Enhancing the Prediction of Missing Targeted Items from the Transactions of Frequent, Known Users

by

Raymond Shunmugam Moodley

A thesis submitted in partial fulfillment for the degree of
Doctor of Philosophy

in the
Faculty of Computing, Engineering and Media, De Montfort University

November 2019

Leicester, United Kingdom

Declaration of Authorship

The content of this submission was undertaken in the School of Computer Science and Informatics, De Montfort University, and supervised by Prof. Francisco Chiclana, Dr Fabio Caraffini and Dr Jenny Carter (University of Huddersfield) during the period of registration. I hereby declare that the materials of this submission have not previously been published for a degree, or diploma at any other university or institution. All materials submitted for assessment are from my own research, except the reference work in any format by other authors, which are properly acknowledged in the content. Part of the research work presented in this submission has been published, or has been submitted for publication in the following papers:

- Moodley, R., Chiclana, F., Caraffini, F. and Carter, J., 2019. Application of uninorms to market basket analysis. *International Journal of Intelligent Systems*, 34(1), pp.39-49.

- Moodley, R., Chiclana, F., Caraffini, F. and Carter, J., 2019. A product-centric data mining algorithm for targeted promotions. *Journal of Retail and Consumer Services*, October 2019.

- Moodley, R., Chiclana, F., Caraffini, F. and Carter, J., 2019. A data mining approach for enhancing pupil attendance at schools. *Knowledge-Based Systems*, under review, November 2019.

Abstract

The ability for individual grocery retailers to have a single view of its customers across all of their grocery purchases remains elusive, and is considered the "holy grail" of grocery retailing. This has become increasingly important in recent years, especially in the UK, where competition has intensified, shopping habits and demographics have changed, and price sensitivity has increased. Whilst numerous studies have been conducted on understanding independent items that are frequently bought together, there has been little research conducted on using this knowledge of frequent itemsets to support decision making for targeted promotions. Indeed, having an effective targeted promotions approach may be seen as an outcome of the "holy grail", as it will allow retailers to promote the right item, to the right customer, using the right incentives to drive up revenue, profitability, and customer share, whilst minimising costs.

Given this, the key and original contribution of this study is the development of the market target (mt) model, the clustering approach, and the computer-based algorithm to enhance targeted promotions. Tests conducted on large scale consumer panel data, with over 32000 customers and 51 million individual scanned items per year, show that the mt model and the clustering approach successfully identifies both the best items, and customers to target. Further, the algorithm segregates customers into differing categories of loyalty, in this case it is four, to enable retailers to offer customised incentives schemes to each group, thereby enhancing customer engagement,

whilst preventing unnecessary revenue erosion. The proposed model is compared with both a recently published approach, and the cross-sectional shopping patterns of the customers on the consumer scanner panel. Tests show that the proposed approach outperforms the other approach in that it significantly reduces the probability of having "false negatives" and "false positives" in the target customer set. Tests also show that the customer segmentation approach is effective, in that customers who are classed as highly loyal to a grocery retailer, are indeed loyal, whilst those that are classified as "switchers" do indeed have low levels of loyalty to the selected grocery retailer.

Applying the mt model to other fields has not only been novel but yielded success. School attendance is improved with the aid of the mt model being applied to attendance data. In this regard, an action research study, involving the proposed mt model and approach, conducted at a local UK primary school, has resulted in the school now meeting the required attendance targets set by the government, and it has halved its persistent absenteeism for the first time in four years. In medicine, the mt model is seen as a useful tool that could rapidly uncover associations that may lead to new research hypotheses, whilst in crime prevention, the mt value may be used as an effective, tangible, efficiency metric that will lead to enhanced crime prevention outcomes, and support stronger community engagement.

Future work includes the development of a software program for improving school attendance that will be offered to all schools, while further progress will be made on demonstrating the effectiveness of the mt value as a tangible crime prevention metric.

Acknowledgements

I would like to thank my supervisors, Prof. Francisco Chiclana, Dr Jenny Carter, and Dr Fabio Caraffini for their help, guidance, and support throughout this project.

I would also like to extend my appreciation to Kantar for allowing me to use their data, and to the leadership of Willen Primary School for working with me to help demonstrate the impact of this project on their school's attendance.

Finally, I want to thank my two wonderful daughters, Emily and Hayley, for their love, support and "humour", and last, but by no means least, my loving wife, Deshnee, for not only her great love, and help, but for always being by my side throughout this project, and indeed, over the last 24 years.

