038	0.036
No	Yes	Yes	0.097	0.094
Yes	No	No	0.717	0.717
Yes	Yes	Yes	0.727	0.728

The results of predicting *bundle entropy* through different machine learning regressors are presented in Table 6.5. Initially, the table displays the default settings results of the models, highlighting their significant performance improvement compared to a dummy regressor ($R^2 = -0.001$, MAPE = 22.13%, MAE = 0.70). Particularly, LR, XGB, and RF emerge as the top-performing models. As a result, these models underwent further tuning using the Gridsearch method to optimise their parameters for improved performance. The refined results are also presented in Table 6.5, with

both tree-based models standing out as the highest-performing models. XGBoost showed an R^2 of 0.83 (MAPE = 7.78%, MAE = 0.25), while RF exhibited a slightly lower R^2 of 0.82 (MAPE = 7.86%. MAE = 0.26).

Table 6.5 indicates that the LR, XGB, and RF models performed exceptionally well even before tuning the parameters using Grid search and k-fold cross-validation methods. The results remained almost the same even after applying these techniques, which suggests that the models were already optimised. Both RF and XGB models showed an R^2 value of 0.82, while the LR model's performance remained the same with an R^2 value of 0.78. However, these minor variations in performance compared to after parameter tuning can be attributed to several reasons. Firstly, due to computational cost, this study could only explore a limited range of values per parameter for each model, potentially missing the optimal parameter values. Secondly, the variability associated with the folds while performing cross-validation can also impact model performance. Lastly, internal randomness within the machine learning models could also contribute to the variations in performance.

The findings show that all the transactional, demographic and psychographic input variables were generally effective in predicting *bundle entropy*. Additionally, the study examined the processing time of the top three models. The results in Table 6.5 show that LR and RF required only a few seconds to complete the task, while XGB took significantly longer, almost 12 minutes of CPU time. Considering these results and the fact that MCR is only available on RF, it was concluded that RF with its tuned parameter is the best-performing model. Therefore, this model was chosen to explore the relative power of all input variables in predicting *bundle entropy*.

Table 6.5: Results of all the machine learning models in predicting bundle entropy as a measure of systematic purchase behaviour using the merged dataset (transactional + survey)

	Using all input variables						
	Default parameters			**Best parameters (Gridsearch)**			
Model	R^2	MAPE	MAE	R^2	MAPE	MAE	CPU times (min)
Dummy Regressor	-0.001	22.13%	0.70	–	–	–	–
Decision Tree	0.65	10.34%	0.35	–	–	–	–
K-Nearest Neighbours	0.64	11.01%	0.34	–	–	–	–
Support Vector Machines	0.63	10.90%	0.31	–	–	–	–
Linear Regression	0.78	8.64%	0.28	0.78	8.89%	0.29	0.003
XGBoost	0.83	7.67%	0.25	0.82	7.78%	0.26	11.712
Random Forest	0.82	7.83%	0.26	0.82	7.86%	0.26	0.218

Variable importance

To address the research imperative described in section 6.1.3, it is essential to evaluate the relative importance of every input variable used in the developed RF regressor. As previously noted, in addition to the standard technique of ranking each feature, which is permutation importance, the RF regressor was subjected to both SHAP and MCR analyses. These analyses were conducted to assess the relative power of each behavioural, demographic, and psychographic input variable.

Figure 6.6 the permutation importance of the RF regressor, which predicts *bundle entropy*. The 45 input variables are sorted in descending order of importance, with the top variables having a more significant impact on the RF's performance. Notably, the variables related to the total number of visits and the distinct items per basket have a more significant impact on the model's performance.

While permutation importance is helpful for understanding variable importance, it does not provide a direction for the relationship like SHAP values. Figure 6.7 shows the SHAP summary plot for the RF regressor predicting *bundle entropy*. The predictor variables are ranked in descending order of

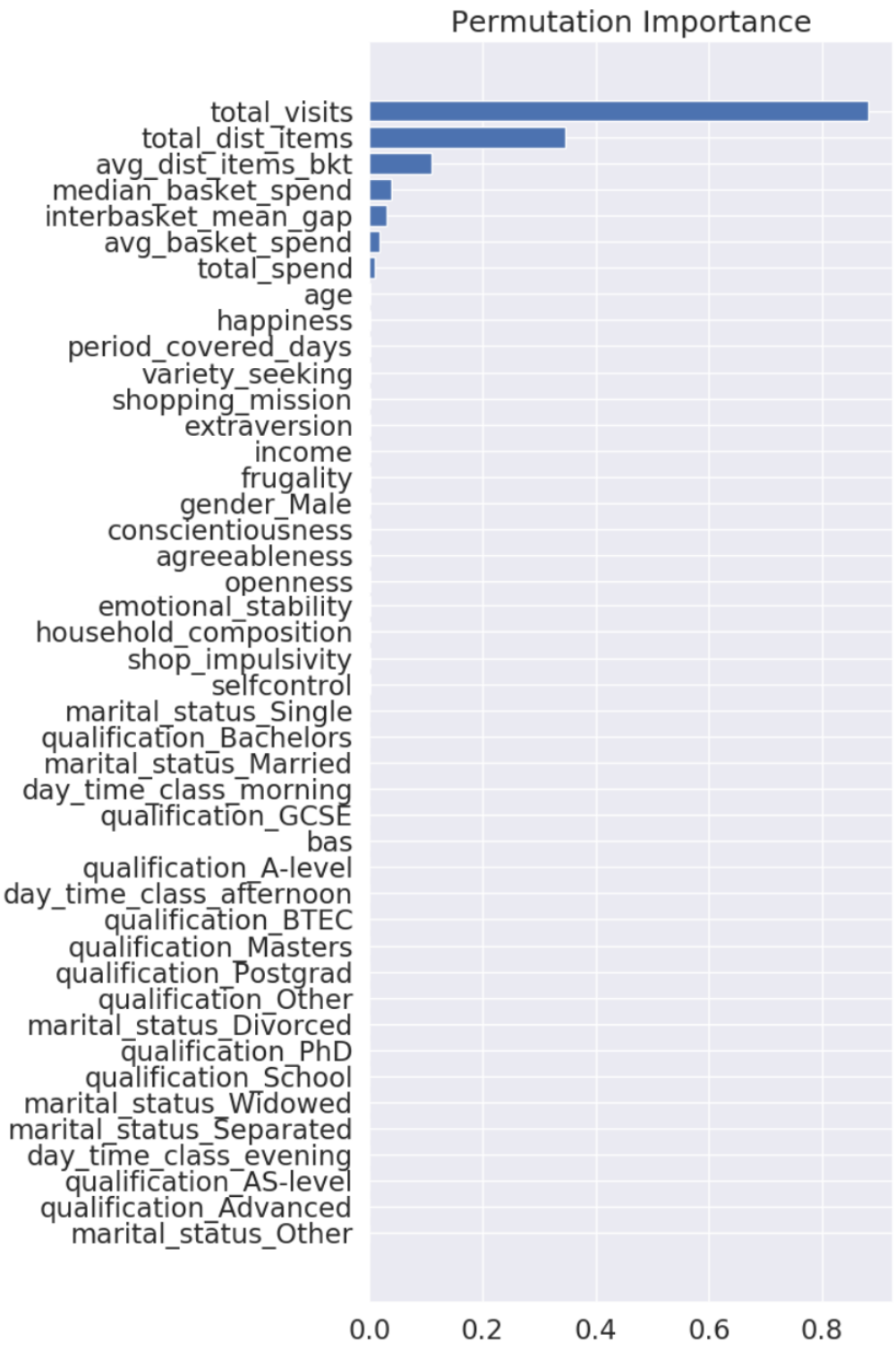


Figure 6.6: Permutation importance for Random Forest regressor predicting *bundle entropy*.

importance, with the same behavioural variables comprising the top 7. The most crucial variable is the number of times a customer visits the store, which is represented as *total visits*. The SHAP plot suggests that higher levels (red colour) of store visits are associated with high *bundle entropy*. In other words, the more an individual purchases, the more likely they will seek alternative or new items.

The second variable that affects *bundle entropy* is the *total distinct items* purchased. This variable measures the number of different items a customer buys across all their purchases. This is an important feature as it provides an idea of how diverse the customer's purchases are. Figure 6.7 indicates that when a customer buys a high number of distinct items, it indicates a higher *bundle entropy*.

Among the demographic variables, just *age* appears to have some level of importance, with younger customers associated with higher *bundle entropy* compared to older customers (see Figure 6.7). Among the top variables, the most important psychographic variable is *variety-seeking*, which aligns with the association between high *bundle entropy* and higher levels of variety. This means that the more an individual seeks variety, the more unsystematic purchase behaviour it will express.

The other variables have small SHAP values compared with the top ones. This means that they have little impact on the model performance. Despite this, some interesting correlations can be observed in Figure 6.7. For instance, the data suggests that males tend to have lower levels of *bundle entropy* than females. Moreover, it appears that individuals who score high on extraversion, which measures their level of sociability, energy, and friendliness, tend to exhibit higher levels of *bundle entropy*. Similarly, the results suggest that individuals who report higher levels of happiness and

openness also tend to exhibit higher levels of *bundle entropy*. Lastly, the data also indicates that households with a larger number of occupants tend to exhibit higher levels of *bundle entropy*.

Figure 6.8 displays the MCR results for RF predicting *bundle entropy* using all the input variables. Each feature's minimum importance is represented as MCR- and visualised as a lollypop on the chart. Only a select few variables have a value above zero, with the *total visits* being the most prominent variable, followed by the *total distinct items* and *average distinct items per basket*. This indicates that these features are essential across all the top-performing models. With less impact on the model's performance are the variables related to spending statistics, such as the *median basket spend*, *average basket spend*, and *total spend*. The last variable with some level of relevance across all RF best-performing models is the *inner basket mean gap*, which represents the average number of days between baskets.

Conversely, variables with MCR- values of zero lack relative predictive power and are thus not necessary in the set of RF best-performing models.

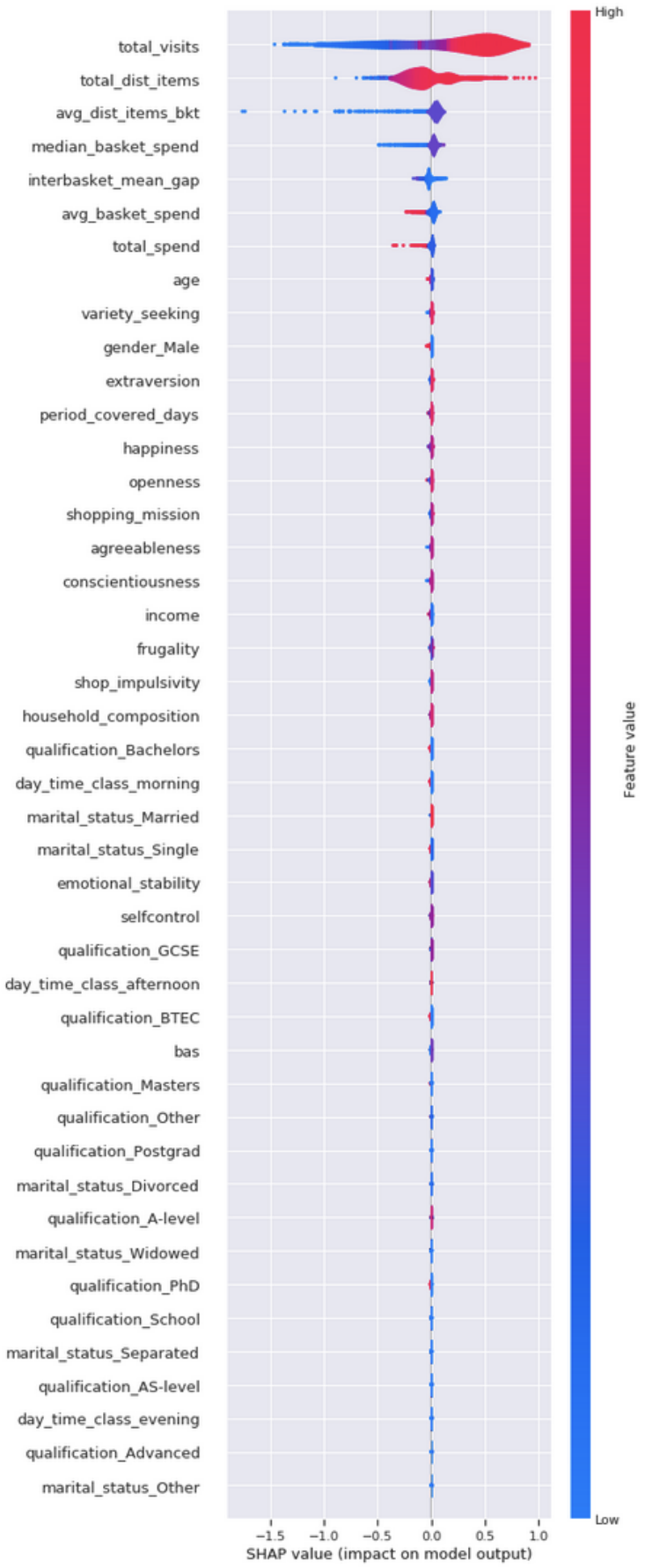


Figure 6.7: SHAP summary plot for RF predicting *bundle entropy* using all input variables.

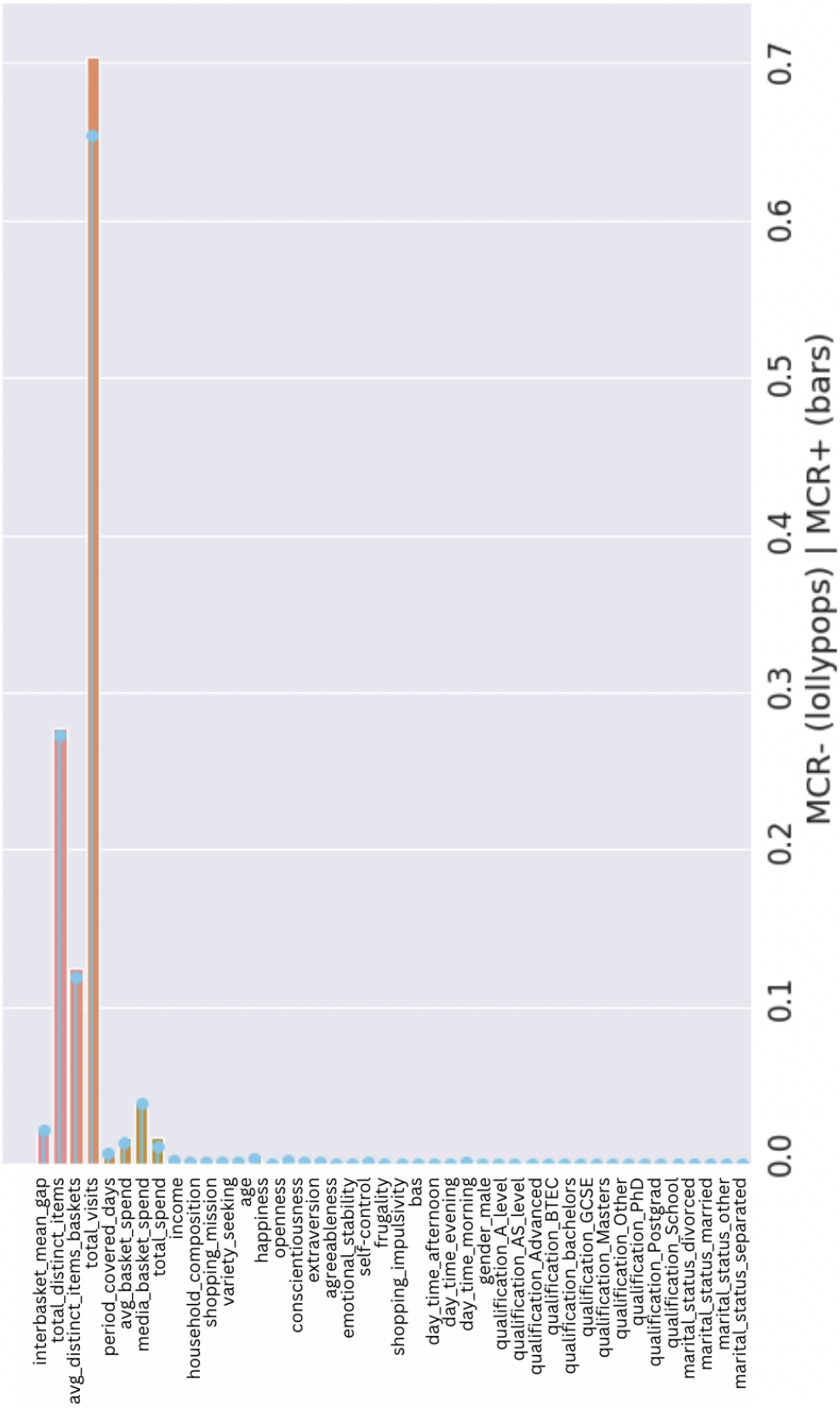


Figure 6.8: MCR chart illustrating feature importance across multiple RF best-performing models for predicting *bundle entropy*

Prediction model using top variables

Both SHAP values and MCR ranked the input variables very similarly. Upon analysing them, it became evident that only a select few variables from the entire set have a significant impact on accurately predicting *bundle entropy*. The remaining variables (mostly demographic and psychographic), on the other hand, have a low or negligible impact. The RF model was then retrained twice to compare two sets of variables. The first time, the RF model was trained using the top three variables, all of which are transactional variables: *total visits*, *total distinct items*, and *average distinct items per basket*. The second time, the RF model was trained using all the demographic and psychographic variables to confirm their relevance in predicting *bundle entropy*. Table 6.6 shows the results of the two approaches.

Table 6.6: Results of the Random Forest model for predicting Systematic Purchase Behaviour using two different sets of variables.

	Input variables					
	Top Transactional			**All demographic and psychographic**		
Model	R^2	MAPE	MAE	R^2	MAPE	MAE
Random Forest	0.93	4.92%	0.16	0.15	20.15%	0.64

Results illustrate a significant difference between the two approaches. The RF model using only the top three transactional variables achieved an R^2 (0.93) even higher than when using all input variables ($R^2 = 0.82$). MAPE and MAE also show better performance. On the contrary, the RF model using all the demographic and psychographic variables only achieved an R^2 of 0.15 with significant errors, MAPE = 20.15%, and MAE = 0.64 (SHAP values for this model can be seen in Appendix I.0.1).

These results suggest that behavioural variables derived from transactional

records are better at predicting *bundle entropy* than demographic and psychographic variables, even if these two are combined together. These results are further discussed in section 6.2.

6.2 Discussion

Until now, very little is known about the characteristics and factors that motivate individuals to behave systematically or unsystematically when it comes to product choices in the retail context.

In section 6.1.3, the study proposed a research imperative. The primary focus of this study was to respond to that imperative using mass transactional data linked to psychographic and demographic predictors of purchase behaviour. The study commenced by showcasing that behavioural variables obtained from vast amounts of transaction records, along with demographic and psychographic variables, can be instrumental in predicting *bundle entropy*. This was reflected by the high performance obtained in all the machine-learning models developed (LR, XGB, and RF). After adjusting their parameters, both the XGB and RF models yielded remarkably similar results, achieving a high R^2 value of 0.82.

6.2.1 Behavioural predictors

In terms of understanding the relative power of the behavioural, demographic, and psychographic variables in predicting *bundle entropy*, several behavioural metrics were identified as important predictors. These variables consistently ranked in the top three in all the variable importance methods applied, including permutation importance, SHAP, and MCR

analysis.

The first variable is the *total visits*, which was shown to be the most important one in predicting *bundle entropy*. From the SHAP results, the more an individual visits the store and purchases, the higher the chances of experiencing high *bundle entropy*. This implies that shoppers who visit the store more frequently tend to purchase a wider variety of items. They are less likely to stick to a particular combination of products and are often on the lookout for new items to try. This is especially true if there are promotions on single products rather than bundles (Mittelman et al., 2014) or if they feel the need to seek variety. This is consistent with earlier research that suggests individuals who make multiple visits to the store tend to seek more variety than those who complete all their shopping in fewer trips (Simonson, 1990). It is also in line with later research that empirically demonstrates that purchase frequency is a significant contributor to explaining variety-seeking (Van Trijp et al., 1996).

The second variable that proved to be highly valuable in predicting *bundle entropy* is the total number of unique items (also known as repertoire size) purchased across all baskets. This suggests that those who have a high number of distinct items in their purchase history tend to be less predictable in their purchasing habits. One possible explanation for this trend is that customers who buy a wider range of items are more likely to encounter external factors, such as situation-specific preferences, or out-of-stock conditions, which may lead them to brand switching (Holbrook, 1984; Van Trijp et al., 1996). Conversely, those who stick to a smaller range of products are less likely to face such situations.