Table of Contents

List of Tables

List of Figures

Nomenclature

Variables

e	neutral element used in uninorms; $0 \leq e \leq 1$
m	fuzzifier, used in FCM; $1 < m < \infty$
$minsup$	minimum support; $0 < minsup \leq 1$
$minconf$	minimum confidence; $0 < minconf \leq 1$

Acronyms

ARM	Association Rule Mining
BHF	British Heart Foundation
BMI	Body Mass Index
CBF	Content Based Filtering
CHD	Coronary Heart Disease
CF	Collaborative Filtering
DES	Discrete Event Simulation
EDM	Educational Data Mining
ETTP	Early Truancy Prevention Program
FCM	Fuzzy-C-Means clustering
HBP	High Blood Pressure
KDD	Knowledge Discovery in Databases
KM	K-Means clustering
MBA	Market Basket Analysis
MFI	Maximal Frequent Itemset
mt	Market Target
NDI	Non-Derivable Itemset
RFM	Recency, Frequency, Monetary
RS	Recommender System
SAD	Separation Anxiety Disorder
SR	School Refusal
S&S	Stop and Search
WPS	Willen Primary School

Chapter 1

Introduction

The need to have an intimate understanding of the customer, with a view of predicting their wants, has always been a key ambition of retailers across the globe. This has become increasingly important in recent years, as a result of increased competition, and advances in technology that now makes attracting and retaining customers more challenging [19][120][154]. Today, it is no longer sufficient to provide customers with just the goods or services that they seek, but it has become essential to accompany these goods and services with an outstanding level of customer service; and not just any generic customer service at that; but rather a tailored one. This is commonly referred to as "providing a unique customer experience" [19][154]. Consequently today, the task of attracting and retaining customers entails both providing high quality goods and services, as well as a unique customer experience for every customer.

The main thrust of this research has been framed against this backdrop with the aim of developing novel data mining models, and an algorithm that helps grocery retailers identify the best items and customers to target, and to support grocery retailers with the prediction of customer behaviour, more specifically, the way in which customers split their purchases across multiple grocery retailers. By understanding both these

aspects, grocery retailers will be able to enhance their endeavours in offering customers a unique customer experience through tailored marketing promotions, thereby: (1) enhancing the effectiveness of their marketing spend, (2) enticing customers to switch grocery retailers, where they split their purchases, and (3) sustaining loyalty amongst its current customers. Figure 1.1 is a high level, flow diagram that outlines the proposed novel algorithm, the interplay between the various unique contributions of this research; and how they support grocery retailers' endeavours in offering a unique customer experience to all of their customers, whilst maximising sales, and minimising costs. These key components are discussed throughout this study. The UK grocery retail sector was used as a case study, and the data used as part of this study was obtained from the third party UK grocery retail sector analyst group, Kantar [90].

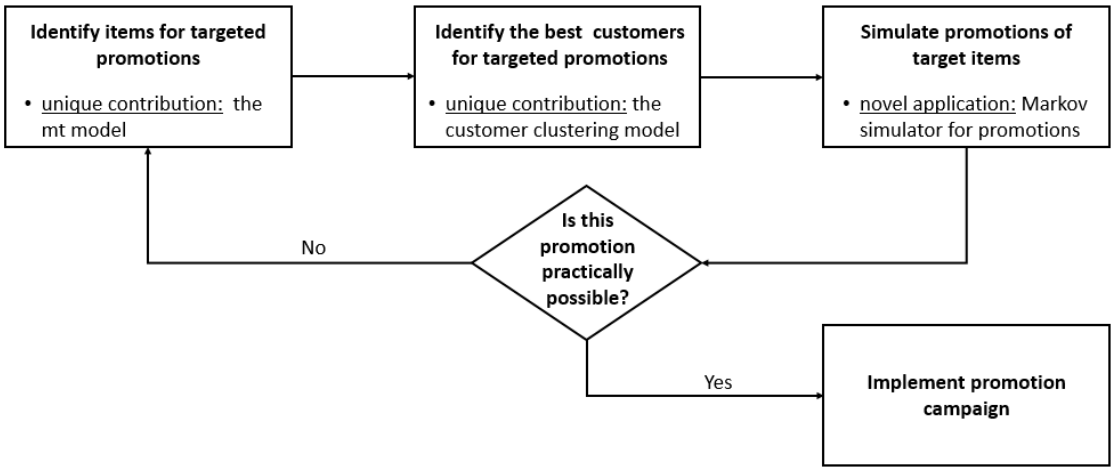


Figure 1.1: High-level flow chart of the proposed data mining models and algorithm

The mathematical models and algorithm developed for the grocery retail sector was found to have broader applicability. Indeed customers are people, and people engage in a wide variety of activity beyond retail. Given this, an additional element of this research focussed on applying the model to other scenarios where predicting behaviour can be beneficial. In this regard, applications in education, medicine and crime prevention were investigated, and is detailed in Chapter 6.

1.1 Overview of the UK Grocery Retail Sector

The UK grocery retail sector is considered central to the approximately 64 million UK residents who remain reliant on it for food and everyday household goods [87]. The analysis in [87] and [91] showed that the UK grocery retail sector accounted for 20 percent of all UK consumer spending in 2019 at $210 billion and is projected to grow to $220 billion by 2023. The sector is made up of several categories of grocery retailers including the "The Big Four" supermarket chains, Tesco, ASDA, Sainsbury's and Morrisons, who collectively account for over 70% of the UK grocery retail market. This is followed by discount brands Aldi and Lidl, the high-end brand Waitrose, and the local, convenient store specialist, Co-op, who collectively account for a further 25% of share. The remaining 5% is made up of local convenient stores, smaller high street brands, independents as well as non-food majors that are increasingly selling groceries and foods like Marks and Spencer [87][91].

Despite the projected growth, the sector has been undergoing major shifts in recent years with all major players scrambling to retain or grow market share [13]. These shifts include: (1) customers splitting their purchases between retailers more frequently than ever before, (2) stiff competition and disruptive practices amongst grocery retailers, (3) the rise in online retailing, and (4) the use of smaller, more frequent shopping trips as a result of tighter household budgets [13]. Whilst some players, like the discount chains, are responding to these shifts by perfecting their pricing strategies, others like The Big Four, have acknowledged pricing pressure and are focussing on non-price competition [172]. Whatever the focus, all retailers are now placing a greater emphasis on customer relationship management and customer data analytics [59][172].

1.1.1 Loyalty Programs and Frequent Users

Loyalty programs being offered by grocery retailers in the UK have become increasingly popular over the past few years, thanks to both the attractive rewards being offered to customers, and the successful marketing campaigns undertaken by large grocery retailers [42]. Indeed in the UK, loyalty program penetration, i.e. customers belonging to at least one loyalty program, is at 90% and ranks amongst the highest in Europe [42]. Customers and UK grocery retails both benefit from these programs, with customers being rewarded with vouchers for increased spending, while UK grocery retailers use this as a powerful tool to identify, reward and retain profitable customers [42][101]. The findings in [42] are relevant for this study, particularly in understanding the behaviour of frequent, known customers, which is sometimes referred to as known users throughout this study:

1. Loyalty programs primarily attract existing customers.
2. Loyalty programs have greater increases in purchasing behaviour in Low/ Medium volume customers.
3. Customers see rewards from loyalty programs as an investment in relationships.
4. Loyalty programs can increase product sales by between 4% and 25%.

It is thus clear that grocery retailers with loyalty programs benefit immensely from these programs, including from the wealth of data that they gather from each subscribed customer. However, those without loyalty programs have also realised the value of data and have since become creative in their data gathering efforts. In this regard, they continue to leverage data sold by banks, suppliers, and other third parties [72]. One example of a reliable data source is banking card transaction data, which grocery retailers are able to purchase from banks, without customers' personal data. Using this data, grocery retailers are able to assign unique "dummy" customer details

to each bank card used across all of its stores and this serves as a means for future identification of individual customers. Hence, all future transactions made using a known bank card, across any of the grocery retailer's stores are recorded against that specific customer, and it is from this that grocery retailers are able to understand customer shopping patterns and preferences with their stores [72].

Given this, the definition of a known customer or user that was used in this study is: a customer or user that is identifiable to the grocery retailer, and whose individual transactions, contained in the grocery retailer's database, can be assigned to this customer or user. Hence for this study, a frequent, known customer is defined as a customer that has visited the grocery retailer's stores at least twice during a given time period. In this study, the time period was one calendar year.

1.1.2 Promoting Items (Target Items and Target Customers)

The concepts of sales promotions as outlined in [126] was found to be comprehensive and was used throughout this study. According to Ozer et al. [126], sales promotion is defined as "an action-focused marketing event whose purpose is to have a direct impact on customer behaviour". Further, it is typically a temporary offering and in the form of either one or a combination of: (1) retailer promotions which are offered to consumers by retailers to increase the sale of an item or category or store, (2) trade promotions which are offered to stores so that they can offer discounts to their customers and stimulate that sales channel, and (3) consumer promotions which manufacturers offer to consumers to stimulate sales at a given point. It should be noted that permanent price reductions are different to sales promotions in that they are indefinite, and whilst they may create the same initial customer switching reaction as sales promotion, customers are less-likely to return to "old ways" in the case of permanent price reductions as prices remain reduced and loyalty increases [141][154][155].

The notion of "targeting" that has been used throughout this study is as outlined in [55] and [140], where a specific item or set of items is being promoted to a specific customer base. Consequently, not all items are on promotion at the same time, as this will erode profitability and create industry-wide pricing pressure [133], and not all customers are targeted with the same offer at the same as this has been shown to be ineffective, as it wastes resources and annoys customers who receive irrelevant marketing material ("false positives") [110][116][182]. Hence, a target item is defined as an item that is either on a sales promotion or has recently had a permanent price reduction. A target customer is defined as a customer who receives marketing material for target items based on their individual, historic shopping preferences. These definitions were used throughout this study.