Finally, the average number of distinct items per basket, which is derived from the second-best predictor mentioned above, consistently appears as

the third most important predictor of *bundle entropy*. Although the SHAP values for this variable are not as definitive as those for the previous two predictors, they do suggest that lower values of the average distinct items per basket are associated with lower *bundle entropy*, which means that they tend to be more consistent (repeat purchase) in their item choices over time. This relationship is expected since the more distinct items in your basket, the higher the chances of switching some of them. Previous studies have suggested that individuals tend to seek out maximum enjoyment from their product consumption over time. To achieve this, they balance the feeling of satiation that arises from systematic choices with the stimulation that comes from trying different options (Sevilla et al., 2019). In essence, the chances of maintaining a sense of novelty and excitement while avoiding monotony and boredom increase if the average number of unique items in the basket increases as well. These results are also in line with previous studies where basket sizes have been found to be related to purchase stability (Koll and Plank, 2022).

The MCR analysis corroborates that "Total distinct items" and "Average distinct items per basket" exhibit lower predictive capabilities for *bundle entropy* compared to "Total visits". Nevertheless, it is crucial to incorporate these variables in any model that demonstrates comparable or superior performance to the RF model, as both MCR- and MCR+ render identical values.

Based on the results shown across all the variable importance methods employed in this study, it is evident that behavioural variables are dominant in predicting the *bundle entropy* in comparison to demographic and psychographic variables. This is in line with research done by Bellman et al. (1999), where it was stated that the most important predictors of purchase habits across different retail channels are variables related to past

purchases and not demographics. Ajzen (2002) reinforces this notion that past purchase behaviours account for a significant amount of variance in later behaviours.

These results were further confirmed by training the RF model under two different approaches. The first approach involved using only the top behavioural variables, while the second approach included all the demographic and psychographic variables. The outcomes revealed a significant increase in the model's performance with the first approach, indicating the importance of behavioural variables in predicting the *bundle entropy*. On the other hand, the model's performance declined noticeably with the second approach, emphasising the relatively less important role of demographic and psychographic variables in predicting the *bundle entropy*. Thus, this study highlights the pivotal role played by behavioural variables in predicting the *bundle entropy*, while the significance of demographic and psychographic variables becomes relatively less crucial. This further confirms more recent research that states that behavioural data can usually be a better predictor of purchase behaviour than other predictors, such as demographics (Asniar and Surendro, 2019). This suggests that retailers with access to more purchase records can more accurately forecast the *bundle entropy* of their customers, as these records serve as reliable behavioural indicators.

Finally, based on the SHAP and MCR analysis, the study discovered that by using only the top transactional variables, the RF model's performance could be increased by 10%, leading to an R^2 value of 0.93 (please see Table 6.6). In contrast, the model that solely relied on demographic and psychographic variables experienced a significant drop in its R^2 value, descending from 0.82 to 0.15%. This is in line with literature that has shown some evidence that the contribution of past purchase behaviours can increase

models' performance by up to 50% more (Bosnjak et al., 2007).

The increase in the R^2 value can be attributed to eliminating irrelevant variables that potentially restrained the RF model's training with all variables. By removing these noisy variables, the model could concentrate on the most critical features, namely transactional variables, which resulted in a better fit for the data.

One probable reason for the model's improved performance is the curse of dimensionality, which can significantly hamper models with a high number of noisy variables. This issue is usually overcome by removing some of the irrelevant or noisy variables (Köppen, 2000).

It is essential to consider that while the behavioural variables are vital for the model's prediction accuracy, the demographic and psychographic variables still hold valuable insights into understanding the relationship with *bundle entropy*. Their directional relationship makes them an essential piece of this complex purchase behaviour. These variables are discussed in depth below.

6.2.2 Demographics predictors

According to the study's findings, demographic predictors become less significant when behavioural predictors are available. Among the demographic predictors assessed, the RF regression model revealed that 'Age' was the most significant predictor (see Figure 6.7, 6.8). This suggests that older individuals are more likely to have high *bundle entropy* and, therefore, more predictable than their younger counterparts. These findings are consistent with previous studies that have identified *age* as a key predictor in various consumer behaviours, such as online shopping propensity (Bellman

et al., 1999; Sorce et al., 2006) and product choice behaviour (Rich and Jain, 1968). Moreover, recent studies have found that younger individuals tend to exhibit higher levels of impulsive buying behaviour, making them less predictable than older individuals (Kanwal et al., 2022). Overall, this suggests that retailers who would like to promote new arrivals or test new selling strategies, such as cross-selling or up-selling, should focus on younger individuals.

It was unexpected to find that *gender* was among the top 10 predictors of *bundle entropy* (see Figure 6.7). This is because previous literature suggests that *gender* is usually not a significant determinant of different purchase behaviours, especially when other demographic variables are considered (Bellman et al., 1999; Li and Russell, 1999; Skatova et al., 2019). Numerous studies have indicated that there are differences in impulsive purchasing behaviour between men and women, depending on the situation (Kanwal et al., 2022). When shopping online, men tend to display more impulsive tendencies, whereas in physical stores for fashion apparel, they tend to be less impulsive and more predictable than women. This variance is particularly apparent when compared to younger women. The results of this study are in line with these findings, as this study found that men demonstrate lower levels of *bundle entropy* in physical environments, while women exhibit higher levels of *bundle entropy*.

In regards to the *income* variable, results show that the RF models do not rely much on it to predict *bundle entropy*. Hence, it is considered a weak predictor. This is in line with earlier research (Dubois and Duquesne, 1993; Peters, 1970), which indicates that the correlation between *income* and purchasing behaviour may vary depending on the type of product being considered. For high-involvement or luxury products, *income* is a reliable predictor, especially when it is combined with social class. However, for

low-involvement products, like the ones used in this study, it may not be a significant predictor (Schaninger, 1981).

The direction in the relationship between *income* and *bundle entropy* was anticipated. Figure 6.7 shows that individuals with lower income*incomes* are less systematic in their product choices (high *bundle entropy*) compared to individuals with high *incomes*. Hence, lower-income buyers tend to switch their product choice combinations over time. This might be partially explained because existing research shows that individuals with money restrictions tend to be less loyal to products (Klopotan et al., 2016). Hence, more susceptible to sales and promotion that might affect their systematic behaviour over time. This contradicts the literature that finds that budget restrictions at a certain level can lead to repeat choices (Carlson et al., 2015).

The less important demographic predictors were variables related to individuals' qualifications (e.g. School, Postgraduate, PhD), household size and marital status (e.g. Divorced, Widowed). This corroborates several earlier studies (e.g. Peters1970, Slocum and Mathews1970,Myers et al.1971) that have found that social class, which is usually linked to individuals' qualifications, is not as good a predictor as other demographics, such as *income*. Although later studies found that individuals' education levels are significant predictors, this was only tested to predict *online* purchase behaviour on less than 1,000 participants (Li and Russell, 1999), not offline as in this study. Contrary to expectations, household composition has a low impact on predicting *bundle entropy*. This aligns with prior investigations that have demonstrated a negligible impact of household composition on repurchasing behaviour (Koll and Plank, 2022). Although household composition may not be a strong predictor of *bundle entropy*, larger households tend to have higher *bundle entropy*, as expected based on the correlation

between household size and *bundle entropy*.

6.2.3 Psychographic predictors

Overall results, especially SHAP values, show that psychological traits are better predictors of *bundle entropy* than shopping motivations, with the exception of variety-seeking (see Figure 6.7).

As expected, the direction of the relationship between variety-seeking and *bundle entropy* is direct. Hence, high values of variety-seeking are associated with high *bundle entropy*. The results illustrate that variety-seeking is the most important predictor among the motivational psychological ones. This is supported by the positive correlation between them in the data (see Figure 6.4). A similar relationship has been found in previous studies between variety-seeking and other purchase behaviours, such as impulsive shopping (Sharma et al., 2010b). According to the results, variety-seeking is in the top 10 predictors of *bundle entropy*. However, it is important to note that this predictor may have more relevance at the product level rather than the customer level. In other words, the need for variety may be specific to certain product categories rather than being a general trait of an individual. This idea is supported by the work of Van Trijp et al. (1996) and others (e.g. Mazursky et al., 1987, Givon, 1984), who argues that the desire for variety is a phenomenon that is product category-specific. As a result, an individual may exhibit a high *bundle entropy* for all their purchases, but may not necessarily have the same level of *bundle entropy* for every product category.

According to the study's results, while psychological traits may not play a massive role in predicting *bundle entropy*, extroversion, openness, agree-

ableness, and conscientiousness still hold significance. Figure 6.7 shows that individuals with higher levels of extroversion, openness, agreeableness, and conscientiousness tend to exhibit high *bundle entropy*. The findings are consistent with prior research on related shopping behaviours, which highlights its association with unsystematic product choices. In the case of extroversion, an expected positive relationship was found with *bundle entropy*. This suggests that individuals with extroverted personalities tend to have a lower degree of systematic purchasing habits compared to those who are more introverted. In other words, this suggests that extroverts may be more inclined to make impulsive purchases rather than planned ones, and they are less likely to follow a structured approach while making buying decisions. This negative relationship between extroversion and different consistent purchase behaviours is in line with extant literature. For example, Verplanken and Herabadi (2001) found a negative relationship between extroversion and planned buying in offline contexts, while Lin (2010) found negative associations between extroversion and brand loyalty, repeat purchases, and commitment behaviours (Bove and Mitzifiris, 2007).

The results on agreeableness are contrary to the studies that found a positive relationship with brand loyalty (Lin, 2010; Smith, 2012). This suggests that individuals who exhibit agreeable traits tend to have a positive relationship with brand loyalty. They perceive their interactions with companies as being honest, ethical, and reliable. However, more recent studies suggest that this may not always be the case. New research has revealed that individuals with high levels of agreeableness and openness are more prone to compulsive shopping behaviours (Ali et al., 2022). In other words, those who enjoy exploring novel products and have an open-minded approach towards shopping experiences are more likely to engage in compulsive buying, leading to high *bundle entropy*, which is in line with

the results of this study.

It was expected that conscientious people, who tend to be more organised and reliable, would have a negative relationship with *bundle entropy*. However, the results demonstrated the opposite. Conscientiousness and *bundle entropy* had a positive correlation. While the relationship was not strong (see Figure 6.7), it contrasts with some prior research suggesting a positive relationship between conscientiousness and brand loyalty (Smith, 2012). On the other hand, the correlation matrix (see Figure 6.4) showed a negative relationship between conscientiousness and shopping impulsivity, which is consistent with previous studies on credit card usage among colleague students (Pirog and Roberts, 2007). A potential explanation for the study's findings is that the historical records might have captured individuals' conscientiousness towards budgeting rather than a particular brand, which might affect their product selection depending on various factors such as price, promotions, and coupons.

6.2.4 Practical implications

The study analysed machine learning models' effectiveness in predicting *bundle entropy*. The results showed that these models can account for a minimum of 78% of the variance in *bundle entropy*. Additionally, the study discovered that the model's performance could be significantly improved by identifying the most appropriate model and adjusting its parameters accordingly. As a result of this process, the select RF model's accuracy was boosted to 92%.

The ability to predict *bundle entropy* has practical implications for both academics and practitioners alike. By being able to identify individuals

with high levels of *bundle entropy* behaviour, companies can strategically target sales or promotions on alternative products, as these individuals tend to have highly unsystematic product choices. For example, retailers can tailor exclusive offers, price discounts, promotions, coupons or early access to new arrivals to these individuals as they have been shown to be extroverted, agreeable, and open-minded individuals. Furthermore, the research findings have revealed that higher levels of *bundle entropy* values are associated with young individuals and women. Therefore, businesses looking to promote new products or services should specifically target these groups in their marketing efforts.

In contrast, individuals who exhibit low *bundle entropy* tend to shop for the same products over time, making sales or promotions on alternative or new products pointless. Retailers should, therefore, focus on offering discounts and promotions on the products that these customers already purchase and trust. It was also found that men tend to have lower *bundle entropy* than women. Hence, marketing efforts promoting repeat purchases should target them.

Moreover, retailers can incorporate information on individuals with low *bundle entropy* levels into their supply chain models, inventory, or demand forecast models, as these individuals are highly predictable.

From a methodological perspective, the study demonstrates that the exploratory method of using machine learning regressor models to identify individuals with high or low *bundle entropy* levels is an effective approach that can be extended to investigate other complex purchase behaviour outcomes over time. The results highlight the potential of utilising machine learning models, combined with techniques of variable importance, to gain valuable insights into various aspects of consumer behaviour. By gaining a

deeper understanding of different aspects of purchase behaviour, businesses can gather more information, which can help develop targeted marketing strategies and optimise their offerings to meet the needs of their customers better.

6.3 Conclusion

The findings from this study contribute to the increasing body of evidence that demographic and psychographic variables predict and explain a small proportion of consumer purchase outcomes (e,g., Rich and Jain 1968, Bellman et al. 1999, Sorce et al. 2006). Moreover, it provides empirical evidence that behavioural variables (derived from past purchases) can have a significant impact on predictive power.

These contributions are timely as the extant literature lacks a truly parsimonious model that successfully predicts *bundle entropy* as a consequence of novel linked data sets. SPB is a behavioural outcome, assessed by *bundle entropy*, that reflects consistency in product choices over time, as opposed to static indicators, such as repeat purchases, frequencies, and repertoire size.

This study has responded to the research imperative and showed that machine learning models relying on behavioural variables outperform those relying on demographic and psychographic variables. It also highlights the predictive variables of customers with low *bundle entropy*: being male, being older, spending more but in fewer visits, buying fewer unique items per visit, seeking less variety, being less extroverted, and being less open and agreeable. Notably, many demographic and psychographic variables are irrelevant and may cause noise that harms the model's performance.

According to the results, using the top behavioural variables alone can account for 10% more of the variance of *bundle entropy*, increasing the R^2 value from 0.82 to 0.93. Conversely, a model built solely on demographic and psychographic variables has a much lower R^2 of only 0.15.

From an explanatory viewpoint, the study provides valuable insights into the relationship between personality traits and *bundle entropy*. The findings show that personality traits hold similar significance levels as *age* and *gender* in predicting *bundle entropy*. It further reveals that personality traits are more relevant than education and marital status in this aspect. These findings suggest that despite having low predictive power, personality variables can provide valuable insights into individual characteristics that influence consumer purchase outcomes. A comprehensive understanding of both behavioural and personality traits can enable retailers to customise their marketing strategies and product offerings to better resonate with their intended audience.

Contents

Chapter 7

Discussion & Conclusion

This thesis adopted a positivist approach and implemented three interconnected studies using quantitative methods. It incorporated both synthetic and real-world datasets from various companies and sources (e.g. loyalty card data, nutrition data, survey data), including information ranging from product to individual levels. One of the main objectives of this thesis was to address the growing demand for new measures (Marr, 2015; Guidotti et al., 2015; Rathore, 2018; Netto and Slongo, 2019) capable of handling the vast and diverse datasets available today to uncover complex buying dynamics and support marketing decisions (Mintz et al., 2021) within, but not limited to, the retail context.

This thesis encompasses a wide range of topics and research methodologies, resulting in numerous significant contributions outlined in Table 7.1. It is important to note that this chapter provides a broad discussion of the findings from a general perspective, as each individual study has its own dedicated discussion section where findings and contributions are thoroughly described.

The thesis begins by addressing the primary research objective, which is to establish a precise and effective measure for systematic product choices across multiple purchase baskets. The results demonstrate that the proposed measure (*bundle entropy*) can yield reliable outcomes and provide clear interpretations from both synthetic and real-world transactional data, outperforming existing metrics.

The measure's success with real-world datasets supports the second study's exploration of its application in addressing various research questions related to consumer spending, product choice combinations, and healthy shopping behaviours in both online and offline contexts. The research findings uncover notable and novel disparities in individual shopping behaviour across channels, revealing generally higher spending and more systematic and health-conscious product choices in online shopping compared to offline.

Studies 1 and 2 laid the groundwork for the final study. In this study, traditional statistical and machine learning models and techniques were utilised to explore the relative power of demographic, psychographic, and behavioural factors in explaining *bundle entropy* as a measure of systematic purchase behaviour (SPB). The findings support the idea that previous purchasing patterns provide better insights into future buying behaviours than individuals' internal and demographic characteristics.

The findings also demonstrate the effectiveness of novel variable importance techniques, such as Shapley Additive exPlanations (SHAP) values and Model Class Reliance (MCR), in analysing and predicting entropic behaviour. The final model effectively captures a significant portion of the variance in predicting *bundle entropy* and offers clear insights into the most influential factors. The study highlights the critical need for thorough

testing and transparent explanations of data analysis methods to develop accurate and reliable prediction models that can effectively interpret complex buying behaviours.

The insights obtained from each study played a crucial role in shaping and justifying the subsequent studies within the thesis.

Table 7.1: Table with the contributions of this thesis

No.	Type of contribution	Description	Study
1	New measurement	The study introduces a new measure called *bundle entropy*, which accurately measures systematic purchasing behaviours across multiple baskets.	Study 1a
2	Properties recommendations	The study proposes and establishes new criteria that any measurement of similarity between baskets should adhere to.	Study 1a
3	Significant findings	Results demonstrate that *bundle entropy* outperforms current measurements when applied to real-world data.	Study 1b
4	Practical interpretability	Bundle entropy provides a simple and intuitive method for effectively interpreting its findings and insights.	Study 1b
5	Significant findings	Results demonstrate the potentially hidden purchasing patterns that can be uncovered with the proposed measure.	Study 2
6	Significant findings	The study results demonstrate substantial discrepancies in purchasing behaviour across retail channels within subjects. The findings suggest that the same household spends more, exhibits more consistency in their product choices, and has a tendency to purchase healthier items online rather than offline.	Study 2
7	Practical recommendations	Recommendations for retailers utilising bundle analysis to effectively target both systematic and non-systematic customers.	Study 2
8	New experimental design	New experimental design to model systematic purchasing behaviour and test variable importance using a novel explainability tool, Model Class Reliance for Random Forest, on real-world data.	Study 3
9	Significant findings	Results demonstrate that behavioural variables are better predictors of systematic purchase behaviour compared to demographic and psychographic variables.	Study 3

Overall, the study's findings align with previous discoveries, which have

stated that large transactional data, particularly data from loyalty card members, holds great potential as a valuable source of information for exploring different consumer buying behaviours (Ehrenberg, 1988; Bawa et al., 1989; Van Trijp et al., 1996; Vilcassim and Chintagunta, 1995; Bousofiane, 1996; Smith et al., 2004). Furthermore, the findings contribute to existing knowledge by identifying specific variables (See Figures 6.7 and 6.8) derived from transactional data that have the greatest impact on systematic behaviour.

The studies also validate earlier findings that emphasise the significance of transactional data above other types of data in comprehending basket compositions (Julander, 1992; Mild and Reutterer, 2003), predictability (Guidotti et al., 2015, 2017), and cross-channel purchasing patterns (Chu et al., 2008; Pozzi, 2012). Moreover, the findings extend current knowledge by ranking transactional, demographic and psychographic input variables based on their relative power to predict systematic choices (See Figures 6.7 and 6.8). These findings underscore the importance of such data for both researchers and industry professionals in gaining deeper insights into customer behaviour and tailoring marketing strategies accordingly.

Furthermore, the findings underscore the potential and significance of linking products with their nutritional information to gain direct insights into dietary behaviour from individuals' actual purchases (Eyles et al., 2010; Thomas et al., 2011; Trivedi et al., 2016; Appelhans et al., 2017; Vepsäläinen et al., 2022) rather than relying on self-reported or survey data.

Lastly, the studies contribute towards the utilisation of innovative methodologies (e.g. machine learning) and the integration of multiple large datasets (e.g. *transactional* and *survey* datasets) to develop novel approaches that can uncover hidden drivers of complex buying behaviours and predict them

accurately (Asniar and Surendro, 2019; Noori Hussain et al., 2023; Guidotti et al., 2015; Chu et al., 2010).

7.1 Big data in consumer research

The three research studies shed light on the numerous obstacles and restrictions associated with leveraging big transactional shopping data for consumer research. To begin with, individuals' shopping baskets are notably chaotic, comprising a diverse array of items spanning from perishable goods to non-food products. As a result, advanced data skills are imperative to effectively manage, clean, extract valuable insights, and accurately forecast data in a meticulous manner.

Additionally, the studies highlight the challenges related to how retailers store, manage, and process their data. For instance, in the course of the research, it was noted that certain datasets utilised lacked essential product description details. This absence of information posed a challenge when attempting to match products with their respective nutritional information. Moreover, insufficient information to differentiate between household and individual purchases also posed a significant challenge. These limitations and obstacles are thoroughly examined in the subsequent section.

Despite these challenges, the studies offer valuable insights into consumer behaviour, providing a deeper understanding of purchasing disparities across retail channels, healthy preferences, and the determinants of systematic product choices.

Throughout the three studies, the thesis employs a diverse set of traditional statistical methods, data mining, and machine learning techniques

to explore a wide range of research objectives, inquiries, and necessities surrounding consumer purchasing dynamics. This approach reflects the revolution and potential for these novel methods to complement each other by considering various data sources, leading to a more comprehensive understanding of consumer buying behaviour (Vanhala et al., 2020; Mustak et al., 2021) and ultimately creating competitive advantages (Erevelles et al., 2016).