1.2 Customer Analytics and Market Basket Analysis (MBA)

The benefits of data analytics in retail is now mainstream, and in line with this, all major UK grocery retailers have embraced data analytics, in some form or another, in an attempt to develop their customers, enhance loyalty and consequently increase their sales and/or profitability [59][154]. Given this, customer data analytics, in particular MBA, has become increasingly popular from a commercial perspective, giving rise to the use of two popular techniques namely: Associated Rule Mining (ARM) and Clustering (or Cluster Analysis) [19][120][154]. MBA may be defined broadly as the process of studying individual customer's or a collection of customers' grocery shopping transactions with the aim of gaining valuable insight on shopping patterns, customer behaviour, etcetera [5][120]. In this regard, MBA is typically conducted by individual grocery retailers themselves, but is also conducted, albeit at a smaller scale, by third party analysts and academics [77][90].

ARM was popularised by the work in [5] where the concepts of support and confidence were used to define associations between two or more unrelated items contained within a grocery retailer's transaction database. The support of an item (X) or supp(X) is defined as the ratio of the number of transactions containing X to the total number of transactions, and the confidence of purchasing item (X) leading to the purchasing of item (Y) or conf($X \rightarrow Y$) is the ratio of the number of transactions containing both (X) and (Y) to the number of transactions containing (X). Mathematically, the concept of support is equivalent to the principle of probability, hence supp(X) = $P(X)$, where $P(X)$ is the probability of item (X) being present in a particular transaction within a transaction database. Similarly, conf($X \rightarrow Y$) may also be expressed in terms probability, with conf($X \rightarrow Y$) = $P(X,Y)/P(X)$. Items are said to be frequent if their support in the transaction database is equal to or exceed a minimum, user-defined support threshold, commonly referred to as *minsup*. Similarly, items are considered to be associated if their confidence exceeds a user-defined minimum confidence threshold, commonly referred to as $minconf$. It should be noted that confidence is not commutative. For example, if $P(X)$ is considerably greater than $P(Y)$ for a given transaction database, then conf($X \rightarrow Y$) may not be associated but conf($Y \rightarrow X$) may be associated due to the differing influences of their respective denominators.

Clustering is described in [77] as the process of grouping together data objects that are similar to each other, and are collectively dissimilar to other objects in other groups. Within MBA, clustering has typically been used to group similar products, transactions and customers with its benefits usually being realised when it has been combined with ARM. For example, in [118], it was shown that the efficiency of collaborative filtering recommender systems was enhanced when transactions within a large, sparse

dataset were clustered prior to performing ARM to find associated items. Similarly in [140], customers were clustered prior to ARM being performed to find associated items within each customer cluster for product targeting.

Given the above, and the competitive nature of the sector, it is of little surprise that much of the work conducted by the individual grocery retailers remain highly confidential, with their primary source of data being their own transaction databases. Data analytics is clearly seen as a competitive advantage in the grocery retail sector and the sharing of techniques, results and data could result in a grocery retailer neutralising its differentiation from its competitors [59]. Consequently, there is little published literature on the exact nature of the analytical techniques used by individual retailers. However, researchers continue to develop algorithms, models and frameworks, in some cases independent of grocery retailers, with the aim that these solutions will likely be leveraged by grocery retailers to enhance their businesses, and by other practitioners to enhance their operations in their respective applications [59][77][172].

1.3 Motivation for this Study: Gaps in MBA-related Association Rule Mining (ARM) and Clustering Research

Data-driven decision making is becoming increasingly important across most sectors as organisations look to harness data to better support its stakeholders [135]. Whilst there is a plethora of research in both ARM and Clustering related to MBA, their focus has largely been on optimising algorithms to enhance processing speed, minimizing computer memory usage, optimising the quality of clusters, and finding new extensions and applications for ARM [152][180]. In this regard, MBA techniques have proved to be very useful in finding associations and segmenting markets for targeted promotions, all of which have been underpinned by robust mathematical principles

(for example: Apriori, Bayes Theorem, etcetera) that have been leveraged extensively in developing ARM algorithms [5][77]. However, these algorithms usually provide a large set of associations and selecting the best associations from this set for marketing purposes, i.e. identifying target items for promotions, is not always easy as grocery retailers typically sell thousands of individual items, have limited marketing budgets, and are continuously under pressure to maximise sales and profitability [87][126]. Thus, successful marketing campaigns are reliant on both successfully selecting the right target items, and successfully identifying the right customers that will be most receptive to these promotions [120][140][141].