Study 3 further emphasises the potential of leveraging large datasets and numerous potential explanatory variables in predictive models to enhance knowledge discovery and minimise the impact of unmeasured confounding variables in consumer research (Pearl, 2022). It also illustrates how overfitting, a common issue in predictive models dealing with large datasets, was addressed through the use of cross-validation (refer to Figure 6.5). This method directly assesses the ability of the reported relationships to generalise, ensuring that the effects are not merely a result of the sample's characteristics (Yarkoni and Westfall, 2017). This study underscores the importance of distinguishing between extracting the importance of variables for making predictions and doing so for providing explanations (Grömping, 2009). It involves the selection of a model that can offer further insights into the most relevant drivers, as opposed to choosing a model based solely on its predictive performance.

Overall, this thesis asserts that the analysis of big transactional datasets holds significant potential for comprehending complex consumer purchasing patterns and the driving factors behind them. It underscores the need for further research utilising big data, as we have only begun to tap into its capabilities.

However, it also warns that failing to address issues such as noisy data, out-

liers, overfitting, and highly correlated variables may lead to erroneous or imprecise findings and predictions, ultimately limiting their true potential.

7.1.1 Representativeness in big data

This thesis emphasises the critical role of representative data in effectively addressing the research questions and imperatives examined across the three studies.

The concept of representatives encompasses a wide range of definitions and approaches that vary across different fields. As a result, representatives continue to be a topic of ongoing debate and discussion. Gobo (2004) offers a comprehensive explanation of several approaches to representative sampling in social science research. In the field of environmental studies, representativeness can be defined in various ways depending on how the study is conducted (Warren, 2005). In scientific studies, representative samples are sometimes not recommended and should be avoided (Rothman et al., 2013). A data-driven definition of representativeness is more closely related to the internal and external inclusion criteria for a specific scenario, which can change depending on the context (Corpas and Seghiri, 2010).

The multiplicity of definitions arises from the subjective nature of representativeness, which most cases, is directly tied to the specific question or phenomenon the sample aims to address. From this viewpoint, representativeness is an operational definition, meaning a sample is representative of a specific question but may not be for another (Ramsey and Hewitt, 2005). This perspective aligns with how representativeness is defined in machine learning, where it refers to a proportionate match to a target population, enabling generalizable predictions (Chasalow and Levy, 2021).

This thesis adopted the latter view, where the research questions and imperatives guided the sampling selection across all the studies within this thesis. Consequently, the methodologies employed in each study focus on defining inclusion criteria to identify a sample that accurately reflects the overall phenomenon under examination. The research demonstrates that utilising large transactional datasets allows for the identification of smaller datasets that can accurately estimate individuals' different buying behaviours.

While alternative methods such as digital and non-digital interviews and surveys could have been employed, it is widely acknowledged that individuals may provide inaccurate information and forget their purchases and consumption over time, making these methods less reliable than actual purchase data. Moreover, traditional methods are expensive to implement and pose scalability challenges (Lefever et al., 2007). While online iterations of these methods might be debatable superior (Sethuraman et al., 2005), they suffer from lower response rates and validity compared to offline approaches (Siva et al., 2019). This suggests that relying solely on traditional methods is inadequate for thoroughly grasping the complexities of purchasing behaviour at a national scale and over extended periods. Despite this, traditional methods still hold significant value, as they have the ability to capture information not present in transactional records. Consequently, they can serve to complement other data sources and enrich our understanding of the behaviour under scrutiny, as evidenced in Study 3 (Chapter 6).

In contrast, big transactional data can encompass a wide array of information, including specific behaviours and dietary patterns. However, it is crucial to meticulously define inclusion criteria to ensure the representativeness of these extensive datasets and mitigate biases. Although the sub-samples may constitute only a small fraction of the original data, their

precision and representativeness are pivotal in generating high-quality scientific findings. Access to unprecedented behavioural data presents an opportunity to explore various aspects of consumer behaviour at a national level and over time.

The insights gained from addressing this thesis's research questions and imperatives have paved the way for formulating recommendations, offering guidance on best practices for collecting and analysing large transactional purchase data for consumer research.

7.2 Recommendations

In order to bring together and strengthen the knowledge gained from the separate discussions conducted in each study, the following recommendations are presented as a convenient and efficient way to comprehend the key findings for future research in the covered topics.

These recommendations are intended to acknowledge and optimise the effective utilisation of *bundle entropy* in consumer research. Furthermore, they aim to promote the adoption of large transactional data and emerging datasets to enhance the accuracy and comprehensibility of predictive models for consumer research. Finally, this section provides some methodological recommendations when applying machine learning and techniques of variable importance to big data for consumer research.

Each recommendation is drawn from one or more studies and specifies the primary stakeholders to whom they are most relevant.

1. **Standardise the definitions of product choice predictability among related concepts and measures**

- One of the challenges of this thesis was to search for relevant literature on agreed definitions in regards to an individual product choice behaviour stability over time. The study found that there are several concepts, such as basket predictability, purchase uncertainty, and systematic behaviour, that are closely related. The most recent one, and the one used in this thesis, is systematic purchase behaviour, introduced by Guidotti et al. (2015). However, the author did not provide a clear definition of the concept, leading to confusion, misuse of the measure and biases in decision-making arising from its use. Therefore, future studies should establish clear definitions for any new measure and the parameter, in case they have, since wrong parameterizations can yield diverse findings, potentially impacting business decisions (Briggs, 2006; Mansilla et al., 2022). Study 1 proposed a straightforward definition and intuitive interpretation as a starting point.
- **Study**: 1a.
- **Relevant to**: Researchers.

2. **Establish clear properties when proposing new measures**

   - Basket Revealed Entropy (BRE) is defined as a measure of systematic choices across purchases (Guidotti et al., 2015). However, it was introduced through a purely empirical approach without a solid theoretical foundation. In contrast, *bundle entropy* was developed with clearly defined properties that enable intuitive interpretations. Future research proposing similar or derived measures from *bundle entropy* to assess systematic behaviour should explicitly state their properties, whether in consumer research or other fields. This will ensure transparency

and facilitate the replication of findings, thereby enhancing the measure's reliability.

- **Study**: 1a
- **Relevant**: Researchers

3. **Test new measures on different datasets for real-world applicability**

   - It is crucial to test new measures across diverse datasets to establish their validity, reliability, transparency, and real-world applicability (DeVellis and Thorpe, 2021). Certain measures, such as Key Performance Indicators (KPIs), are integral to most companies and institutions. Therefore, it is essential to rigorously test future measures to ensure their quality and credibility, ultimately leading to valuable conclusions. In this thesis, *bundle entropy* was tested on four different data sets, one synthetic and three real-world shopping data sets. Additionally, it is important to compare new measures against benchmarks to evaluate potential advantages and broader use.
   - **Study**: 1a, 1b, 2.
   - **Relevant**: Researchers, academics, and practitioners.

4. **Further explore the real-world applicability of *bundle entropy***

   - By assigning a predictability score to a household or consumer, it is possible to integrate it into retail segmentation, descriptive, and predictive analytics (Wen et al., 2018). This creates greater possibilities for personalising responses and offers to customers. The framing and messaging of direct-to-consumer marketing communications are increasingly informed by behavioural

and propensity scores, ensuring that communication is consistent with consumer needs or demands. Moreover, with the advance in technology and the business opportunity that they offer, consumer buying behaviours are more dynamic than ever. Hence, there is an increase need for measures that intent to capture and inform complex buying patterns, especially over time. Overall, retailers can enhance customer satisfaction and loyalty by adopting more insightful measures than traditional measures. This thesis argues that *bundle entropy* holds significant potential for comprehending diverse social behaviours across various fields of study that are still to be explored.

- **Study**: 1b, 2.
- **Relevant**: Researchers, academics, practitioners, and policymakers.

5. **From big data to smaller data: Strategies for improving data quality through representative sampling**

    - As discussed previously, real-world big transactional data is inherently noisy, and determining the appropriate sample size is critical for obtaining meaningful results. The studies in this thesis employed a consistent approach to select the most representative sample possible. While there is no standardised process for sample selection, we argue that certain steps are indispensable for obtaining a representative sample:

        A) **Formulate a clear research question**: This is crucial as it guides the selection of the most relevant data (Chasalow and Levy, 2021). Without a clear research question, it is impossible to determine which data will yield the most effective results.

B) **Establish clear data-driven inclusion criteria**: These criteria should be aligned with the research question (Corpas and Seghiri, 2010). Depending on the stringency of the inclusion criteria, a substantial portion of the raw data will be eliminated. However, it is important to recognise that high-quality data (data that represents the behaviour under study) leads to more accurate results than noisy data.

C) **Identify potential outliers**: Thoroughly searching for potential outliers is crucial, as their presence could significantly impact the analyses if left unidentified.

D) **Sanity check**: Once the sample data is selected, conducting a sanity check through data exploration might yield valuable insights that prompt a reassessment of the inclusion criteria. Think of it as an iterative process.

- **Study**: 1a, 1b, 2, 3.
- **Relevant**: Researchers.

6. **Within subject marketing strategies**

- Personalised marketing involves tailoring marketing efforts to individuals with similar characteristics and behaviours (Arora et al., 2008). However, the results from Study 2 reveal significant variations in buying behaviours depending on the retail setting (online vs. offline). This finding supports previous studies with the idea that companies should integrate considerations of their various offering channels into their segmentation processes (Melis et al., 2015; Andrews and Currim, 2004). The findings extend current knowledge by demonstrating that the same household/individual tends to exhibit more consistent healthy behaviours in the online environment as compared to offline.

This suggests the potential for developing customised strategies that accommodate individual dynamics and predictability across both online and offline shopping channels. As a result, marketing efforts can be enhanced to better align with consumer preferences and behaviours.

- **Study**: 2
- **Relevant**: Researchers, academics, and practitioners.

7. **Machine learning models and explainability techniques considerations when modelling consumer behaviour**

    - Study 3 showed the great potential of utilising innovative machine learning and explainability tools to effectively model consumer systematic behaviour. Despite their promise, it's important to carefully consider a few factors when working with these new tools. Before any analysis or modelling, it is crucial to thoroughly process the data, this includes but is not limited to ensuring consistency across different datasets, performing data cleansing (e.g. missing values and anomalies), transforming necessary variables (e.g. normalising, encoding), reducing variables if necessary, sampling the data ensuring representativeness, among others to ensure overall data quality and ultimately robust machine learning models. Additionally, it is essential to use proper model specification and evaluation techniques, such as K-fold cross-validation and out-of-sample testing, to prevent overfitting and ensure that emerging findings can be applied to a wider population. Additionally, baseline models are essential for assessing the performance of the models being tested and for avoiding underperforming models.
    - **Study**: 3.

- **Relevant**: Researchers, academics, and practitioners.

7.3 Limitations

7.3.1 The nature of big transactional data

The accessibility to massive amounts of transactional data is a relatively recent development. While it is widely acknowledged that increasing the amount of data can enhance our comprehension of various behaviours, the rapid expansion in the size of these data sets poses challenges for consumer research. However, recent and ongoing research has been addressing some of these challenges and limitations (Clarke et al., 2021; Jenneson et al., 2022, 2023; Rains and Longley, 2021; Mansilla et al., 2024a). As this thesis relies heavily on large transactional datasets as the primary source of data, it is important to acknowledge the inherent limitations associated with these datasets. These limitations include, but are not limited to:

- Differentiating between household and individual purchases.
- Purchase of items not consumed.
- Situational purchases (e.g. birthdays, visits of relatives with different dietary requirements).
- Purchases for someone else (e.g. care support workers).
- Inconsistency in loyalty card usage.
- Loyalty cards being used by multiple individuals.
- Assessing whether the store is the main store where consumers make the majority of their purchases.

- And dealing with biased demographics of card-holders.

7.3.2 Time period covered by the data sets

Some of the datasets utilised in this thesis are more dated than others, yet they still offer significant value to the research, providing a robust foundation for reliable conclusions. The comprehensive analysis conducted with these datasets allows for meaningful interpretations and valuable findings. Nevertheless, due to the unavailability of more recent data, future research could explore whether the observed patterns still hold true, especially in the post-pandemic era. This further exploration would help determine the continued relevance and applicability of the findings in a rapidly changing retail landscape.

7.3.3 Scope and size of the sample data

While all the studies had access to large transactional data, the studies only made use of small samples from the raw data set. This is attributed to the previous discussion about the challenges of sampling selection and representativeness in datasets rife with real-world noise. Despite their limited size, these samples effectively represented individual buying and choice behaviour relevant to their respective research questions. Additionally, the studies highlighted that individual loyalty card data, when combined with products' nutritional information and surveys containing demographic and psychographic details, offered valuable insights into captured shopping behaviour. However, it is important to note that the limitation of a small sample size may not be representative when the data collection is scaled. Furthermore, while national sales data captures the entire customer popula-

tion of the store, the absence of linked individual demographics for studies 1 and 2 introduces uncertainties about the composition of the captured population. This lack of information limits the ability to generalise the findings of studies 1 and 2 to specific consumer segments, potentially overlooking variations in purchasing behaviour and preferences related to demographic and psychographic factors.

Lastly, it is essential to acknowledge that this research did not explore external co-variates that could potentially influence the patterns observed in bundled product choices across the studies. Factors such as marketing promotions, seasonality, or regional differences could significantly shape consumer behaviour, yet this study did not explicitly examine them.

7.3.4 Nutritional data access

Specifically, Study 2 delved into the relationship between bundled product choices and their average healthiness. The analysis uncovered significant differences depending on whether the shopping was conducted in person or online. It is important to note that the analysis was limited to soft drinks due to the availability of nutritional information in the retailer's Application Programming Interfaces (API) for that specific category. Therefore, the findings may not hold true for other product categories, and the observed differences may not accurately represent broader household product choices.

7.3.5 Survey data

Study 3 utilised a comprehensive survey dataset to delve into the psychological factors influencing individuals within the transactional data. It is

crucial to note that the survey was originally designed to evaluate the psychological factors impacting various purchase behaviour outcomes rather than focusing specifically on understanding systematic purchase behaviour. Despite this limitation, the dataset serves as a reliable and extensive source of psychological information about individuals. It has proven instrumental in uncovering valuable insights into the psychological and demographic characteristics of individuals with both high and low systematic purchase behaviour.

7.4 Future Research Directions

Various measures can be utilised to estimate the predictability of shopping baskets and systematic choices, which are interconnected concepts. The predictability of an individual's shopping choices is directly associated with the systematic nature of their selections. However, selecting the appropriate measure is not always straightforward and depends on the research objective and the behaviour being measured.

In Study 1, the focus was on developing a clear and intuitive measure to assess how predictable an individual's product choices are across their purchases. There are existing measures that address this phenomenon, but their definition of predictability varies, leading to differences in approach. For example, measures like BRE look for an exact match at the sub-basket level (Guidotti et al., 2015), while Basket Level Entropy (BLE) looks for an exact match at the basket level (Nicolas-Sans and Ibáñez, 2021) and Item Entropy (IE) at the item level. In this research, predictability is based on the extent to which it is possible to predict the composition of a basket or sub-basket. *Bundle entropy* can be used to measure systematic product

choices across multiple baskets, considering product choices at different levels, such as complete baskets or bundles of items. However, there is uncertainty regarding the extent of its application and consistency across different settings, given the nature of its development and the datasets in which it was tested. Therefore, future studies should explore various applications where *bundle entropy* can be useful in the retail context but also in consumer research in general. For example, it could be used to assess diet consistency from transactional data over time and to evaluate the consistency of nutrient intake based on product choices. The unique characteristics of transactional data and the potential to link it with other datasets, such as demographics, psychographics, and nutritional information, offer numerous possibilities for applying *bundle entropy*. Thus, future studies should also incorporate more varied types of datasets to ensure the measure's applicability, reliability and consistency across distinct settings.

Bundle entropy, as a direct and easily interpretable measure, has the potential to be compared with other direct measures, extending beyond those explored in Study 2 (products' nutritional value). In the future, studies could investigate the connection between SPB (via *bundle entropy*) and measures of food and nutrient deprivation. Similarly, future studies could also explore the relationship between SPB and obesity among adults and students. Furthermore, various studies have developed models to predict child obesity (George, 2002; Long et al., 2023), respiratory diseases (Dolan et al., 2023a), ovarian cancer (Dolan et al., 2023b), and plastic bag usage (Lavelle-Hill et al., 2020) using transactional data and purchase patterns. Therefore, *bundle entropy* could also be used as an additional input variable to measure product choice uncertainty, potentially improving predictions since a measure of choice uncertainty is not present in those models.

Study 2 offers fresh insights into individuals' consistent healthy choices,

both online and offline. However, as previously mentioned in the limitations, these healthy patterns were only assessed for soft drink products due to the lack of nutrient information for other product categories. Future studies could explore different sources of nutritional information for products to align them and determine if the findings of Study 2 differ across product categories, as suggested by Van Trijp et al. (1996). In the UK, there are several databases containing nutritional information for millions of products, such as the UK Composition of Foods Integrated Data Set[1] and Nutritics[2]. The challenge lies in the fact that each source has its own standards, often requiring the creation of specific rules or heuristics to align products. Nonetheless, some studies have attempted this massive task and successfully aligned several product categories for specific retailers' databases (Long et al., 2023). Therefore, future studies could examine those methodologies and evaluate consistent healthy choices across other product categories, providing new insights for academics, marketers, and food policymakers.

Findings from Study 3 opened the door to explore whether the main drivers of systematic purchase behaviour are the same in the online setting as they are in the offline one. With online and offline channels offering different customer experiences, it might be that the drivers of systematic behaviour that did not have relevant predictive power, such as psychographic and demographic variables, might do in the online setting. Additionally, the top drivers of SPB, as shown in Study 3, took into account all product categories. Future explorations should evaluate if these high-performance predictors remain stable or change across product categories or even departments.

[1]https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid

[2]https://www.nutritics.com/en/

Study 3 is equipped with an extensive 45 predictors, which greatly enhances its predictive capabilities. It is worth noting that factors such as price discounts, promotional initiatives, discount vouchers, and bundled deals have long been acknowledged for their ability to influence consumer behaviour and habits (Van Heerde et al., 2003; Yan et al., 2017; Haans and Gijsbrechts, 2011; Fader and Lodish, 1990). Therefore, future research could enrich Study 3 by integrating these variables, which were not available for the current study, and evaluating their overall predictive power when combined with historical purchasing behaviours, demographic data, and psychographic variables.

Supply-side factors, which encompass a range of elements, including product availability, pricing strategies, and promotional campaigns, influence consumer purchasing behaviours and, therefore, play a key role in systematic purchase patterns. For instance, when products are out of stock or when supply is constrained in certain regions, consumers may alter their typical buying behaviours, leading to unexpected shifts in purchasing patterns.

Additionally, strategically timed promotions can create an illusion of predictability, enticing consumers to buy in larger quantities or at specific times, which may not reflect their usual habits. These insights underscore the essential need to integrate supply-side data into the analysis of purchasing behaviour whenever possible. By examining metrics such as shelf availability, regional pricing variations, and detailed promotional calendars, researchers and marketers can gain a comprehensive understanding of the external forces at play. This analysis not only enhances the interpretation of consumer decision-making but also sheds light on the underlying predictability in consumer behaviour that may otherwise remain obscured.

Future research could further address these complexities by developing models that explicitly incorporate supply-side variables into the analysis of SPB. Such work could investigate the extent to which external conditions mediate or amplify SPB.

7.4.1 Advance in data analytics for consumer research

The rapid advancement of data science and data analytics methods has not only revolutionised various fields but also significantly impacted consumer research. These innovative methods have enabled in-depth analysis of new data sets, facilitating the capture of extensive and real-time consumer behaviour data (Erevelles et al., 2016). However, there are still challenges, such as the need for new methods to effectively link loyalty card data to different complementary datasets like national aggregations and nutrient information. This is necessary to produce precise and meaningful outputs regarding complex purchasing patterns, hidden correlations, and consumer well-being. This area of work represents an important opportunity to fully utilise integrated datasets in consumer research (Dekimpe, 2020). Key priorities for future research methods include:

- In recent years, predictive models have emerged as powerful tools, not only for making accurate predictions but also for providing valuable insights. Different machine learning techniques such as the ones used in this thesis (MCR (Fisher et al., 2019; Smith et al., 2020) and SHAP values (Rodríguez-Pérez and Bajorath, 2020)) have been combined to enhance model interpretability and identify crucial variables for further actions (Ljevar et al., 2021; Dolan et al., 2023a; Long et al., 2023). However, since many models developed within consumer research aim to explain individuals' behaviours rather than

just predict them, additional advancements are needed in this area to make machine learning models a more effective alternative for explaining model predictions. Some of these advances include variable importance methods capable of tracking changes in the relevance of variables over time and detecting feature drift. Additionally, there is a need for the development of variable importance models that consider Rashomon sets (sets of alternative models that fit equally well on the same data) for other machine learning models apart from the existing ones (Fisher et al., 2019; Smith et al., 2020; Gunasekaran et al., 2022).