The notion of combining variables to form interestingness measures from which decisions can be made, has been well-studied over the years. Indeed, thirty eight such measures were noted in [65] and [104] and twenty one were compared in [161]. It should be noted that interestingness measures are highly dependent on the context in which they are being applied, hence the need for the plethora of measures. Consequently, while one measure may be best suited to one context, it may to be totally unsuitable to another [161]. In this regard, the cosine measure was found to be best suited to retail settings [161]. A common thread across most measures was that they largely centred on Piatetsky and Shaprio's principles for rules interestingness [65][129][161]. The Piatetsky and Shaprio principles are outlined as follows:

1. Statistical Independence: rules are not interesting if they are statistically independent, i.e. the observed probability of a combination $P(A, C)$ is equal to the theoretical probability, $P(A, C) = P(A) \cdot P(C)$, where $P(A)$, $P(C)$ and $P(A, C)$ are the probabilities, or support of items (A), (C), and (A, C) respectively.

2. Monotonicity: measures that combine probabilities should be strictly monotonic with respective to the variable being adjust while all else remains fixed.

The decision making scenarios, based on the concepts outlined in several works including in [5], [65] and [161], may be classified as: (1) same antecedent, i.e. $(A \rightarrow C)$ versus $(A \rightarrow D)$, (2) same consequent, i.e. $(A \rightarrow D)$ versus $(C \rightarrow D)$, and (3) independent sets, i.e. $(A \rightarrow C)$ versus $(B \rightarrow D)$, where (A), (B), (C) and (D) are itemsets contained in a grocery retailers transaction database. Clearly decisions involving types (1) and (2) are specific cases of the more generalised case of type (3).

1.3.1 Gaps in Decision Making using MBA

Research, as part of this study, has found that no consistent, robust model exists to support the decision making process for selecting target items [115]. Indeed, there may be proprietary models that are being used by various firms that are not in the public domain. Tests conducted in [115], which formed part of this study, using type (1) decision making showed that the uninorm outperformed the cosine measure as well as three other popular measures for interestingness. However, the uninorm and other measures were found to be inconsistent with regards to both obeying the Piatetsky and Shaprio principles and from a real-life, practical perspective, when extending this measures to model type (3) decision making [115]. Given this, there remains opportunities to advance the body of knowledge around decision making support systems to enhance the selection of target items, leveraging ARM, clustering and MBA. It is against this backdrop that this study adds value to the greater body of academic research in the field of data analytics, more specifically, data-driven decision making.

1.4 Research Hypothesis

The hypothesis of this research was: an algorithm, based on a mathematical model developed from existing research, that optimises the selection of associated items for

targeted promotions, and categorises frequent, known users, using their transaction history, will enhance the prediction of missing targeted items from these known users' future transactions that contain frequent itemsets.

1.5 Research Aims

There were four specific aims of this study, the first three related to validating the research hypothesis whilst the fourth was related to extending the model and algorithm to other fields. The aims were:

1. Leverage existing mathematical frameworks to develop a mathematical model to optimise the selection of associated items, i.e. selecting the best item to target for promotion.

2. Incorporate the mathematical model into a computer-based algorithm that includes categorising frequent, known users to predict missing targeted items from their transactions, i.e. targeting the right customers with the right target items.

3. Validate the predictive accuracy of the developed algorithm by comparing the predictions made by the algorithm against historical, real-life data and comparing its performance against similar published models.

4. Demonstrate the use of the mathematical model and the algorithm in other applications, and outline possible areas for future research.

1.6 Unique contributions of this Research

This research has made unique contributions not only to the field of artificial intelligence, and more specifically data analytics and MBA, but also to several everyday sectors including grocery retail, education, public health and crime prevention. It

has provided a compact mechanism which practitioners can leverage to enhance their understanding of the underlying dynamics of common issues within their respective sectors. Several case studies have been conducted as part of this study to illustrate this point. The unique contributions are outlined as follows:

- At the heart of the unique contributions of this study was the development of the mathematical model, referred to as the "market target" or mt model, which can aid decision making, particularly where the choice between alternatives is not obvious. The principle behind the mt model is that it measures the "effort" required to achieve the desired outcome from the current state. Hence by comparing the output of the mt model, the mt value, of different alternatives, practitioners are thus able to select the optimal alternative and take the necessary action to achieve the relevant outcome.

- The second major, unique contribution of this study was the development of a customer clustering approach, using fuzzy c-means clustering. This approach segmented the target customer population, based on their loyalty to the grocery retailer, which then created a platform from which the grocery retailer could launch customised incentive programs to each cluster.

- The third major, unique contribution of this study was the development of an algorithm to enhance the prediction of missing targeted items from the transactions of frequent, known customers. The algorithm leverages the mt model and the clustering approach, and was able to identify the best group of customers for targeted marketing promotions for specific items. The algorithm demonstrated superior performance, when compared to the published model in [140], which has a similar objective.

Other novel contributions of this study were largely centred on the diverse applicability

of the mt model and targeting algorithm. These were:

1. Improving school attendance by isolating the most problematic school session and addressing poor attendance through improved attendance rewards mechanisms.

2. Optimising public health campaigns and research by quantifying the true impact of the underlying causes of diseases, suggesting treatment priorities, and fostering rapid hypotheses development for future research programs.

3. Enhancing the efficiency of crime prevention initiatives by providing an easy-to-use, tangible metric that can be rolled out to "on the ground" police officers, thereby enhancing the effectiveness of operations and the cohesion of police with communities through the reduction of "false positives".

1.7 Research Ethics

Research ethics was an important consideration for this study given that it leveraged proprietary and sensitive data. In this regard, there was strict adherence to the ethical requirements of De Montfort University (DMU), as outlined in [41], and the data privacy and data handling requirements of third party data providers. Ethics approval was sought from the DMU ethics committee prior to commencement of this study. The committee required that a non-disclosure agreement be signed with any third party providing proprietary and/or sensitive data.

1.7.1 Shopping Data

Grocery shopping data was provided by Kantar for the sole use of this study [90]. In this regard, the data that was provided was highly confidential and proprietary to Kantar. The data contained identifiable product brands, pricing information and

grocery retailer names. Given this, a non-disclosure agreement was signed between DMU and Kantar which stipulated that the data was to be used for the sole purpose of this research and that all publications should contain anonymised data representations only. Further, the data is not to be shared with any third party or be used for other research without the permission of Kantar.

1.7.2 School Attendance Data

School attendance data was obtained from Willen Primary School (WPS) to demonstrate the application of the market target model to other fields. In this regard, the author worked closely with the school leadership to ensure that all data provided was completely anonymised. Further, the school leadership team ensured that the data sharing was fully compliant with all of WPS's policies, including compliance with the General Data Protection Regulation (GDPR).

1.7.3 Handling of Data

All data was handled with extreme care throughout the duration of this study. All grocery retail data was initially anonymised and then processed further. Each grocery retailer, or store, was assigned an arbitrary, non-identifiable, number, e.g. Store 20, and referred to in this manner throughout the study. All items and customers were treated similarly. Permission will be sought from both Kantar and WPS to retain the data and to use it as part of the future work outlined in this study or as part of similar studies. If permission is denied, then all data will be returned to their respective owners.

1.8 Dissertation Structure

The remainder of this dissertation is organised as follows:
Chapter 2 provides a critical review of previous work done in consumer retail, artificial intelligence, and data analytics. The aim of this review was to both identify gaps in research that would be filled by this study, and identify concepts that could be leveraged as part of this study.

Chapter 3 discusses the research philosophy and methodology that was adopted as part of this study to develop the unique contributions that added to the body of knowledge on data mining and grocery retail. In this regard, it provides: (1) a review of well-known research methods, (2) provides a justification for the research approach adopted as part of this study, and (3) provides the principles on which the proposed algorithm was validated.

Chapter 4 details both the development of the mathematical model and algorithm that formed the basis of the contributions of this study. The chapter provides a "first principles" development of the mt model which is underpinned by contributions of previous works.

Chapter 5 is dedicated to the UK grocery retail case study where a detailed account of the use of mt model and algorithm in the grocery retail sector is provided including the experimental approach, results and discussion. The chapter also details simulation tests that were conducted to demonstrate the effectiveness of the model in achieving the business objective of targeted promotions. Further, the predictive accuracy of the mt model and the algorithm was compared to both another similar published model, and the actual shopping patterns of the customers on the scanner panel data provided by Kantar. Finally, and for completeness, the processing performance of the model

and algorithm was also tested on synthetic datasets.

Chapter 6 details novel applications of the mt model and algorithm to other applications including in education, public health and crime prevention. The chapter also highlights innovative approaches that were taken as a consequence of the results generated from the application of the mt model in school attendance, and the impact it has made in improving attendance and learning as a whole.

The dissertation concludes with Chapter 7 were the findings and results are summarised together with the research limitations, the wider impacts of this study, and the opportunities for future work.

Chapter 2

Literature Review

2.1 Introduction

Data analytics research remains very active thanks to the recent explosion of data being produced by people and organisations across the globe as a result of widespread online transacting and social media use [72][82][160]. The primary objective of this study was to add to the growing body of research, in particular, research on decision making between two independent alternatives. In this regard, a key contribution of this study was the development of a novel algorithm for targeted promotions in grocery retail. Whilst the main thrust of this research is grocery retail, the model and algorithm has been applied to other fields as well. An application-specific literature review for these other fields is included in Chapter 6. The objective of this chapter is to present a critical review of existing literature and is divided into four sections:

1. **Data Analytics:** this section starts broadly with a review of the overarching concept of knowledge discovery and data mining (KDD). It then hones in on Market Basket Analysis, more specifically Association Rule Mining and Cluster Analysis, which are essential elements of this research. For completeness, Recommender Systems, which remains a "hot topic" in online transacting, is then discussed.

This is then followed by a review of common algorithms used in ARM and Cluster Analysis.

2. **Data Analytics in Grocery retail:** grocery retail is central to this study, and in line with this, a critical review of the data analytics research related to grocery retail is presented, including commentary on the techniques that are commonly used, together with the current challenges and areas that require future focus.