- In the limitations section, it was highlighted that while some studies have suggested methods for addressing the inherent limitations of large transactional data in specific contexts (Jenneson et al., 2022; Rains and Longley, 2021), there is still a need for further progress in establishing standardised procedures to mitigate these limitations and enhance the reliability of transactional data, such as loyalty card data and digital foot prints in general.

- Many retailers do not maximise the potential of their data, often overlooking the collection of relevant information. By gaining a comprehensive understanding of data capabilities, retailers can invest in enhancing data collection, management, and processing. This will ensure that their data becomes a reliable and accurate resource for improving decision-making and driving progress across various business areas. For instance, retailers could standardise product names and descriptions, ensure accurate product categorisation, reduce redundancies, and collect complete product nutrient information. The latter has been shown to be useful in enhancing the understanding of various social behaviours, related consequences, and diseases (Man-

silla et al., 2024a; Long et al., 2023; Clark et al., 2021; Jenneson et al., 2023). Moreover, these standardizations could be implemented across retailers to combine their datasets, improving representativeness and enabling more impactful research for social good.

Advancing in the mentioned points will significantly contribute to a deeper understanding of consumer behaviour. Retailers can utilise these potential findings to enhance various aspects of their business and improve the overall customer experience in multiple ways. Customers who recognise the benefits and enhance their overall experience might be motivated to use their loyalty cards more frequently, potentially increasing the number of new loyalty card members. Overall, this will enable data collection from a broader range of people, thus reducing data skewness and sparsity, which have been recognised as common challenges with big shopping data (Chen et al., 2013; Hsu et al., 2004).

7.5 Conclusion

In conclusion, the thesis utilised a positivist approach and applied quantitative methodologies across three interconnected studies on consumer buying behaviour. It introduced a new metric, *bundle entropy*, to evaluate systematic purchase behaviour over time and explored its potential applications and drivers. The research also suggests clear properties that every measure should accord when measuring predictability across purchases. The research also aimed to promote the usability of big transactional datasets and novel data analytics technologies due to their huge potential to uncover novel hidden patterns for consumer research.

The findings revealed that the proposed measure surpassed existing metrics

in terms of accuracy and consistency across multiple transactional data sets, establishing it as a reliable tool for measuring individual systematic choices. This measure can identify different combinations of products systematically bought together across purchases and provide a single score for intuitive interpretation.

Furthermore, cross-reference analyses demonstrated the measure's usability in understanding how uncertainty in product choices relates to other behaviours, such as healthy choices, across retail settings. This has significant applications in customer segmentation, improving target marketing efforts and customer experience.

Each of the three studies outlined the practical implications of the findings and methods used, contributing to an improved understanding of consumer buying patterns and predictability. The research also discussed practical implications for business and marketing decision-making, highlighting the novel insights that can be derived from effectively combining transactional data with complementary data sets.

The thesis also provided tangible recommendations and considerations to guide future efforts focused on integrating novel data sources into consumer behaviour research. Additionally, it discussed the limitations presented in each of the studies and proposed potential future works to develop the ideas presented further. Ultimately, the thesis encouraged the exploration of novel methods from different fields, such as data mining, advanced analytics, and machine learning, to enhance and uncover hidden patterns in current and emerging big data sets rich with behavioural information.

In summary, this study not only contributes to the comprehension of consumer purchasing patterns by introducing the *bundle entropy* metric and its real-world applications but also showcases and encourages the imple-

mentation and the power of using big data and advanced data analytics for a more comprehensive understanding of consumer behaviour dynamics.

Bibliography

Abdi, H. (2007). Kendall Rank Correlation Coefficient. *The Concise Encyclopedia of Statistics*, 2:508–510.

Ailawadi, K. L., Pauwels, K., and Steenkamp, J.-B. E. M. (2008). (electronic) Private-Label Use and Store Loyalty. *Journal of Marketing*, 72:19–30.

Ajzen, I. (2002). Residual Effects of Past on Later Behavior: Habituation and Reasoned Action Perspectives. *Personality and Social Psychology Review*, 6(2):107–122.

Akaika, H. (1985). *Prediction and entropy*. Springer US.

Al Hamli, S. S. and Sobaih, A. E. E. (2023). Factors Influencing Consumer Behavior towards Online Shopping in Saudi Arabia Amid COVID-19: Implications for E-Businesses Post Pandemic. *Journal of Risk and Financial Management*, 16(1).

Alexander, P. J. (1997). Product variety and market structure: A new measure and a simple test. *Journal of Economic Behavior and Organization*, 32(2):207–214.

Ali, F., Zubair, M., and Ali, A. (2022). The Big Five dyad congruence and compulsive buying : A case of service encounters. *Journal of Retailing and Consumer Services*, 68(March):103007.

Allenby, G. M. and Lenk, P. J. (1994). Modeling household purchase behavior with logistic normal regression. *Journal of the American Statistical Association*, 89(428):1218–1231.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

André, Q., Carmon, Z., Wertenbroch, K., Crum, A., Frank, D., Goldstein, W., Huber, J., van Boven, L., Weber, B., and Yang, H. (2018). Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data. *Customer Needs and Solutions*, 5(1-2):28–37.

Andrews, R. and Currim, I. (2004). Behavioural differences between consumers attracted to shopping online versus traditional supermarkets: implications for enterprise design and marketing strategy. *International Journal of Internet Marketing and Advertising*, 1(1):38–61.

Anesbury, Z., Nenycz-Thiel, M., Dawes, J., and Kennedy, R. (2016). How do shoppers behave online? An observational study of online grocery shopping. *Journal of Consumer Behaviour*, 15(3):261–270.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., and Ridella, S. (2012). The 'K' in K-fold Cross Validation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 441–446.

Appelhans, B. M., French, S. A., Tangney, C. C., Powell, L. M., and Wang, Y. (2017). To what extent do food purchases reflect shoppers' diet quality and nutrient intake? *International Journal of Behavioral Nutrition and Physical Activity*, 14(1):1–10.

Apperley, L. J., Blackburn, J., Erlandson-Parry, K., Gait, L., Laing, P., and

Senniappan, S. (2022). Childhood obesity: A review of current and future management options. *Clinical Endocrinology*, 96(3):288–301.

Arasu, B. S., Seelan, B. J. B., and Thamaraiselvan, N. (2020). A machine learning-based approach to enhancing social media marketing. *Computers & Electrical Engineering*, 86:106723.

Arboleda-Florez, M. and Castro-Zuluaga, C. (2023). Interpreting direct sales' demand forecasts using SHAP values. *Production*, 33.

Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., Joshi, Y., Kumar, V., Lurie, N., Neslin, S., and others (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters*, 19:305–321.

Ascarza, E., Iyengar, R., and Schleicher, M. (2016). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research*, 53(1):46–60.

Asniar and Surendro, K. (2019). Predictive analytics for predicting customer behavior. *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, pages 230–233.

Aydinli, A., Lamey, L., Millet, K., ter Braak, A., and Vuegen, M. (2021). How Do Customers Alter Their Basket Composition When They Perceive the Retail Store to Be Crowded? An Empirical Study. *Journal of Retailing*, 97(2):207–216.

Barber, N., Kuo, P. J., Bishop, M., and Goodman, R. (2012). Measuring psychographics to assess purchase intention and willingness to pay. *Journal of Consumer Marketing*, 29(4):280–292.

Basu, S., McKee, M., Galea, G., and Stuckler, D. (2013). Relationship of soft drink consumption to global overweight, obesity, and diabetes: A cross-national analysis of 75 countries. *American Journal of Public Health*, 103(11):2071–2077.

Baumeister, R. F. (2002). Yielding to Temptation: Self-Control Failure, Impulsive Purchasing, and Consumer Behavior. *Journal of Consumer Research*, 28(4):670–676.

Bawa, K., Landwehr, J. T., and Krishna, A. (1989). Consumer Response to Retailers' Marketing Environments: An Analysis of Coffee Purchase Data. *Journal of Retailing*, 65(4):471–495.

Behe, B. K., Huddleston, P. T., Childs, K. L., Chen, J., and Muraro, I. S. (2020). Seeing through the forest: The gaze path to purchase. *Plos one*, 15(10):e0240179.

Belk, R. W. (1986). What Should ACR Want to Be When It Grows Up? *Advances in Consumer Research*, 13:423–424.

Bell, D. R. and Lattin, J. M. (1998a). Shopping behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. *Marketing Science*, 17(1):66–88.

Bell, D. R. and Lattin, J. M. (1998b). Shopping behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. *Marketing Science*, 17(1):66–88.

Bellman, S., Lohse, G. L., and Johnson, E. J. (1999). Predictors of online buying behavior. *Communications of the ACM*, 42(12):32–38.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

Source: Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300.

Berger, J., Draganska, M., and Simonson, I. (2007). The influence of product variety on brand perception and choice. *Marketing Science*, 26(4):460–472.

Bhattacharya, C. B. (1997). Is your brand's loyalty too much, too little, or just right?: Explaining deviations in loyalty from the Dirichlet norm. *International Journal of Research in Marketing*, 14(5):421–435.

Bian, R., Murray-Tuite, P., and Wolshon, B. (2023). Predicting grocery store visits during the early outbreak of COVID-19 with machine learning. *Transportation Research Record*, 2677(4):79–91.

Boone, D. S. and Roehm, M. (2002). Retail segmentation using artificial neural networks. *International journal of research in marketing*, 19(3):287–301.

Bosnjak, M., Galesic, M., and Tuten, T. (2007). Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach. *Journal of Business Research*, 60(6):597–605.

Boussofiane, S. (1996). Exploiting data analysis in customer loyalty. *JOURNAL OF TARGETING MEASUREMENT AND ANALYSIS FOR MARKETING*, 5:11–19.

Bove, L. and Mitzifiris, B. (2007). Personality traits and the process of store loyalty in a transactional prone context. *Journal of Services Marketing*, 21(7):507–519.

Boztuğ, Y. and Reutterer, T. (2008). A combined approach for segment-specific market basket analysis. *European Journal of Operational Research*, 187(1):294–312.

Bray, G. A., Nielsen, S. J., and Popkin, B. M. (2004). Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *American Journal of Clinical Nutrition*, 79(4):537–543.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Breimann, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. *Pacific Grove, Wadsworth*.

Briggs, R. (2006). Marketers who measure the wrong thing get faulty answers. *Journal of Advertising Research*, 46(4):462–468.

Brunelle, C. and Grossman, H. (2022). Predictors of online compulsive buying : The role of personality and mindfulness. *Personality and Individual Differences*, 185(August 2021):111237.

Bryman, A. (2016). *Social research methods*. Oxford university press.

Brynjolfsson, E. and Smith, M. D. (2000). Frictionless commerce? A comparison of Internet and conventional retailers. *Management Science*, 46(4):563–585.

Budescu, D. V. and Budescu, M. (2012). How to measure diversity when you must. *Psychological Methods*, 17(2):215–227.

Campo, K., Lamey, L., Breugelmans, E., and Melis, K. (2021). Going Online for Groceries : Drivers of Category-Level Share of Wallet Expansion. *Journal of Retailing*, 97(2):154–172.

Candrian, C. and Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, 134.

Carlson, K. A., Wolfe, J., Blanchard, S. J., Huber, J. C., and Ariely, D. (2015). The budget contraction effect: How contracting budgets lead to less varied choice. *Journal of Marketing Research*, 52(3):337–348.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Chang, C. (2011). The Effect of the Number of Product Subcategories on Perceived Variety and Shopping Experience in an Online Store. *Journal of Interactive Marketing*, 25(3):159–168.

Chasalow, K. and Levy, K. (2021). Representativeness in statistics, politics, and machine learning. In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–89. Association for Computing Machinery, Inc.

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., and Zhou, X. (2013). Big data challenge: A data management perspective. *Frontiers of Computer Science*, 7(2):157–164.

Chen, S.-S., Choubey, B., and Singh, V. (2021). A neural network based price sensitive recommender model to predict customer choices based on price effect. *Journal of Retailing and Consumer Services*, 61:102573.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., and others (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chen, Y., Liu, H., Wen, Z., and Lin, W. (2023). How Explainable Machine Learning Enhances Intelligence in Explaining Consumer Purchase Behavior: A Random Forest Model with Anchoring Effects. *Systems*, 11(6).

Chintagunta, P. K. (1999). Variety seeking, purchase timing, and the 'lightning bolt' brand choice model. *Management Science*, 45(4):486–498.

Chintagunta, P. K., Dubé, J.-P., and Singh, V. (2003). Balancing Profitability and Customer Welfare in a Supermarket Chain. *Quantitative Marketing and Economics*, 1:111–147.

Chow, S.-Y., Chen, C.-W., and Chang, W.-S. (2004). A study of relationship between human personality and brand personality: sports shoes as an example. *Chung Hua Journal of Management*, 5(3):1–16.

Chu, J., Arce-Urriza, M., Cebollada-Calvo, J., and Chintagunta, P. (2010). An Empirical Analysis of Shopping Behavior Across Online and Offline Channels for Grocery Products: The Moderating Effects of Household and Product Characteristics. *Journal of Interactive Marketing*, 24(4):251–268.

Chu, J., Chintagunta, P., and Cebollada, J. (2008). A comparison of within-household price sensitivity across online and offline channels. *Marketing Science*, 27(2):283–299.

Cimana, E. (2013). *Online Grocery Shopping in Sweden : Identifying key factors towards consumer's inclination to buy food online*. PhD thesis.

Clark, S. D., Shute, B., Jenneson, V., Rains, T., Birkin, M., and Morris, M. A. (2021). Dietary patterns derived from UK supermarket transaction data with nutrient and socioeconomic profiles. *Nutrients*, 13(5):1–21.

Clarke, H., Clark, S., Birkin, M., Iles-Smith, H., Glaser, A., and Morris, M. A. (2021). Understanding barriers to novel data linkages: topic modeling of the results of the LifeInfo survey. *Journal of medical Internet research*, 23(5):e24236.

Conolly, A., Craig, S., and Gebert, S. (2019). Health Survey for England 2018 overweight and obesity in adults and children. *London: Health and Social Care Information Centre*.

Corpas, G. and Seghiri, M. (2010). Size matters: A quantitative approach to corpus representativeness. Technical report.

Costa, P. T. and McCrae, R. R. (1992). *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Dahl, F. A., Grotle, M., Saltyte Benth, J., and Natvig, B. (2008). Data splitting as a countermeasure against hypothesis fishing: With a case study of predictors for low back pain. *European Journal of Epidemiology*, 23(4):237–242.

Danaher, P. J., Wilson, I. W., and Davis, R. A. (2003). A Comparison of Online and Offline Consumer Brand Loyalty. *Marketing Science*, 22(4):461–476.

Degeratu, A. M., Rangaswamy, A., and Wu, J. (2000). Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *International Journal of Research in Marketing*, 17(1):55–78.

Dekimpe, M. G. (2020). Retailing and retailing research in the age of big data analytics. *International Journal of Research in Marketing*, 37(1):3–14.

Deng, X., Kahn, B. E., Unnava, H. R., and Lee, H. (2016). "Wide" Variety: Effects of Horizontal Versus Vertical Display on Assortment Process-

ing, Perceived Variety, and Choice. *Journal of Marketing Research*, 53(5):682–698.

DeVellis, R. F. and Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Di Crosta, A., Ceccato, I., Marchetti, D., la Malva, P., Maiella, R., Cannito, L., Cipi, M., Mammarella, N., Palumbo, R., Verrocchio, M. C., Palumbo, R., and Domenico, A. D. (2021). Psychological factors and consumer behavior during the COVID-19 pandemic. *PLoS ONE*, 16(8 August):1–23.

Dolan, E., Goulding, J., Marshall, H., Smith, G., Long, G., and Tata, L. J. (2023a). Assessing the value of integrating national longitudinal shopping data into respiratory disease forecasting models. *Nature Communications*, 14(1):7258.

Dolan, E. H., Goulding, J., Tata, L. J., and Lang, A. R. (2023b). Using Shopping Data to Improve the Diagnosis of Ovarian Cancer: Computational Analysis of a Web-Based Survey. *JMIR cancer*, 9(1):e37141.

Donnelly, R., Ruiz, F. J., Blei, D., and Athey, S. (2021). Counterfactual inference for consumer choice across many product categories. *Quantitative Marketing and Economics*, 19(3-4):369–407.

Dorfman, R. (1979). A formula for the Gini coefficient. *The review of economics and statistics*, pages 146–149.

Downs, S. M., Bloem, M. Z., Zheng, M., Catterall, E., Thomas, B., Veerman, L., and Wu, J. H. (2017). The Impact of Policies to Reduce trans Fat Consumption: A Systematic Review of the Evidence. Technical report.

Droomer, M. and Bekker, J. (2020). Using machine learning to predict the next purchase date for an individual retail customer. *South African Journal of Industrial Engineering*, 31(3):69–82.

Duarte, V., Zuniga-Jara, S., and Contreras, S. (2022). Machine Learning and Marketing: A Systematic Literature Review.

Dubois, B. and Duquesne, P. (1993). The market for luxury goods: income versus culture. *European Journal of Marketing*, 27(1):35–44.

Ehrenberg, A. S. (1988). Repeat-Buying. In *Repeat-Buying: facts, theory and applications*, pages 31–78.

Erevelles, S., Fukawa, N., and Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, 69(2):897–904.

Evans, M. (1999). Food retailing loyalty schemes–and the Orwellian Millennium. *British Food Journal*, 101(2):132–147.

Eyles, H., Jiang, Y., and Ni Mhurchu, C. (2010). Use of Household Supermarket Sales Data to Estimate Nutrient Intakes: A Comparison with Repeat 24-Hour Dietary Recalls. *Journal of the American Dietetic Association*, 110(1):106–110.

Fader, P. S. and Lodish, L. M. (1990). A Cross-Category Analysis of Category Structure and Promotional Activity for Grocery Products. *Journal of Marketing*, 54(4):52.

Fader, P. S. and Schmittlein, D. C. (1993). Excess Behavioral Loyalty for High-Share Brands: Deviations from the Dirichlet Model for Repeat Purchasing. *Journal of Marketing Research*, 30(4):478–493.

Faraway, J. J. and Augustin, N. H. (2018). When small data beats big data. *Statistics and Probability Letters*, 136:142–145.

Fisher, A., Rudin, C., and Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. (Vi).

Food Standards Agency (2006). The Nutrient Profiling Model London: Food Standards Agency.

Food Standards Agency (2007). Front of pack (FoP) nutrition label for pre-packed products sold through retail outlets. Technical report, Food Standards Agency.

Foxall, G. R. (2001). Foundations of consumer behaviour analysis. *Marketing Theory*, 1(2):165–199.

Frisbie, G. A. (1980). Ehrenberg's Negative Binomial Model Applied to Grocery Store Trips. *Journal of Marketing Research*, 17(3):385–930.

Froehlich, J. and Krumm, J. (2008). Route Prediction from Trip Observations. In *Society of Automotive Engineers (SAE) 2008 World Congress, April 2008*.

Garg, N., Wansink, B., and Inman, J. J. (2007). The influence of incidental affect on consumers' food intake. *Journal of Marketing*, 71(1):194–206.

George, B. (2002). The relationship between lottery ticket and scratch-card buying behaviour, personality and other compulsive behaviours. *Journal of Consumer Behaviour: An International Research Review*, 2(1):7–22.

Ghafoor, A., Dean, A., and Abbas, N. (2015). Impact of Demographic Factors on Impulse Buying Behavior of Consumers in Multan-Pakistan. Technical Report 22.

Gil, R., Korkmaz, E., and Sahin, O. (2020). Can free-shipping hurt online retailers? *Quantitative Marketing and Economics*, 18(3):305–342.

Givon, M. (1984). Variety Seeking through Brand Switching. *Marketing Science*, 3(1):1–22.

Gobo, G. (2004). Sampling, representativeness and generalizability. *Qualitative research practice*, 405:426.

Goodhardt, G., Ehrenberg, A., and Chatfield, C. (1984). The Dirichlet : A Comprehensive Model of Buying Behaviour. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):621–655.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Green, P. E. (1977). A new approach to market segmentation. *Business Horizons*, 20(1):61–73.

Grewal, D., Iyer, G. R., and Levy, M. (2004a). Internet retailing: Enablers, limiters and market consequences. *Journal of Business Research*, 57(7):703–713.

Grewal, D. and Levy, M. (2007). Retailing research: Past, present, and future. *Journal of Retailing*, 83(4):447–464.

Grewal, D., Levy, M., and Lehmann, D. R. (2004b). Retail Branding and Customer Loyalty: An overview. *Journal of Retailing*, 80(4).

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319.

Grömping, U. (2015). Variable importance in regression models.

Guadagni, P. M. and Little, J. D. C. (2008). A logit model of brand choice calibrated on scanner data. *Marketing Science*, 27(1):29–48.

Guidotti, R., Coscia, M., Pedreschi, D., and Pennacchioli, D. (2015). Behavioral entropy and profitability in retail. *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, pages 1–10.

Guidotti, R., Gabrielli, L., Monreale, A., Pedreschi, D., and Giannotti, F. (2018). Discovering temporal regularities in retail customers' shopping behavior. *EPJ Data Science*, 7(1).

Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F., and Pedreschi, D. (2017). Market basket prediction using user-centric temporal annotated recurring sequences. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2017-Novem:895–900.

Gunasekaran, A., Chen, M., Hill, R., and McCabe, K. (2022). Method Agnostic Model Class Reliance (MAMCR) Explanation of Multiple Machine Learning Models. In *International Conference on Soft Computing and its Engineering Applications*, pages 56–71.

Guo, L.-J. (2003). The effects of personality trait and brand personality on brand preference. *Graduate Institute of Management Science, National Chiao Tung University, Hsinchu*, pages 1–10.

Gupta, S., Chintagunta, P., Kaul, A., and Wittink, D. R. (1996). Do Household Scanner Data Provide Representative Inferences from Brand Choices: A Comparison with Store Data. *Journal of Marketing Research*, 33(4):383–398.

Haans, H. and Gijsbrechts, E. (2011). One-deal-fits-all? On Category Sales Promotion Effectiveness in Smaller versus Larger Supermarkets. *Journal of Retailing*, 87(4):427–443.

Hair, J. F. (2009). Multivariate data analysis.

Harris, L. (1986). A transaction data study of weekly and intradaily patterns in stock returns. *Journal of Financial Economics*, 16(1):99–117.

Hart, S., Smith, A., Sparks, L., and Tzokas, N. (1999). Are loyalty schemes a manifestation of relationship marketing? *Journal of marketing management*, 15(6):541–562.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Henry, P. (2002). Systematic variation in purchase orientations across social classes. *Journal of Consumer Marketing*, 19(5):424–438.

Holbrook, M. B. (1984). Situation-specific ideal points and usage of multiple dissimilar brands. *Research in Marketing*.

Hossain, M. A., Akter, S., and Yanamandram, V. (2020). Revisiting customer analytics capability for data-driven retailing. *Journal of Retailing and Consumer Services*, 56.

Hsu, C.-N., Lavrač, N., Motoda, H., and Fawcett, T. (2004). Mining Skewed and Sparse Transaction Data for Personalized Shopping Recommendation. Technical report.

Huang, Y. and Oppewal, H. (2006). Why consumers hesitate to shop online: An experimental choice analysis of grocery shopping and the role of delivery fees. *International Journal of Retail and Distribution Management*, 34(4-5):334–353.

Huber, J. and Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, 36(4):1420–1438.

Hult, G. T. M., Sharma, P. N., Morgeson, F. V., and Zhang, Y. (2019). Antecedents and Consequences of Customer Satisfaction: Do They Differ Across Online and Offline Purchases? *Journal of Retailing*, 95(1):10–23.

Huyghe, E., Verstraeten, J., Geuens, M., and Van Kerckhove, A. (2017). Clicks as a healthy alternative to bricks: How online grocery shopping reduces vice purchases. *Journal of Marketing Research*, 54(1):61–74.

Islam, J. U., Rahman, Z., and Hollebeek, L. D. (2017). Personality factors as predictors of online consumer engagement : an empirical investigation. *Marketing Intelligence & Planning*.

Jadhav, V. and Khanna, M. (2016). Factors influencing online buying behavior of college students: A qualitative analysis. *Qualitative Report*, 21(1):1–15.

Jani, D. and Han, H. (2014). Personality, satisfaction, image, ambience, and loyalty: Testing their relationships in the hotel industry. *International Journal of Hospitality Management*, 37:11–20.

Javed Awan, M., Mohd Rahim, M. S., Nobanee, H., Yasin, A., and Khalaf, O. I. (2021). A big data approach to black friday sales. *MJ Awan, M. Shafry, H. Nobanee, A. Yasin, OI Khalaf et al.," A big data approach to black friday sales," Intelligent Automation & Soft Computing*, 27(3):785–797.

Jenneson, V., Clarke, G. P., Greenwood, D. C., Shute, B., Tempest, B., Rains, T., and Morris, M. A. (2022). Exploring the geographic variation in fruit and vegetable purchasing behaviour using supermarket transaction data. *Nutrients*, 14(1).

Jenneson, V., Greenwood, D. C., Clarke, G. P., Rains, T., Tempest, B.,

Shute, B., and Morris, M. A. (2023). Supermarket Transaction Records in Dietary Evaluation: The STRIDE study: Validation against self-reported dietary intake. *Public Health Nutrition*, 26(12):2663–2676.

John, O. P. and Srivastava, S. (1999). The Big Five trait taxonomy.

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Julander, C.-R. (1992). Basket analysis: a new way of analysing scanner data. *International Journal of Retail \& Distribution Management*, 20(7).

Jung, B. Y., Choi, M. S., Youn, H. Y., and Song, O. (2010). Vertical handover based on the prediction of mobility of mobile node. In *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2010*, pages 534–539.

Kahle, L. R., Beatty, S. E., and Homer, P. (1986). Alternative measurement approaches to consumer values: the list of values (LOV) and values and life style (VALS). *Journal of consumer research*, 13(3):405–409.

Kahn, B. E., Kalwani, M. U., and Morrison, D. G. (1986). Measuring Variety-Seeking and Reinforcement Behaviors Using Panel Data. *Journal of Marketing Research*, 23(2):89.

Kahn, B. E. and Lehmann, D. R. (1991). Modeling Choice Among Assortments. *Journal of Retailing*, 67(3):274–299.

Kanwal, M., Burki, U., Ali, R., and Dahlstrom, R. (2022). Systematic review of gender differences and similarities in online consumers' shopping behavior. *Journal of Consumer Marketing*, 39(1):29–43.

Kaur, M. and Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, 85(Cms):78–85.

Khodabandehlou, S. and Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, 19(1-2):65–93.

Kim, E., Kim, W., and Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems*, 34(2):167–175.

Kim, J. and Forsythe, S. (2008). Adoption of virtual try-on technology for online apparel shopping. *Journal of Interactive Marketing*, 22(2):45–59.

Kirchler, E. (1988). Household Economic Decision Making. In van Raaij, W. F., van Veldhoven, G. M., and Wärneryd, K.-E., editors, *Handbook of Economic Psychology*, pages 258–292. Springer Netherlands, Dordrecht.

Kirchler, E. (1995). Studying economic decisions within private households: A critical review and design for a couple experiences diary. *Journal of Economic Psychology*, 16(3):393–419.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, 1(1):1–12.

Klassen, M. L. and Glynn, K. A. (1992). Catalog loyalty. *Journal of Direct Marketing*, 6(3):60–67.

Klopotan, I., Vrhovec-Žohar, K., and Mahič, E. (2016). Impact of Income on Customers' Loyalty: Are Customers with Higher Income more Loyal? *Business Systems Research Journal*, 7(1):81–88.

Koll, O. and Plank, A. (2022). Do shoppers choose the same brand on the

next trip when facing the same context? An empirical investigation in FMCG retailing. *Journal of Retailing*.

Köppen, M. (2000). The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8.

Koschate-Fischer, N., Cramer, J., and Hoyer, W. D. (2014). Moderating Effects of the Relationship Between Private Label Share and Store Loyalty. *Journal of Marketing*, 78:69–82.

Krumm, J. (2010). Ubiquitous advertising: The killer application for the 21st century. *IEEE Pervasive Computing*, 10(1):66–73.

Kuhn, T. S. (1997). *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago.

Kulkarni, G., Ratchford, B. T., and Kannan, P. K. (2012). The Impact of Online and Offline Information Sources on Automobile Choice Behavior. *Journal of Interactive Marketing*, 26(3):167–175.

Larson, J. S., Bradlow, E. T., and Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of research in Marketing*, 22(4):395–414.

Lavelle-Hill, R., Goulding, J., Smith, G., Clarke, D. D., and Bibby, P. A. (2020). Psychological and demographic predictors of plastic bag consumption in transaction data. *Journal of Environmental Psychology*, 72(July):101473.

Lavelle-Hill, R., Smith, G., Mazumder, A., Landman, T., and Goulding, J. (2021). Machine learning methods for "wicked" problems: exploring the complex drivers of modern slavery. *Humanities and Social Sciences Communications*, 8(1):1–11.

Lavelle-Hill, R. E., Smith, G., and Murayama, K. (2023). Machine Learning Meets Traditional Statistical Methods in Psychology: Challenges and Future Directions. *OSF Preprints*, 30.

Lawlor, L. R. (1980). Overlap, similarity, and competition coefficients. *Ecology*, 61(2):245–251.

Lefever, S., Dal, M., and Matthíasdóttir, (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology*, 38(4):574–582.

Leo, B. (2001). Random forests. *Random Forests*, pages 1–122.

Li, H. and Russell, M. G. (1999). The Impact of Perceived Channel Utilities, Shopping Orientations, and Demographics on the Consumer's Online Buying Behavior. *Journal of Computer-Mediated Communication*, 5(2).

Li, J., Pan, S., Huang, L., and Zhu, X. (2019). A machine learning based method for customer behavior prediction. *Tehnicki Vjesnik*, 26(6):1670–1676.

Lian, J. W. and Yen, D. C. (2014). Online shopping drivers and barriers for older adults: Age and gender differences. *Computers in Human Behavior*, 37:133–143.

Lin, C. H. and Lin, H. C. (2009). The effect of mood states on variety-seeking behavior: The moderating role of price promotion. *Social Behavior and Personality*, 37(10):1307–1312.

Lin, L. Y. (2010). The relationship of consumer personality trait, brand personality and brand loyalty: An empirical study of toys and video games buyers. *Journal of Product & Brand Management*, 19(1):4–17.

Ljevar, V., Goulding, J., Smith, G., and Spence, A. (2021). Using model class reliance to measure group effects on non-adherence to asthma medication. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708.

Lo, C., Frankowski, D., and Leskovec, J. (2016). Understanding behaviors that lead to purchasing: A case study of pinterest. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 531–540. Association for Computing Machinery.

Lobstein, T. and Davies, S. (2009). Defining and labelling healthy and unhealthy food. *Public Health Nutrition*, 12(3):331–340.

Long, G., Hogg, M. K., Hartley, M., and Angold, S. J. (1999). Relationship marketing and privacy: exploring the thresholds. *Journal of Marketing Practice: Applied Marketing Science*, 5(1):4–20.

Long, G., Nica-Avram, G., Harvey, J., Mansilla, R., Welham, S., Lukinova, E., and Goulding, J. (2023). Predicting health related deprivation using loyalty card digital footprints. *International Journal of Population Data Science*, 8(3).

Lundberg, S. M., Allen, P. G., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems*, 30.

Ma, L. and Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3):481–504.

Mackenzie, N. and Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in educational research*, 16(2):193–205.

Mansilla, R., Long, G., Welham, S., Harvey, J., Lukinova, E., Nica-Avram, G., Smith, G., Salt, D., Smith, A., and Goulding, J. (2024a). Detecting iodine deficiency risks from dietary transitions using shopping data. *Scientific Reports*, 14(1).

Mansilla, R., Smith, A., Smith, G., and Gouilding, J. (2024b). The relative power of behavioural, demographic, and psychographic variables as predictors of systematic purchase behaviour. In *British Academy of management*.

Mansilla, R., Smith, A., Smith, G., and Goulding, J. (2024c). Systematic purchase behaviour and healthy choices across online and offline channels: Insights from transactional data. *Journal of Business Research*.

Mansilla, R., Smith, G., Smith, A., and Goulding, J. (2022). Bundle entropy as an optimized measure of consumers' systematic product choice combinations in mass transactional data. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1044–1053. IEEE.

Marr, B. (2015). *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*. John Wiley \& Sons.

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., and Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3):588–596.

Mathews, H. L. and Slocum Jr, J. W. (1969). Social Class and Commercial Bank Credit Card Usage. *Journal of Marketing*, 28:71–78.

Matzler, K., Bidmon, S., and Grabner-Kräuter, S. (2006). Individual determinants of brand affect: the role of the personality traits of ex-

traversion and openness to experience. *Journal of product & brand management*, 15(7):427–434.

Mazursky, D., Labarbera, P., and Aiello, A. (1987). When consumers switch brands. *Psychology & Marketing*, 4(1):17–30.

McAlister, L. and Pessemier, E. (1982). Variety Seeking Behavior: An Interdisciplinary Review. *Journal of Consumer Research*, 9(3):311.

McCrae, R. R. and Costa, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36(3):587–596.

McDonald, W. J. (1993). The roles of demographics, purchase histories, and shopper decision-making styles in predicting consumer catalog loyalty. *Journal of Direct Marketing*, 7(3):55–65.

Melis, K., Campo, K., Breugelmans, E., and Lamey, L. (2015). The Impact of the Multi-channel Retail Mix on Online Store Choice: Does Online Experience Matter? *Journal of Retailing*, 91(2):272–288.

Mild, A. and Reutterer, T. (2003). An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. *Journal of Retailing and consumer Services*, 10(3):123–133.

Mintz, O., Gilbride, T. J., Lenk, P., and Currim, I. S. (2021). The right metrics for marketing-mix decisions. *International Journal of Research in Marketing*, 38(1):32–49.

Mitchell, A. (1984). Nine American Lifestyles: Values and Societal Change. *Futurist*, 18(4):4–14.

Mittal, V. and Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of marketing research*, 38(1):131–142.

Mittelman, M., Andrade, E. B., Chattopadhyay, A., and Miguel Brendl, C. (2014). The offer framing effect: Choosing single versus bundled offerings affects variety seeking. *Journal of Consumer Research*, 41(4):953–964.

Morrison, M., Gan, S., Dubelaar, C., and Oppewal, H. (2011). In-store music and aroma influences on shopper behavior and satisfaction. *Journal of Business Research*, 64(6):558–564.

Moshrefjavadi, M. H., Rezaie Dolatabadi, H., Nourbakhsh, M., Poursaeedi, A., and Asadollahi, A. (2012). An Analysis of Factors Affecting on Online Shopping Behavior of Consumers. *International Journal of Marketing Studies*, 4(5).

Mowen, J. C. (2000). *The 3M model of motivation and personality: Theory and empirical applications to consumer behavior*. Springer Science & Business Media.

Mulyono, K. B. and Rusdarti (2020). How psychological factors boost compulsive buying behavior in digital era: A case study of Indonesian students. *International Journal of Social Economics*, 47(3):334–349.

Mustak, M., Salminen, J., Plé, L., and Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, 124(January 2020):389–404.

Myers, J. H., Stanton, R. R., and Huag, A. F. (1971). Correlates of Buying Behavior: Social Class vs . Income. *Journal of Marketing*, 35:8–15.

Nauenberg, E., Basu, K., and Chand, H. (1997). Hirschman–Herfindahl index determination under incomplete information. *Applied Economics Letters*, 4(10):639–642.

Nayyar, R. and Gupta, S. L. (2011). Determinants of Internet Buying Behavior in India. *Asian Journal of Business Research*, 1(2).

Netto, C. F. S. and Slongo, L. A. (2019). Marketing metrics, big data and the role of the marketing department. *Revista de Administração da Universidade Federal de Santa Maria*, 12(3):527–543.

Ngai, E. W. and Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research*, 145:35–48.

Nica-Avram, G., Harvey, J., Smith, G., Smith, A., and Goulding, J. (2021). Identifying food insecurity in food sharing networks via machine learning. *Journal of Business Research*, 131:469–484.

Nicolas-Sans, R. and Ibáñez, D. G. (2021). Customer basket heterogeneity: how to measure it and some possible business applications. *Economic Research-Ekonomska Istrazivanja*, 34(1):2572–2592.

Nikolova, H. D. and Inman, J. J. (2015). Healthy Choice: The Effect of Simplified Point-of-Sale Nutritional Information on Consumer Food Choice Behavior. *Journal of Marketing Research*, 52(6):817–835.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. *Lecture Notes in Engineering and Computer Science*, 2202:380–384.

Noori Hussain, H., Yousif Alabdullah, T. T., Ahmed, E. R., and M. Jamal, K. A. (2023). Implementing Technology for Competitive Advantage

in Digital Marketing. *International Journal of Scientific and Management Research*, 06(06):95–114.

Office for National Statistics, U. (2023). The cost of living, current and upcoming work: February 2023. Technical Report June.

Organization, W. H. (2000). Obesity: preventing and managing the global epidemic: report of a WHO consultation.

Organization, W. H. and others (2017). Obesity and overweight.

Panda, S. K. and Dwivedi, M. (2020). Minimizing Food Wastage Using Machine Learning: A Novel Approach. In *Smart Innovation, Systems and Technologies*, volume 159, pages 465–473. Springer.

Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., and Zingaretti, P. (2020). Deep understanding of shopper behaviours and interactions using RGB-D vision. *Machine Vision and Applications*, 31:1–21.

Pathak, A., Gupta, K., and McAuley, J. (2017). Generating and personalizing bundle recommendations on steam. In *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1073–1076. Association for Computing Machinery, Inc.

Pearl, J. (2022). Comment: understanding Simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl*, pages 399–412.

Peters, W. H. (1970). Relative Occupational Class Income: A Significant Variable in the Marketing of Automobiles. *Journal of Marketing*, 34(2):74.

Pirog, S. F. and Roberts, J. A. (2007). Personality and credit card misuse among college students: The mediating role of impulsiveness. *Journal of Marketing Theory and Practice*, 15(1):65–77.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Pozzi, A. (2012). Shopping cost and brand exploration in online grocery. *American Economic Journal: Microeconomics*, 4(3):96–120.

Punj, G. (2011). Effect of Consumer Beliefs on Online Purchase Behavior: The Influence of Demographic Characteristics and Consumption Values. *Journal of Interactive Marketing*, 25(3):134–144.

Puntoni, S., Reczek, R. W., Giesler, M., and Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing*, 85(1):131–151.

Putrevu, S. . and Ratchford, B. T. (1997). A model of search behavior with an application to grocery shopping. Technical report.

Rahim, H. L., Abidin, Z. Z., and Khairuddin, N. N. (2014). Psychographic Characteristics Influencing Customer Behaviour on Online Purchase Intention. *Aust. J. Basic & Appl. Sci.*, 8(April):248–253.

Rains, T. and Longley, P. (2021). The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services*, 63.

Ramsey, C. A. and Hewitt, A. D. (2005). A methodology for assessing sample representativeness. *Environmental Forensics*, 6(1):71–75.

Rana, S. and Tirthani, J. (2012). Effect of education, income and gender on impulsive buying among Indian consumer an empirical study of readymade garment customers. *Management*, 1(12):145–146.

Ratchford, B., Soysal, G., Zentner, A., and Gauri, D. K. (2022). Online and offline retailing: What we know and directions for future research. *Journal of Retailing*, 98(1):152–177.

Rathore, B. (2018). Metaverse Marketing: Novel Challenges, Opportunities, and Strategic Approaches. *International peer reviewed/refereed academic multidisciplinary journal*, 07(02):72–82.

Ratner, R. K. and Kahn, B. E. (2002). The impact of private versus public consumption on variety-seeking behavior. *Journal of Consumer Research*, 29(2):246–257.

Rayner, M., Scarborough, P., Heart, B., and Health, F. (2009). The UK Ofcom Nutrient Profiling Model. Technical report, Food Standards Agency, London, UK.

Rich, S. U. and Jain, S. C. (1968). Social Class and Life Cycle as Predictors of Shopping Behavior. *Journal of Marketing Research*, 5:41–9.

Rivera, R., Amorim, M., and Reis, J. (2021). Technological Evolution in Grocery Retail: A Systematic Literature Review. In *Iberian Conference on Information Systems and Technologies, CISTI*. IEEE Computer Society.

Robertson, T. S. and Kennedy, J. N. (1968). Prediction of consumer innovators: Application of multiple discriminant analysis. *Journal of Marketing Research*, 5(1):64–69.

Rodríguez-Pérez, R. and Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of computer-aided molecular design*, 34(10):1013–1026.

Rothman, K. J., Gallacher, J. E., and Hatch, E. E. (2013). Why representativeness should be avoided. *International Journal of Epidemiology*, 42(4):1012–1014.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography*, 3(3):268–273.

Russell, G. J. and Kamakura, W. A. (1997). Modeling multiple category brand preference with household basket data. *Journal of Retailing*, 73(4):439–461.

Russell, G. J. and Petersen, A. (2000a). Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3):367–392.

Russell, G. J. and Petersen, A. (2000b). Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3):367–392.

Saberi, S., Kouhizadeh, M., Sarkis, J., and Shen, L. (2019). Blockchain technology and its relationships to sustainable supply chain management. *International journal of production research*, 57(7):2117–2135.

Sandy, C., Gosling, S., and Durant, J. (2013). Predicting Consumer Behavior and Media Preferences: The Comparative Validity of Personality Traits and Demographic Variables. *Psychology & Marketing*, 30(11):937–949.