3. **Decision-making models in MBA:** modelling of the decision-making process between two alternatives is a key theme of this study. Indeed, one objective of this study is to develop a model to enhance item and customer targeting. This section reviews the existing literature on modelling techniques, commonly used models, and the pros and cons of each approach.

4. **Summary:** the chapter concludes with a summary of considerations from existing research that: (1) identifies the main gaps in existing knowledge and discusses how these will be addressed in this study, and (2) highlights key concepts from existing literature and discusses how these will be leveraged as part of this study.

2.2 Data Analytics

2.2.1 Knowledge Discovery and Data Mining

The philosophy of Knowledge Discovery in Databases (KDD) and data mining, outlined over two decades ago in [51], a little later in [67], and more recently in [160], has remained consistent and still widely accepted. KDD is seen as the over-arching process of finding knowledge from data through the use of data mining, with data mining being the application of algorithms to extract patterns in the data [51][67][160]. The restrictions placed on the definition of patterns are important to the understanding of KDD and knowledge [51]. Given its importance, the definition of patterns is de-

tailed as follows: patterns are generally referred to as being able to summarise data or describe, through some language, a subset of the data. Such descriptions should be non-trivial, in that it should not be obvious, and must be valid for new data with a defined degree of certainty. Further, patterns should be beneficial to the user and task, with the quantification, value, of the benefit being its interestingness. Patterns that meet these criteria are referred to as knowledge, a concept central to KDD [160].

MBA has long since been viewed as a cornerstone of modern KDD processing [24][77], in that it produces knowledge and patterns that are beneficial to its primary users, which in this case are usually grocery retailers. Thus, it can be concluded that ARM and FIM are examples of the data mining techniques used to derive patterns, or knowledge, based on predefined values for $minsup$ and $minconf$, which may be considered as interestingness. Hence, knowledge in this instance is the association rules and frequent itemsets that result from the use of ARM and FIM algorithms on a dataset [24].

2.2.2 Market Basket Analysis (MBA)

Several studies have noted that research into MBA has become increasingly popular as a result of the rapid expansion in consumer retail, which has been driven by several factors including, competition and consumer-choice [43][120][140][153]. Researchers view MBA as part of customer knowledge and consisting of three aspects: (1) knowledge for customers, i.e. product knowledge, store layout knowledge, etcetera, (2) knowledge about customers, i.e. who they are, what they buy, when they buy, etcetera, and (3) knowledge from customers, i.e. what they know about products, what they know about the competition, etcetera[95][120][153]. Indeed, MBA has been cited in several studies, as being an effective tool to foster good customer relationship management (CRM)[120][153][164]. In [120], the primary purpose of MBA has been described as a

mechanism to consistently expand customer transaction intensity, transaction value, and individual customer profitability. A similar notion was highlighted in [153], where the underlying idea of MBA is that consumers rarely make purchases in isolation, and by carefully studying their purchases, retailers can develop interventions to influence their purchasing behaviour and thus enhance sales. Similarly, in [164], the focus of MBA was on the effective mining of purchasing patterns, and using these purchasing patterns to predict the future purchasing behaviour of customers.

At a practical level the findings in [43], [69] and [167] all showed that MBA is being used by the most progressive grocery retailers to understand and predict customer purchasing behaviour, with the aim of maximising consumer sales. As a consequence of this, grocery retailers have provided their operational teams with real-time MBA, which has allowed for rapid adjustments to be made to the business, and thus enabled the entire organisation to work smarter, while enhancing revenue and profitability [43][69][167]. Both retailers and third parties, including software and business consulting companies, have now realised the commercial opportunities in MBA. Consequently, they have developed software and tools that grocery retailers can use to leverage customer data, and maximise revenue and profit generation, whilst effectively marketing the "benefits" to the customer to produce a "win-win" scenario [4][43][167].

The success of MBA in retail has seen it being applied innovatively to other fields including bio-medicine, crime prevention, web mining and geo-spatial sciences [52][80] [160][171]. In [171] MBA was used to understand the symptoms and their association with cancer, whilst in [80], MBA was leveraged as part of a crime prevention strategy to detect crime hotspots, and link related criminal incidents. Similarly, in [52] MBA was used to predict weather patterns using meteorological data, while in [77] MBA was used in applications to personalise user experiences based on web-page access

data. Based on the vast volume of research and the array of applications, it is clear that MBA is an effective KDD process. Given this, it was decided that MBA data mining techniques including ARM and clustering will form the core of this study, and thus have been explored in detail in sections that follow.

2.2.2.1 Association Rule Mining in MBA

ARM, first proposed in [5], and detailed more recently in [160], is generally recognised as the *de facto* data analytic technique for MBA [24][140][164]. This is largely due to the substantial growth in research, both in depth and in breadth, as a result of its widespread applicability in retail and various other fields [160][171]. In this regard, ARM research has become so popular that it is often attributed to being a catalyst for modern data mining research [24][160].