Santana, C., Botta, F., Barbosa, H., Privitera, F., Menezes, R., and Di Clemente, R. (2022). Changes in the time-space dimension of human mobility during the COVID-19 pandemic. pages 1–28.

Schaninger, C. M. (1981). Social Class Versus Income Revisited: An Empirical Investigation. *Journal of Marketing Research*, 18:192–208.

Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M. R., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D. S., and others (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*, 193:35–45.

Schneider, S. and Leyer, M. (2019). Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions. *Managerial and Decision Economics*, 40(3):223–231.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Science*, 33(2):188–205.

Seetharaman, P. B. and Chintagunta, P. (1998). A model of inertia and variety-seeking with marketing variables. *International Journal of Research in Marketing*, 15(1):1–17.

Sethuraman, R., Kerin, R. A., and Cron, W. L. (2005). A field study comparing online and offline data collection methods for identifying product attribute preferences using conjoint analysis. *Journal of Business Research*, 58(5):602–610.

Sevilla, J., Lu, J., and Kahn, B. E. (2019). Variety Seeking, Satiation, and Maximizing Enjoyment Over Time. *Journal of Consumer Psychology*, 29(1):89–103.

Shankar, V., Smith, A., and Rangaswamy, A. (2003). Customer satisfaction and loyalty in online and offline environments. *International Journal of Research in Marketing*, 20(2):153–175.

Sharma, P., Sivakumaran, B., and Marshall, R. (2010a). Exploring impulse buying and variety seeking by retail shoppers: Towards a common conceptual framework. *Journal of Marketing Management*, 26(5-6):473–494.

Sharma, P., Sivakumaran, B., and Marshall, R. (2010b). Impulse buying and variety seeking: A trait-correlates perspective. *Journal of Business Research*, 63(3):276–283.

Sharp, B. and Sharp, A. (1997). Loyalty programs and their impact on repeat-purchase loyalty patterns. *International Journal of Research in Marketing*, 14(5):473–486.

Sharp, B., Wright, M., Dawes, J., Driesener, C., Meyer-Waarden, L., Stocchi, L., and Stern, P. (2012). It's a dirichlet world: Modeling individuals' loyalties reveals how brands compete, grow, and decline. *Journal of Advertising Research*, 52(2):203–213.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

Shelby, D. H. (1991). Positiism and paradigm dominance in consumer research: Toward critical pluralism and rapprochement.

Shin, H. and Cho, S. (2006). Response modeling with support vector machines. *Expert Systems with Applications*, 30(4):746–760.

Siddarth, S., Bucklin, R. E., and Morrison, D. G. (1995). Making the Cut: Modeling and Analyzing Choice Set Restriction in Scanner Panel Data. *Journal of Marketing Research*, 32(3):255.

Silvera, D. H., Lavack, A. M., and Kropp, F. (2008). Impulse buying: The role of affect, social influence, and subjective wellbeing. *Journal of Consumer Marketing*, 25(1):23–33.

Simonson, I. (1990). The Effect of Purchase Quantity and Timing on Variety-Seeking Behavior. *Journal of Marketing Research*, 27(2):150.

Siva, M., Nayak, D. P., and Narayan, K. A. (2019). Strengths and weaknesses of online surveys. *IOSR Journal of Humanities and Social Sciences (IOSR-JHSS)*, 24(5):31–38.

Skatova, A., Stewart, N., Flavahan, E., and Goulding, J. (2019). Those Whose Calorie Consumption Varies Most Eat Most. *PsyArXiv*, 44(July):1–18.

Slocum, J. W. and Mathews, H. L. (1970). Social Class and Income as Indicators of Consumer Credit Behavior. *Journal of Marketing*, 34(2):69–74.

Smith, A. (2019). *Consumer behaviour and analytics*. Routledge, London.

Smith, A., Harvey, J., Goulding, J., Smith, G., and Sparks, L. (2021). Exogenous cognition and cognitive state theory: The plexus of consumer analytics and decision-making. *Marketing Theory*, 21(1):53–74.

Smith, A. and Sparks, L. (2004). Consumer surveillance: the case of loyalty card data. *European Advances in Consumer Research*.

Smith, A., Sparks, L., Smith, A., Sparks, L., and Eve, A. A. (2004). All About Eve ? 1376(2004).

Smith, B. (2015). The Transition of Shopping Mall Development. Technical report.

Smith, G., Mansilla, R., and Goulding, J. (2020). Model Class Reliance for Random Forests. In *Advances in Neural Information Processing Systems*, pages 22305–22315.

Smith, G., Wieser, R., Goulding, J., and Barrack, D. (2014). A refined limit on the predictability of human mobility. *2014 IEEE International Conference on Pervasive Computing and Communications, PerCom 2014*, pages 88–94.

Smith, T. (2012). The personality trait predictors of brand loyalty. *Academy of Business Research*, 3:6–21.

Smith, T. A. (2020). The role of customer personality in satisfaction, attitude-to-brand and loyalty in mobile services. *Spanish Journal of Marketing - ESIC*, 24(2):155–175.

Song, C., Qu, Z., Blumm, N., and Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Sorce, P., Perotti, V., and Widrick, S. (2006). Attitude and age differences in online buying. *International Journal of Retail and Distribution Management*.

Stamps, A. E. (2002). Entropy, visual diversity, and preference. *Journal of General Psychology*, 129(3):300–320.

Stigler, S. M. (1981). *Gauss and the invention of least squares*. JSTOR.

Straathof, S. M. (2007). Shannon's entropy as an index of product variety. *Economics Letters*, 94(2):297–303.

Sun, T. and Wu, G. (2014). Trait Predictors of Online Impulsive Buying Tendency : A Hierarchical Approach. *Journal of Marketing Theory and Practice*, 6679.

Tarka, P., Kukar-Kinney, M., and Harnish, R. J. (2022). Consumers' personality and compulsive buying behavior: The role of hedonistic shopping experiences and gender in mediating-moderating relationships. *Journal of Retailing and Consumer Services*, 64.

The Swedish National Food Agency (1980). The Keyhole.

Thomas, M., Desai, K. K., and Seenivasan, S. (2011). How credit card payments increase unhealthy food purchases: Visceral regulation of vices. *Journal of Consumer Research*, 38(1):126–139.

Tian, J., Zhang, Y., and Zhang, C. (2018). Predicting consumer variety-seeking through weather data analytics. *Electronic Commerce Research and Applications*, 28:194–207.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Trivedi, M., Sridhar, K., and Kumar, A. (2016). Impact of Healthy Alternatives on Consumer Choice: A Balancing Act. *Journal of Retailing*, 92(1):65–82.

Uncles, M. (1994). Do you or your customers need a loyalty scheme? *Journal of Targeting, Measurement and Analysis for Marketing*, 2(4):335–350.

Uncles, M., Ehrenberg, A., and Hammond, K. (1995). Patterns of Buyer Behavior : Regularities , Models , and Extensions. *Marketing Science*, 14(3):G71–G78.

Uncles, M. and Hammond, K. (1995). Grocery store patronage. *The International Review of Retail, Distribution and Consumer Research*, 5(3):287–302.

Van Den Poel, D. and Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166(2):557–575.

Van Heerde, H. J., Gupta, S., and Wittink, D. R. (2003). Is 75% of the Sales Promotion Bump Due to Brand Switching? No, only 33% Is. *Journal of Marketing Research*, 40(4):481–491.

Van Kleef, E., Van Trijp, H., Paeps, F., and Fernández-Celemín, L. (2008). Consumer preferences for front-of-pack calories labelling. *Public Health Nutrition*, 11(2):203–213.

Van Trijp, H., Hoyer, W., and Inman, J. (1996). Why switch? Product category-level explanations for true variety-seeking behavior. *Journal of Marketing Research*, 33(3):281–292.

van Wezel, M. and Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, 181(1):436–452.

Vanhala, M., Lu, C., Peltonen, J., Sundqvist, S., Nummenmaa, J., and Järvelin, K. (2020). The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining–driven analysis of previous research. *Journal of Business Research*, 106(May 2018):46–59.

Venkatesh, A. (1992). Postmodernism, Consumer Culture and the Society of the Spectacle. *Advances in consumer research*, 19(1).

Vepsäläinen, H., Nevalainen, J., Kinnunen, S., Itkonen, S. T., Meinilä, J., Männistö, S., Uusitalo, L., Fogelholm, M., and Erkkola, M. (2022). Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *British Journal of Nutrition*, 128(9):1780–1788.

Verhoef, P. C., Kooge, E., and Walk, N. (2016). *Creating value with big data analytics: Making smarter marketing decisions*. Routledge.

Verplanken, B. (2006). Beyond frequency: Habit as mental construct. *British Journal of Social Psychology*, 45(3):639–656.

Verplanken, B. and Herabadi, A. (2001). Individual differences in impulse buying tendency: Feeling and no thinking. *European Journal of Personality*, 15(1 SUPPL.).

Vilcassim, N. J. and Chintagunta, P. K. (1995). Investigating retailer product category pricing from household scanner panel data. *Journal of Retailing*, 71(2):103–128.

Wang, C.-C. and Yang, H.-W. (2008). Passion for online shopping: The influence of personality and compulsive buying. *Social Behavior and Personality: an international journal*, 36(5):693–706.

Wang, Y. M., Lin, H. H., Tai, W. C., and Fan, Y. L. (2016). Understanding multi-channel research shoppers: an analysis of Internet and physical channels. *Information Systems and e-Business Management*, 14(2):389–413.

Warren, J. (2005). Representativeness of environmental samples. *Environmental Forensics*, 6(1):21–25.

Watson, R. (2009). Product variety and competition in the retail market for eyeglasses. *Journal of Industrial Economics*, 57(2):217–251.

Wen, Y. T., Yeh, P. W., Tsai, T. H., Peng, W. C., and Shuai, H. H. (2018). Customer purchase behavior prediction from payment datasets. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 628–636.

Wong, R. C.-W., Fu, A. W.-C., and Wang, K. (2005). Data mining for inventory item selection with cross-selling considerations. *Data mining and knowledge discovery*, 11:81–112.

Wood, M. (1998). Socio-economic status, delay of gratification, and impulse buying. *Journal of economic psychology*, 19(3):295–320.

Wood, R. A., McInish, T. H., and Ord, J. K. (1985). An Investigation of Transactions Data for NYSE Stocks. *The Journal of Finance*, 40(3):723–739.

Wood, S. L. (2002). Future fantasies: A social change perspective of retailing in the 21 st century. Technical report.

World Health Organization (2020). Obesity and Overweight.

Xu, C., Park, J., and Lee, J. C. (2022). The effect of shopping channel (online vs offline) on consumer decision process and firm's marketing strategy. *Internet Research*, 32(3):971–987.

Yan, J., Tian, K., Heravi, S., and Morgan, P. (2017). The vices and virtues of consumption choices: price promotion and consumer decision making. *Marketing Letters*, 28(3):461–475.

Yang, S., Wang, Y., Li, Z., Chen, C., and Yu, Z. (2022). Time-of-day effects on (un)healthy product purchases: Insights from diverse consumer behavior data. *Journal of Business Research*, 152(March 2021):447–460.

Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Zatz, L. Y., Moran, A. J., Franckle, R. L., Block, J. P., Hou, T., Blue, D., Greene, J. C., Gortmaker, S., Bleich, S. N., Polacsek, M., Thorndike, A. N., and Rimm, E. B. (2021). Comparing Online and In-Store Grocery Purchases. *Journal of Nutrition Education and Behavior*, 53(6):471–479.

Zuo, Y. (2016). Prediction of consumer purchase behaviour using Bayesian network: an operational improvement and new results based on RFID data. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 5(2):85–105.

Zuo, Y., Yada, K., and Ali, S. (2016). Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques.

Appendices

Appendix A

Description of the data sets

A.0.1 Dunnhumby data set: The Complete Journey

Variable	Description
household_key	Uniquely identifies each household
basket_id	Uniquely identifies a purchase occasion
day	Day when transaction occurred
product_id	Uniquely identifies each product
quantity	Number of the products purchased during the trip
sales_value	Amount of dollars retailer receives from the sale
store_id	Identifies unique stores
coupon_match_disc	Discount applied due to retailer's match of manufacturer coupon
coupon_disc	Discount applied due to manufacturer coupon
retail_disc	Discount applied due to retailer's loyalty card programme
trans_time	Time of day when transaction occurred
week_no	Week of the transaction. Ranges 1 – 102

Appendix B

PostgreSQL codes

B.0.1 Bundle Entropy

BE is implemented as a custom postgres aggregate function. The implementation (copy and paste into psql to install) is:

```
create extension plpython3u;

CREATE TYPE basket_tuple AS (
   basket_id       TEXT,
   item       TEXT);

CREATE OR REPLACE FUNCTION _state_bundle_entropy(
   prev basket_tuple[], basket_id TEXT, item TEXT)
   RETURNS basket_tuple[] AS
$$
   SELECT array_append(prev, (basket_id, item)::
      basket_tuple);
$$
LANGUAGE 'sql' IMMUTABLE;
```

```
CREATE OR REPLACE FUNCTION _final_bev3_norm(list_in
   basket_tuple[])
   RETURNS NUMERIC AS
$$

import pandas as pd
import numpy as np
from collections import defaultdict
from collections import Counter

data = defaultdict(list)
for record in list_in:
    data[record['basket_id']].append( record['item']
        )

dataset = [ v for k, v in data.items() ]

baskets = [ frozenset(b) for b in dataset ]

if len(baskets) == 1:
    return 0 # degenerate case

unique_baskets = Counter(baskets)

out = []
for b1 in unique_baskets.keys():
   r_b1 = 0
   for b2 in baskets:
```

```
            r_b1 += 1 -  ( max(len(b1 - b2), len(b2 - b1)
               ) / max(len(b1),len(b2)) )

        r_b1 /= len(baskets)

        out.append( ( r_b1, unique_baskets[b1] ))
    rtn = 0
    for b1p in out:

        rtn += (b1p[1]/len(baskets) ) * np.log2(b1p[0])

    return -(rtn / np.log2(len(out)))

    $$
    LANGUAGE plpython3u;

    DROP AGGREGATE IF EXISTS bev3_norm(TEXT, TEXT);
    CREATE AGGREGATE bev3_norm(TEXT, TEXT) (
      SFUNC=_state_bundle_entropy,
      STYPE=basket_tuple[],
      FINALFUNC=_final_bev3_norm,
      INITCOND='{}'
    );
```

B.0.2 Basket Revealed Entropy

BRE is implemented as a custom postgres aggregate function. The implementation (copy and paste into psql to install) is:

```
CREATE TYPE basket_tuple_with_minsup AS (
```

```
   basket_id       TEXT,
   item       TEXT,
   minsup NUMERIC);

CREATE OR REPLACE FUNCTION
   _state_bundle_entropy_with_minsup( prev
   basket_tuple_with_minsup[], basket_id TEXT, item
   TEXT, minsup NUMERIC)
   RETURNS basket_tuple_with_minsup[] AS
$$
   SELECT array_append(prev, (basket_id, item,minsup
      )::basket_tuple_with_minsup);
$$
LANGUAGE 'sql' IMMUTABLE;

CREATE OR REPLACE FUNCTION _final_bre(list_in
   basket_tuple_with_minsup[])
   RETURNS NUMERIC AS
$$

import pandas as pd
import numpy as np
from mlxtend.preprocessing import TransactionEncoder
from collections import defaultdict
from mlxtend.frequent_patterns import fpgrowth
from collections import Counter

data = defaultdict(list)
for record in list_in:
    minsup = record['minsup']
```

```
    data[record['basket_id']].append( record['item']
        )

dataset = [ v for k, v in data.items() ]

# Now we have the dataset in a python format

te = TransactionEncoder()

data_as_sets = [frozenset(d) for d in dataset]

symbols_and_cts = defaultdict(int)

# Cache the single itete = TransactionEncoder()
oht_ary = te.fit(dataset).transform(dataset, sparse=
   True)
sparse_df = pd.DataFrame.sparse.from_spmatrix(
   oht_ary, columns=te.columns_)
frequent_itemsets = fpgrowth(sparse_df, min_support=
   minsup, use_colnames=True)

data_as_sets = [frozenset(d) for d in dataset]

symbols_and_cts = defaultdict(int)

# Compute and cache the single item support
# Guidotti et. al. (incorrectly) compute lift as the
    product of all single item's (within the common
   basket's) support
lst = []
```

```
for x in data_as_sets:
   lst += list(x)
Counter(lst)
single_item_support = {k:v/len(data_as_sets) for k,
   v in Counter(lst).items()}

for basket in data_as_sets:
   intersect_len = 0
   D = []

   for i, x in enumerate(frequent_itemsets.itemsets.
      values):

       # Compute the length of intersection between
          the common pattern and the basket
       # Keep only the baskets with the longest
          intersection length
       if x <= basket:
            inter = len(basket.intersection(x))
            if inter == intersect_len:
                D.append(x)
            elif inter > intersect_len:
                intersect_len = inter
                D = [x]

   #Algorithm 2 in Guidotti et. al.
   # if there is no "common pattern" in the basket,
      the basket is forced to become a common
      pattern
   if len(D) == 0:
```

```
        symbols_and_cts[basket] += 1

    # Otherwise select the "common pattern" with the
       longest intersection lenght
    elif len(D) == 1:
        symbols_and_cts[D[0]] += 1

    # If there is more than one of these compute the
       (incorrect, but Guidotti version) of lift and
       take the lowest
    else:
        D2 = []
        lift_up = 999999
        for common_pattern in D:
            cp_lift = np.prod([ single_item_support[x
               ] for x in common_pattern ])
            if cp_lift == lift_up:
                D2.append(common_pattern)
            elif cp_lift < lift_up:
                lift_up = cp_lift
                D2 = [common_pattern]
        # If no ties for lowest lift assign the
           support of this basket to this common
           pattern
        if len(D2) == 1:
            symbols_and_cts[D2[0]] += 1
        else:
            # Otherwise approtion the support of this
                basket across the remaining patterns
            for dv in D2:
```

```
                symbols_and_cts[dv] += 1/len(D2)

total_rbs_support = np.sum([v for k, v in
   symbols_and_cts.items()])
if len(symbols_and_cts) == 1: # totally predictable
   rtn = 0
else:
   rtn = -np.sum( [ (v/total_rbs_support)*np.log2(v/
      total_rbs_support) for k, v in symbols_and_cts
      .items() if not ( np.log2(v/total_rbs_support)
       == 0 and (v/total_rbs_support) == 0) ] ) / np
      .log2(len(symbols_and_cts))

return rtn
$$
LANGUAGE plpython3u;

DROP AGGREGATE IF EXISTS bre(TEXT, TEXT, NUMERIC);
CREATE AGGREGATE bre(TEXT, TEXT, NUMERIC) (
  SFUNC=_state_bundle_entropy_with_minsup,
  STYPE=basket_tuple_with_minsup[],
  FINALFUNC=_final_bre,
  INITCOND='{}'
);
```

B.0.3 Item Level Entropy

IE is implemented as a custom postgres aggregate function. The implementation (copy and paste into psql to install) is:

```
CREATE OR REPLACE FUNCTION _final_norm_entropy(TEXT
   [])
    RETURNS NUMERIC AS
$$
DECLARE
    cnt NUMERIC;
DECLARE
    rtn NUMERIC;
DECLARE
    cntd NUMERIC;
BEGIN
    cnt := COUNT(*) FROM unnest($1) val;
    cntd := COUNT(DISTINCT val) FROM unnest($1) val;
    IF cntd < 2 THEN
       RETURN 0;
    END IF;
    SELECT INTO rtn -SUM(p)
    FROM (
      SELECT ((count(*)/cnt) * log(2,count(*)/cnt)/
         log(2,cntd)) as p, val
      FROM unnest($1) val
      GROUP BY val
    ) a;
    RETURN rtn;
END;
$$
LANGUAGE 'plpgsql' IMMUTABLE;

DROP AGGREGATE IF EXISTS norm_entropy(TEXT);
CREATE AGGREGATE norm_entropy(TEXT) (
```

```
    SFUNC=array_append,
    STYPE=TEXT[],
    FINALFUNC=_final_norm_entropy,
    INITCOND='{}');
```

B.0.4 Basket Level Entropy

BLE is implemented as a custom postgres aggregate function. The implementation (copy and paste into psql to install) is:

```
It requires the previous code to be run to define
    basket_tuple and _state_bundle_entropy:

CREATE OR REPLACE FUNCTION _final_joint_entropy(
    list_in basket_tuple[])
    RETURNS NUMERIC AS
$$
DECLARE
     rtn NUMERIC;

BEGIN
 SELECT INTO rtn norm_entropy(basket::TEXT)
 FROM (
      SELECT basket_id, array_agg(DISTINCT item ORDER
          BY item) as basket
      FROM unnest($1)
      GROUP BY 1
) x ;
RETURN rtn;
END;
```

```
$$
LANGUAGE 'plpgsql' IMMUTABLE;

DROP AGGREGATE IF EXISTS joint_entropy(TEXT, TEXT);
CREATE AGGREGATE joint_entropy(TEXT, TEXT) (
  SFUNC=_state_bundle_entropy,
  STYPE=basket_tuple[],
  FINALFUNC=_final_joint_entropy,
  INITCOND='{}'
);
```

Appendix C

Bundle entropy properties

C.0.1 Proofs that bundle entropy meets properties P0-P2.

Given $BE(\mathcal{B})$ is defined as:

$$\frac{1}{\log_2 |B|} \times \sum_{b_k \in B} p(b_k) \left[-\log_2 \left(\sum_{b_q \in B} p(b_q) \frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)} \right) \right]$$

P0.a: When $b_0 = b_1 = \ldots = b_n$ then $\frac{|b_k \cap b_q|}{max(|b_k|,|b_1|)} = 1$, resulting in $BE(\mathcal{B}) = 0$.

P0.b: When $b_0 \cap b_1 \cap \ldots \cap b_n = \emptyset$ then:

$$BE(\mathcal{B}) = \sum_{b_k \in B} p(b_k)[-log_2(p(b_k))]$$

As $|b_k \cap b_q| = 0$ except when $b = q$. As each b_k is unique with a probability of $\frac{1}{|B|}$, then the term excluding the normalisation term sums to $log_2|B|$, resulting in a value of 1.

P1: When $\mathcal{B}' = [b_0 \cup \Gamma, b_1 \cup \Gamma, \ldots,] = [b'_0, b'_1, \ldots,]$

where Γ is any non-zero arbitrary combination of items and $\exists b_k : \Gamma \not\subset b_k$.

Given: $p(b_k) = p(b'_k)$

And:

$$\begin{aligned} \frac{|b'_k \cap b'_q|}{max(|b'_k|, |b'_q|)} &= \frac{|(b_k \cup \Gamma) \cap (b_q \cup \Gamma)|}{max(|b_k \cup \Gamma|, |b_q \cup \Gamma|)} \\ &\geq \frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)} \end{aligned}$$

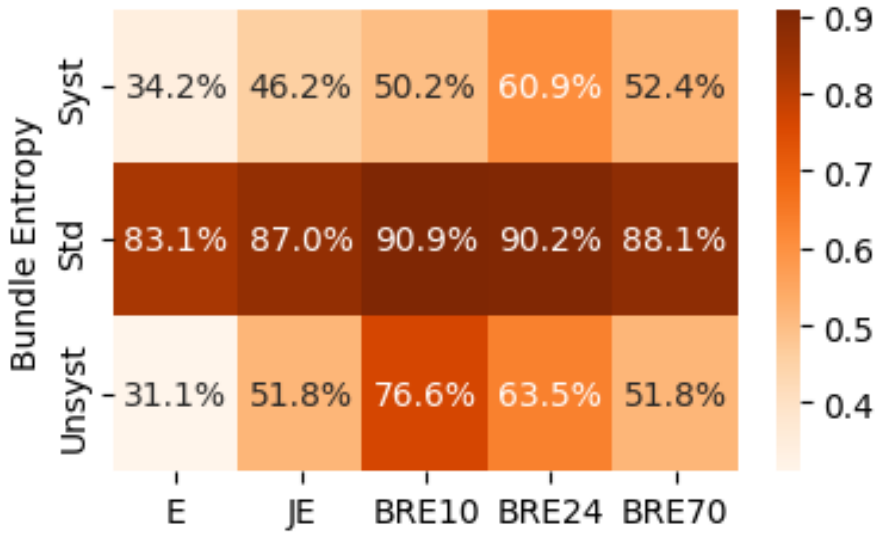
With the inequality strict when $\Gamma \not\subset b_k$ and $b_k \neq b_q$, which by definition must be true at least once or all baskets are the same, resulting in P0.a. Via the summation and negative log and since $|B'| \leq |B|$ as baskets are represented as sets and adding identical sets to all sets in an existing collection of sets can only reduce the number of distinct sets in the collection, then $BE(\mathcal{B}') < BE(\mathcal{B})$.

P2: Holds via the $P1$ proof, mapping $b_0 \cup \Gamma$ to b_0 & Γ' to Γ.

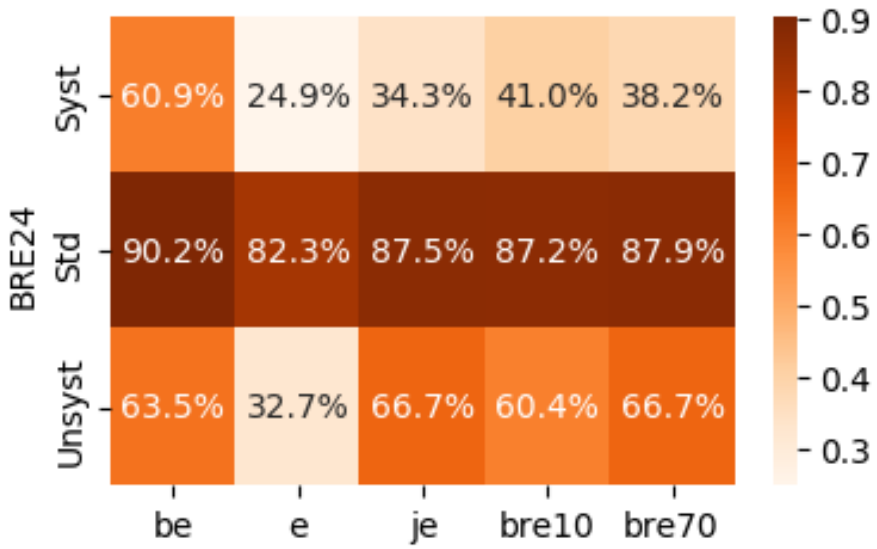
Appendix D

Dunnhumby: Customer classifications

D.0.1 Not-normalised: Percentage of customers share with respect to Bundle Entropy purchase patterns classifications.



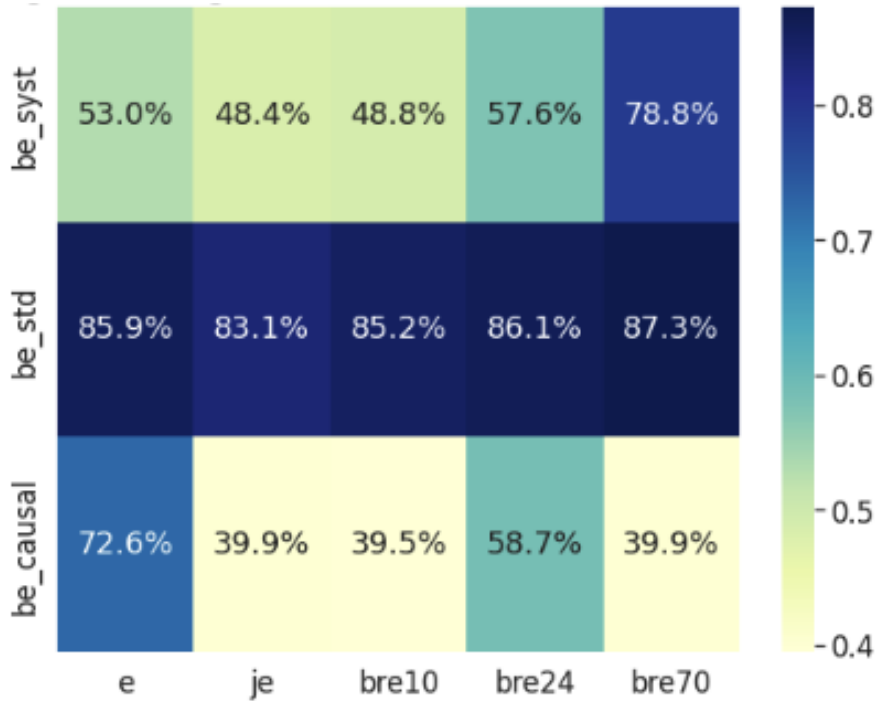
D.0.2 Not-normalised: Percentage of customers share with respect to Basket Revealed Entropy purchase patterns classifications.



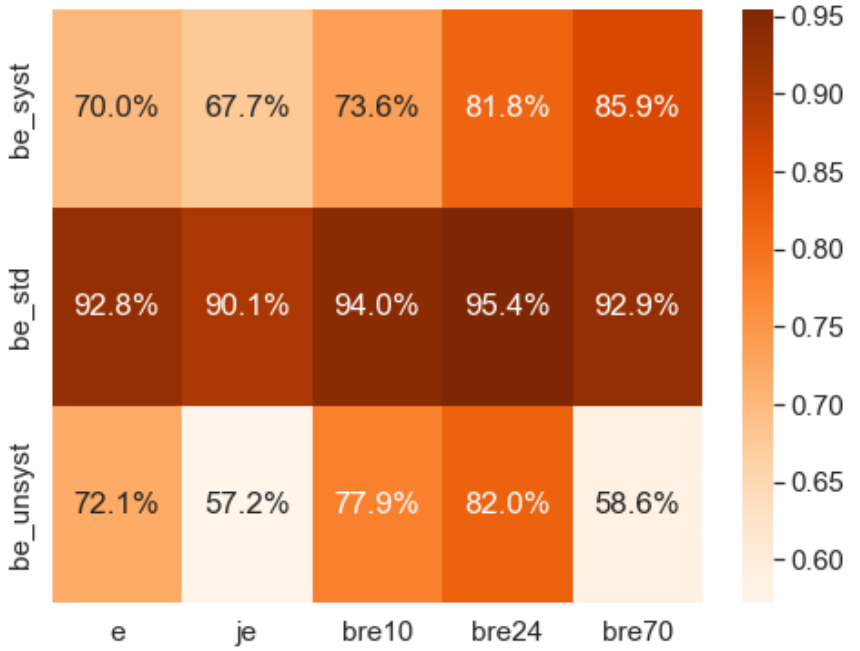
Appendix E

UK grocery retailer: Customer classifications

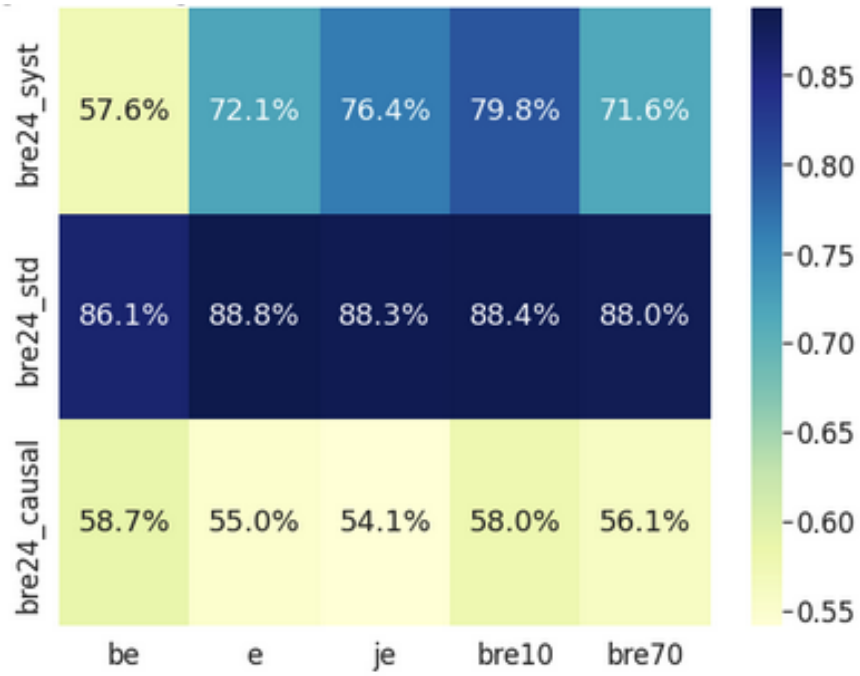
E.0.1 Normalised: Percentage of customers share with respect to Bundle Entropy purchase.



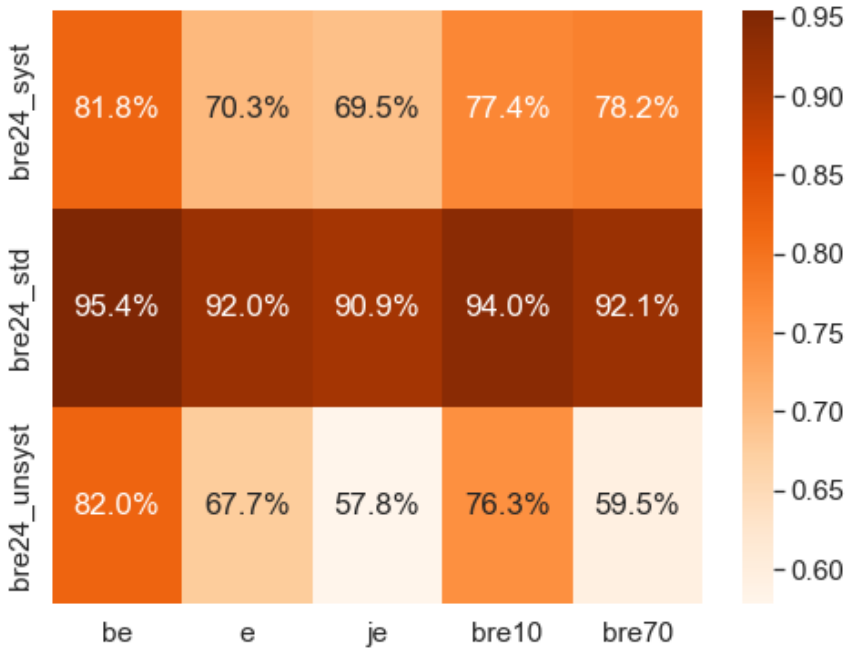
E.0.2 Not-normalised: Percentage of customers share with respect to Bundle Entropy purchase.



E.0.3 Normalised: Percentage of customers share with respect to Basket Revealed Entropy purchase.



E.0.4 Not-normalised: Percentage of customers share with respect to Basket Revealed Entropy purchase.



Appendix F

Health Score Calculation

F.0.1 Resources to compute health scores.

Total ‘A’ pts = [energy pts] + [sat. fat pts] + [sugars pts] + [sodium pts] (F.1)

Total ‘C’ pts = [fruit, vegetable & nut pts] + [fibre pts] + [protein pts] (F.2)

Overall score = [total ‘A’ pts] – [Total ‘C’ pts] (F.3)

Overall score = [total ‘A’ pts] – [fibre pts + fruit, vegetable & nut pts] (F.4)

$$\text{Scale overall score} = (-2 * \text{Overall score}) + 70 \tag{F.5}$$

Table F.1: Total 'A' table points

Points	Energy (kj)	Sat Fat (g)	Total Sugar (g)	Sodium (mg)
0	≤ 335	≤ 1	≤ 4.5	≤ 90
1	>335	> 1	> 4.5	> 90
2	>670	> 2	> 9	> 180
3	>1005	> 3	> 13.5	> 270
4	>1340	> 4	> 18	> 360
5	>1675	> 5	> 22.5	> 450
6	>2010	> 6	> 27	> 540
7	>2345	> 7	> 31	> 630
8	>2680	> 8	> 36	> 720
9	>3015	> 9	> 40	> 810
10	>3350	> 10	> 45	> 900

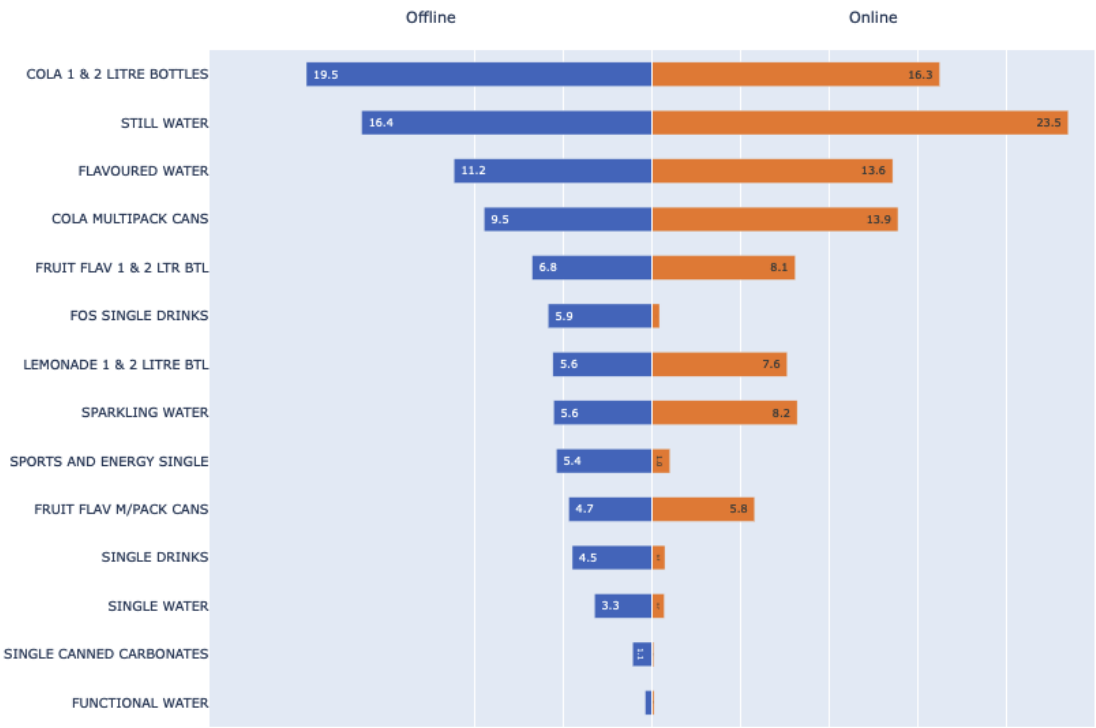
Table F.2: Total 'C' table points

Points	Fruit & Nuts (%)	NSP Fiber (g)	AOAC Fiber (g)	Protein (g)
0	≤ 40	≤ 0.7	≤ 0.9	≤ 1.6
1	>40	> 0.7	> 0.9	> 1.6
2	>60	> 1.4	> 1.9	> 3.2
3	-	> 2.1	> 2.8	> 4.8
4	-	> 2.8	> 3.7	> 6.4
5	>80	> 3.5	> 4.7	> 8.0

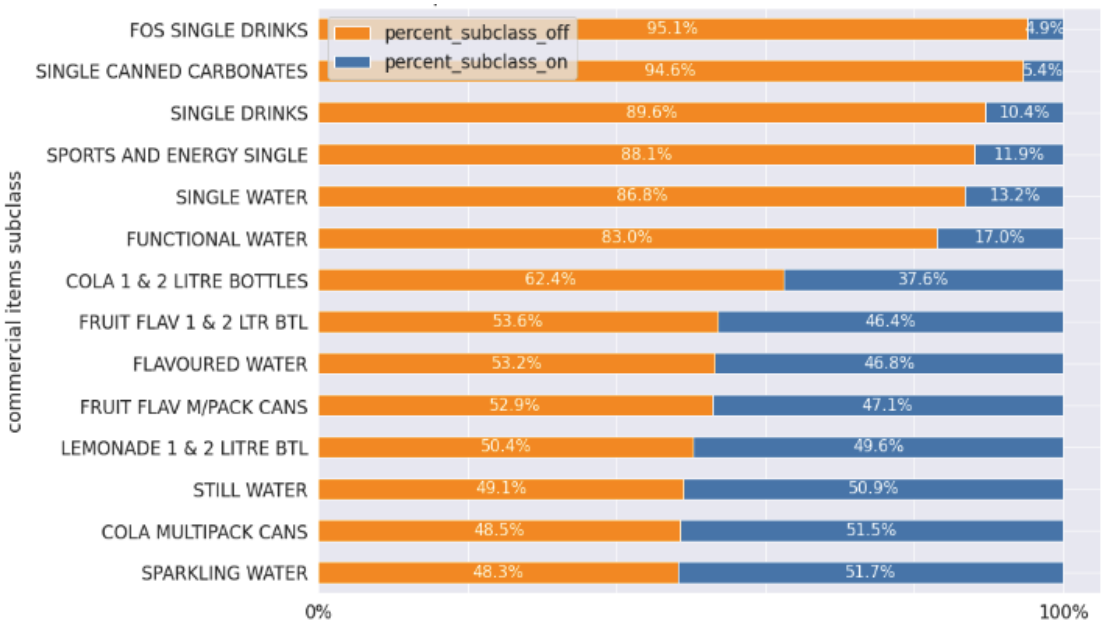
Appendix G

Soft drinks stats

G.0.1 Percentage of each type of soft drink out of the total items sold in each channel



G.0.2 Percentage of soft drinks sold online and offline



Appendix H

Survey About Demographic and Psychographic Questions

H.0.1 List of questions to customers from a leading UK grocery and pharmacy chain.

Please note: the store, loyalty card, and customer panel name have all been anonymised in this version of the questionnaire.

Participant Information Sheet

OVERVIEW

Thank you for your interest in this study. This research is being conducted by the ***<<store name>>*** customer panel in collaboration with the University of Nottingham. The study is designed to explore how and why people exhibit different consumption patterns across their daily lives. You will be invited to answer a number of questions about yourself (such as demographic information - age range, gender, income bracket and some other questions) and about various everyday behaviours that you think characterise you and your personality. *Please note, you may have answered some of these types of questions before – we are asking them again just to ensure we have the most up-to-date information!*

As ever, the information you provide is for research purposes only. Your responses in the survey will be combined with those from other ***<<store name>>*** *customer panel* members, and only these anonymised responses will be shared with the University of Nottingham. Therefore, **at no time** will any of your personally identifiable information (including name and contact details) be shared with the University of Nottingham or with any other third party.*

Anonymised responses in the survey will be linked to shopping patterns, which we will observe via the loyalty card records. This association will allow us to explore how personality and everyday behaviours affect real-world consumption patterns.

THE SURVEY

This survey will take up to 20 minutes to complete. Your participation in the survey is completely voluntary. You may withdraw from participation at any time without consequences by simply leaving the survey.

After analysis, the results and conclusions of the study may be shared by University in the form of presentations and open publications with academic and industry partners – but these will only contain averaged results and **never** will any raw data or personal details be made available.

Please proceed to the next page to fill in the consent form.

**We always follow the MRS guidelines when storing your data, however it is important to ensure you are entirely comfortable with taking part in this research and that you know that your responses will not be used in any other way*

Consent form

Please note that this survey will skip to the end if you disagree with any of the following statements.
SINGLE CODE

1. I confirm that I have read and understood all information provided about this study. (Y/N)
2. I agree that data gathered in this study will be stored anonymously and securely, and will be used for market and academic research purposes only. (Y/N)
3. I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason. (Y/N)
4. I understand that all personal information will remain confidentially within the ***<<store name>>*** customer panel archives, and that no personally identifiable information will be shared with

University of Nottingham or any other third party, and that no data will be made available that can allow me to be personally identified in the results of this research. (Y/N)

5. I am at least 18 years of age. (Y/N)
6. I agree to take part in this study. (Y/N)

PART 1

Please answer the question below as accurately as possible.

SECTION 1: DEMOGRAPHIC QUESTIONS

Firstly, we'd like to ask you some questions about you and your household…

S1. Qualifications

What is your highest level of education?

1. Completed some secondary school
2. GCSE(s) or equivalent
3. BTEC or equivalent
4. AS-level
5. A-level or equivalent
6. Bachelor's degree
7. Other postgraduate qualification (e.g. postgraduate diploma)
8. Master's degree
9. Ph.D., law or medical degree
10. Other advanced degree beyond a Master's degree
11. Other – please specify

S2. Occupation

Which of the following best describes the occupation of the main income earner in your household?

If you are a student living away from home, please indicate the occupation of the main income earner at your family home.

If the main income earner in your household is now retired, and is not entirely reliant on the State Pension, please tell us the occupation he/she used to have.

1. Semi or unskilled manual work
   e.g. Manual workers, all apprentices to be skilled trades, Caretaker, Park Keeper, non-HGV driver, Shop Assistant
2. Skilled manual worker
   e.g. Skilled Bricklayer, Carpenter, Plumber, Painter, Bus/ Ambulance Driver, HGV Driver, AA Patrolman, Pub/Bar Worker etc.
3. Supervisory or clerical/ junior managerial/ professional/ administrative
   e.g. Office worker, Student Doctor, Foreman with 25+ employees, Salesperson, etc.
4. Intermediate managerial/ professional/ administrative
   e.g. Newly qualified (under 3 years) Doctor, Solicitor, Board Director small organisation, Middle Manager in large organisation, Principle Officer in Civil Service/Local Government
5. Higher managerial/ professional/ administrative
   e.g. Established Doctor, Solicitor, Board Director in a large organisation (200+ employees, top level Civil Servant/Public Service Employee
6. Student
7. Casual worker - not in permanent employment
8. Housewife/ Homemaker
9. Retired and living on State Pension

10. Unemployed or not working due to long-term sickness
11. Full-time carer of other household member
12. Other – please specify

S3. Income

What is your total annual household income before tax?

1. Less than £25,000
2. £25,000 to £34,999
3. £35,000 to £49,999
4. £50,000 to £74,999
5. £75,000 to £99,999
6. £100,000 to £149,999
7. £150,000 or more
8. Prefer not to say

S4. Marital status

What is your marital status?

1. Single, never married
2. Married, civil or domestic partnership/living with a partner
3. Widowed
4. Divorced
5. Separated
6. Other – please specify

S5. People

Please could you confirm which household situation best applies to you?

1. Living on my own (no children or children have left home)
2. Living on my own with children under 18
3. Living with partner/spouse (no children or children have left home)
4. Living with partner/spouse with children under 18
5. Living with other adult family members (i.e. aged 18 or older) e.g. adult children, parents and/or elderly relatives
6. Living with other adults that are non-family members e.g. friends/flatmates

[ASK THOSE WHO RESPOND 5 AT S5]

S6. Children

Which of the following adult family members do you live with?

Please select all that apply

1. With partner/spouse
2. With adult children (all aged 18 or older)
3. With parent(s)
4. With other adult family member(s)
5. None

[ASK THOSE WHO RESPOND 6 AT S5]

S7. Children

Which of the following adult non-family members (e.g. housemates) do you live with?

Please select all that apply

1. Friend(s)
2. Houseshare
3. Flatmate(s)
4. Landlord
5. Other – please specify

6. None

[ASK THOSE WHO RESPOND 2 OR 4 AT S5]

S8. Children

How many children aged 18 or under are there living in your household?

1. None
2. One
3. Two
4. Three
5. Four
6. Five or more

SECTION 2: PERCEIVED SHOPPING HABITS AND DIET

Now, we'd like to ask a few questions about the way that you shop…

S9. Shopping

How often would you say you visit any <<store name>> shop at all?

1. Every day
2. Twice a week
3. More than twice a week
4. Once a week
5. Once a fortnight
6. Once a month
7. Once every 2-3 months
8. Once every 4-6 months
9. Every 6 months or less
10. Don't know

S10. Shopping

Which of the following products do you shop for in <<store name>>?

Please select all that apply

1. Pharmacy items (only available over the counter)
2. General medicines (e.g. pain relief)
3. Vitamins or supplements
4. Everyday toiletries (e.g. Shower gel, Toothpaste, Deodorants, Femcare)
5. Facial Skincare
6. Cosmetics
7. Men's toiletries and beauty products (e.g. shaving products)
8. Photo development/ photography items
9. Holiday products (e.g. suntan lotion)
10. Lunchtime food (e.g. meal deal) and snacks
11. Baby items (including clothing, nappies and food)
12. Footcare
13. Haircare
14. Electrical Beauty
15. Toys
16. Gifts
17. Fragrance
18. Prescriptions
19. Other - please specify
20. None of these

S11. Shopping

How often do you tend to use your loyalty card when buying something in <<store name>>?

1. Every time
2. Most of the time
3. Sometimes
4. Hardly ever
5. Never

S12. Food

Do you have any specific dietary requirements?

Please select all that apply

1. Vegetarian
2. Vegan
3. Pescaterian
4. Diabetic
5. Gluten intolerant
6. Kosher
7. Lactose intolerant
8. Allergic to nuts
9. Allergic to fish
10. None
11. Other – please specify

SECTION 3A: TIME PREFERENCES

Now we'd like to ask you a question about your time preferences…

S13. ***If offered the following financial alternatives which would you rather have?***

1. 1. £40 now
OR
2. £70 in 3 months

S14. ***If offered the following financial alternatives which would you rather have***

1. £25 now
   OR
2. £70 in 3 months

S15. ***If offered the following financial alternatives which would you rather have***

1. £55 now
   OR
2. £70 in 3 months

SECTION 3B – WELLBEING

Now we're going to ask you some questions about your general wellbeing…

Please select one response for each statement below

Overall, how satisfied are you with your life nowadays?	**0 – Not satisfied at all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10 – Completely satisfied**

Overall, how happy did you feel yesterday?	**0 – Not happy at all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10 – Completely happy**

PART 2

We are going to show you some statements that people have made about shopping/spending and general behaviours. Each item is a statement that you may either agree or disagree with, and to a certain extent characterise your everyday behaviours.
For each item, simply indicate how much you agree or disagree with what the item says. Please be as accurate and honest as you can be!

SECTION 5: SHOPPING MOTIVATIONS

Now, we'd like to ask you about your shopping preferences…

How much do you agree or disagree with each item?
Please click on the bar below and slide to the number you want to select, where 1 is 'Disagree strongly' and 7 is 'Agree strongly'.

	Disagree strongly	*Disagree moderately*	*Disagree a little*	*Neither agree nor disagree*	*Agree a little*	*Agree moderately*	*Agree strongly*
1. I would rather stick to a brand I usually buy than try something I am not very sure of.	1	2	3	4	5	6	7
2. I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchases.	1	2	3	4	5	6	7
3. If I like a brand I rarely switch from it just to try something new.	1	2	3	4	5	6	7

4. I would not mind paying more in order to get a high quality product.	1	2	3	4	5	6	7
5. I only buy products I trust.	1	2	3	4	5	6	7
6. Product Quality is extremely important to me.	1	2	3	4	5	6	7
7. I am pleased about the way I look.	1	2	3	4	5	6	7
8. When I feel good about my looks, I am happier and have a better outlook on life.	1	2	3	4	5	6	7
9. Whenever I see a mirror, I "check myself out" to see how I look.	1	2	3	4	5	6	7
10. Environmental considerations affect the products that I purchase.	1	2	3	4	5	6	7
11. I am concerned about climate change (also known as global warming).	1	2	3	4	5	6	7

SECTION 4: PERSONALITY

We'd like to learn a bit more about how you are as a person…

How much do you agree or disagree with the following statements…

I see myself as…..	*Disagree strongly*	*Disagree moderately*	*Disagree a little*	*Neither agree nor disagree*	*Agree a little*	*Agree moderately*	*Agree strongly*
1. Extraverted, enthusiastic	1	2	3	4	5	6	7
2. Critical, quarrelsome	1	2	3	4	5	6	7
3. Dependable, self-disciplined	1	2	3	4	5	6	7
4. Anxious, easily upset	1	2	3	4	5	6	7
5. Open to new experiences, complex	1	2	3	4	5	6	7
6. Reserved, quiet	1	2	3	4	5	6	7
7. Sympathetic, warm	1	2	3	4	5	6	7
8. Disorganised, careless	1	2	3	4	5	6	7
9. Calm, emotionally stable	1	2	3	4	5	6	7
10. Conventional, uncreative	1	2	3	4	5	6	7

SECTION 6: SELF-CONTROL

We're now going to ask you a few questions about what motivates you in general…

Please click on the bar below and slide to the number you want to select, where 1 is 'Very false for me' and 6 is 'Very true for me'.

	Very false for me	*Somewhat false for me*	*Neither false nor true of me*	*Somewhat true of me*	*Very true for me*

1. I am good at resisting temptation.	*1*	*2*	*3*	*4*	*5*
2. I have a hard time breaking bad habits.	*1*	*2*	*3*	*4*	*5*
3. I am lazy.	*1*	*2*	*3*	*4*	*5*
4. I say inappropriate things.	*1*	*2*	*3*	*4*	*5*
5. I do certain things that are bad for me, if they are fun.	*1*	*2*	*3*	*4*	*5*
6. I refuse things that are bad for me.	*1*	*2*	*3*	*4*	*5*
7. I wish I had more self-discipline.	*1*	*2*	*3*	*4*	*5*
8. People would say that I have iron self- discipline.	*1*	*2*	*3*	*4*	*5*
9. Pleasure and fun sometimes keep me from getting work done.	*1*	*2*	*3*	*4*	*5*
10. I have trouble concentrating.	*1*	*2*	*3*	*4*	*5*
11. I am able to work effectively toward long-term goals.	*1*	*2*	*3*	*4*	*5*
12. I often act without thinking through all the alternatives	*1*	*2*	*3*	*4*	*5*
13. Sometimes I can't stop myself from doing something, even if I know it's wrong	*1*	*2*	*3*	*4*	*5*

SECTION 7: SHOPPING IMPULSIVITY

We're now going to ask you some questions about how you like to make purchases…

How much do you agree or disagree with each item?

Please select one response for each statement below.

	Very inaccurate	*Moderately inaccurate*	*Neither accurate or inaccurate*	*Moderately accurate*	*Very accurate*
1. I often buy things spontaneously	*1*	*2*	*3*	*4*	*5*
2. "Just do it" describes the way I buy things	*1*	*2*	*3*	*4*	*5*
3. I often buy things without thinking	*1*	*2*	*3*	*4*	*5*
4. "I see it, I buy it" describes me	*1*	*2*	*3*	*4*	*5*
5. "Buy now, think about it later" describes me	*1*	*2*	*3*	*4*	*5*
6. Sometimes I feel like buying things on the spur-of-the-moment	*1*	*2*	*3*	*4*	*5*
7. I buy things according to how I feel at the moment	*1*	*2*	*3*	*4*	*5*
8. I carefully plan most of my purchases	*1*	*2*	*3*	*4*	*5*
9. Sometimes I am a bit reckless about what I buy	*1*	*2*	*3*	*4*	*5*

SECTION 8: FRUGALITY

Now, we're going to ask you just a few questions around your general attitudes towards spending…

How much do you agree or disagree with each item?
Please select one response for each statement below.

	Very inaccurate	*Moderately inaccurate*	*Slightly inaccurate*	*Slightly accurate*	*Moderately accurate*	*Very accurate*
1. I am willing to wait on a purchase I want so that I can save money	*1*	*2*	*3*	*4*	*5*	*6*
2. Making better use of my resources makes me feel good	*1*	*2*	*3*	*4*	*5*	*6*
3. Many things that are normally thrown away are still quite useful	*1*	*2*	*3*	*4*	*5*	*6*
4. I believe in being careful in how I spend my money	*1*	*2*	*3*	*4*	*5*	*6*
5. If you can re-use an item you already have, there's no sense in buying something new	*1*	*2*	*3*	*4*	*5*	*6*
6. I discipline myself to get the most from my money	*1*	*2*	*3*	*4*	*5*	*6*
7. If you take good care of possessions, you will definitely save money in the long run	*1*	*2*	*3*	*4*	*5*	*6*
8. There are things I resist buying today so I can save for tomorrow	*1*	*2*	*3*	*4*	*5*	*6*

SECTION 9. GENERAL IMPULSIVITY

Now, we're going to ask you some more general questions about you…

Please click on the bar below and slide to the number you want to select, where 1 is 'Very False for me' and 6 is 'Very true for me'.

	Very false for me	*Somewhat false for me*	*Somewhat true of me*	*Very true for me*
1. I go out of my way to get things I want.	*1*	*2*	*3*	*4*
2. When I'm doing well at something I love to keep at it.	*1*	*2*	*3*	*4*
3. I'm always willing to try new things if I think it will be fun.	*1*	*2*	*3*	*4*
4. When I get something I want, I feel excited and energized.	*1*	*2*	*3*	*4*
5. When I want something I usually go all-out to get it.	*1*	*2*	*3*	*4*
6. I will often do things for no other reason than that they might be fun.	*1*	*2*	*3*	*4*
7. If I see a chance to get something I want I move on it right away.	*1*	*2*	*3*	*4*
8. When I see an opportunity for something I like I get excited right away.	*1*	*2*	*3*	*4*
9. I often act on the spur of the moment.	*1*	*2*	*3*	*4*
10. When good things happen to me, it affects me strongly.	*1*	*2*	*3*	*4*
11. I crave excitement and new sensations.	*1*	*2*	*3*	*4*
12. It would excite me to win a contest.	*1*	*2*	*3*	*4*
13. When I go after something I use a "no holds barred" approach.	*1*	*2*	*3*	*4*

SECTION 10. COMMENTS

F1. Finally, do you have any other thoughts or feelings which you would like to share with us on this topic or about this survey?
[open answer]

F2. Thanks very much for taking part in this survey.

As your views are so important to us, we'd be grateful if you could answer a couple of quick questions telling us what you thought about it.

Your answers will be taken on board and considered when devising future panel surveys.

We ask these survey health questions to find out how your experience of this particular survey compares to other surveys we run - for example to check it's not too long or repetitive. We look at the results of each survey to try to learn from the feedback and improve our surveys moving forward. We greatly appreciate your feedback - but feel free to skip this section this time if you would prefer.

1. Yes, happy to do so

2. No thanks, maybe next time

F3. Taking everything into account, including how much you enjoyed it and how relevant it was to you personally, how would you rate this survey?

Very poor **Excellent**

0	1	2	3	4	5	6	7	8	9	10

F4. And to what extent, if at all, do you feel this survey was long and\or repetitive?

Not at all long and repetitive **Extremely long and repetitive**

0	1	2	3	4	5	6	7	8	9	10

F5. If you have any ideas, comments or suggestions you would like to share about the subject or design of today's survey, please write them below.

[open answer]

Appendix I

SHAP values

I.0.1 SHAP plot for random forest using demographic and psychographic predictors.

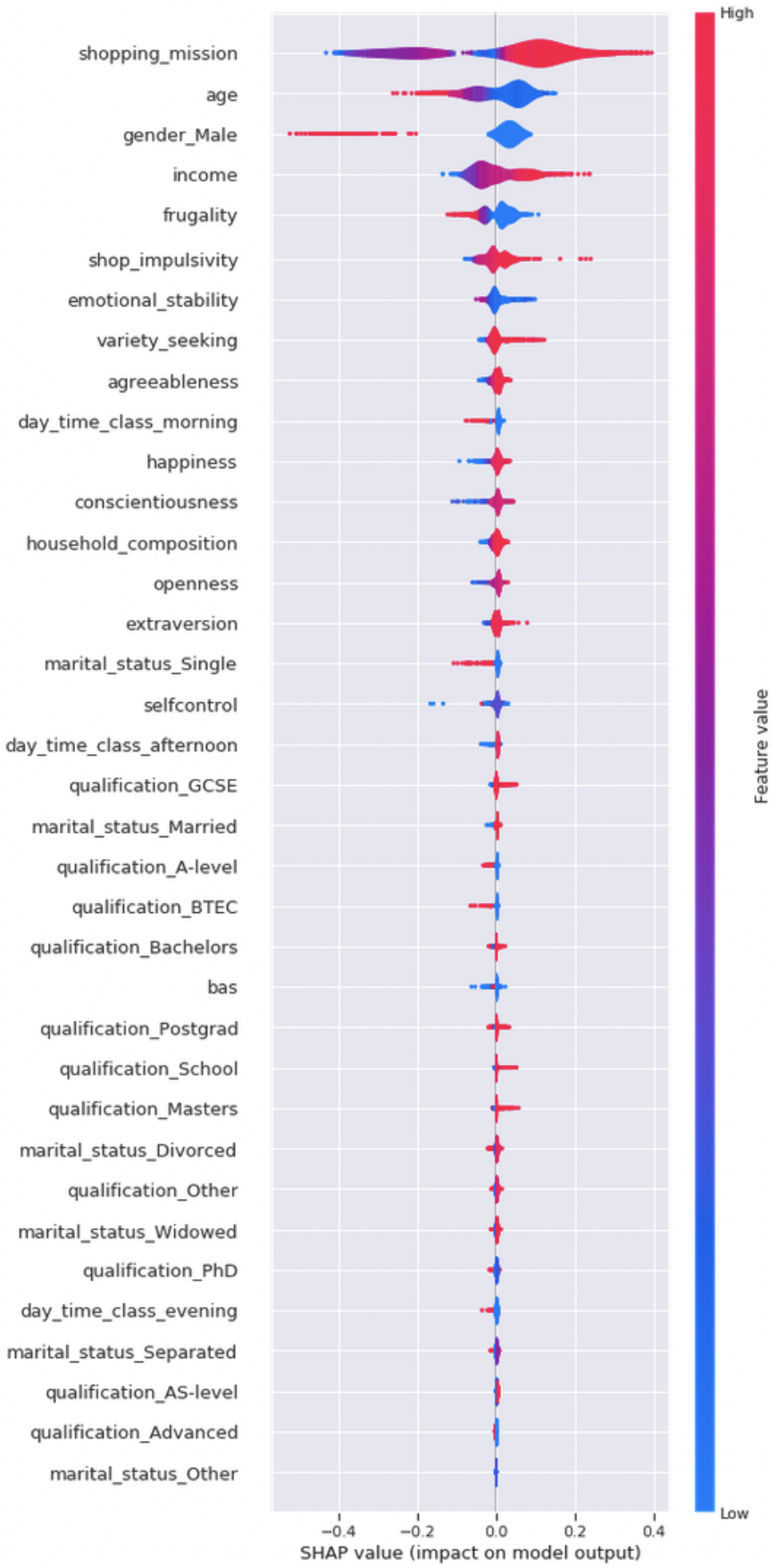


Figure I.1: SHAP summary plot for RF using demographic and psychographic variables to predict SPB.