The original concepts outlined in [5] is now known as "traditional ARM" as it deals with intra-transactional ARM, and is typically seen as the trivial case for more complex ARM models like inter-transaction ARM, frequent itemsets obtained from multiple transactions, and sequential pattern mining, FIM where the order in which the items appear in a transaction is important [6][183]. Note that the triviality of "traditional ARM" arises as both inter-transactional and sequential pattern mining can be reduced to a single transaction by either extending the timeframe or combining transactions or both [52] [183]. Whilst it is useful to understand which items made up a frequent itemset, it was very clear to researchers and grocery retailers alike, that understanding how the itemset was put together was equally important [120]. This was one of the main goals of ARM, that is, to determine the antecedents of a frequent itemset, and to use this information to predict customer behaviour, which in turn informed grocery retailers on store layout, promotions, etcetera [152]. In this regard, the notion of confidence, as defined initially by [5], remains the basis for most ARM

analysis [152].

Sequential Pattern Mining

Mining sequential patterns gained interest following its introduction by Agrawal and Srikant, in [6]. Research into sequential pattern mining, like ARM, extended in several directions with the key areas of focus being the development of algorithms to enhance the process, and finding new applications [64][78][186][188]. Unlike "traditional ARM", sequential pattern mining may be considered to be more restrictive, in that it focusses on mining sequential and not necessarily consecutive, sequences of itemsets in data. Given that the order in which the sequence occurs is important, frequent items purchased in different orders may be discarded, thus sometimes presenting a false view of the underlying reality [160]. Whilst there may be several suitable applications for sequential pattern mining, including medicine, crime prevention, and predicting weather patterns, it is not ideal for grocery retail as it may increase the frequency of "false negatives" [160].

Given this, it was concluded that sequential pattern mining was not relevant for this study, as it was far too constrained, in that grocery shopping patterns, in the current UK context, are not generally rigidly sequential and are dependent on several other factors including convenience, price, and "memory-jogs" whilst in store [141][154][155].

Inter-transactional ARM

Research into inter-transactional ARM sought to address the lack of contextualisation that often results with both "traditional ARM" and sequential pattern mining. This is particularly noticeable in applications that involve prediction, e.g. weather prediction and stock-price movements [52]. It was shown in [52], [53] and[107] that adding contextualisation to transactions, through the use of multiple dimensions to describe the

properties of transactions, enhanced the applicability of inter-transactional ARM as a predictive tool. The concept of inter-transactional ARM has been well-represented using of the one-dimensional example outlined in [52], [53], and [107], and is detailed as follows: Transaction t_1 made today in store 1, containing item i_1, and transaction t_2 made tomorrow in store 2, containing item i_2, leads to transaction t_3 being made the following day in store 3, containing item i_3, i.e. $t_1(i_1)|_1,\ t_2(i_2)|_2 \rightarrow t_3(i_3)|_3$. From this simple example, it is clear that the major drawback for grocery retail shopping is its specificity, in that generalized patterns can be missed. Inter-transactional ARM creates items by grouping together location and product pairs, for example *item i = (store 1, product 1)*. Consequently in the ARM phase, algorithms search to find associated items to *item i* either in the same transaction or other transactions by the same user, for example *item j = (store 2, product 2)*. Hence an association rule may be *item i* $\rightarrow$ *item j*. If however in the following day the customer purchases both *products 1 and 2* from *store 1*, then the algorithms outlined in [53] will not consider this as an associated item set. As a result, this approach, whilst relevant to other applications, does not adequately represent the modern shopping context as outlined in Section 1.1, and was not pursued further.

Hybrid Inter-transactional / Sequential Pattern ARM

The approach taken by [183] may be in many ways considered as a combination of both inter-transactional ARM and sequential pattern mining. The findings in [183] makes two important contributions to this study. Firstly it served as confirmation that models using the combination of sequential pattern mining and inter-transactional ARM also reduce to "traditional ARM" when all dimensionality is removed, and secondly it indirectly highlighted the competitive dynamics that exists in the UK grocery retail sector, and through this, how some frequent items may be ignored by all grocery retailers. A key drawback of the Online Shopping Pattern algorithm, OSP, proposed in

[183], is that it mined frequent patterns where the products had to be purchased from the same site or store. Hence, there is a possibility that items purchased frequently, but from different stores on each occasion may be missed.

From a practical retail perspective, Yang et al. in [183], highlighted that one application of their analysis could be a group of providers coming together, sharing their data, and working in synergy, e.g. airlines working with car-rental companies and hotels to provide a holistic "vacation service". Whilst this is certainly possible, and may be beneficial to all providers for some sectors, it is unlikely, if not illegal, that it would be possible in the highly competitive UK grocery retail sector, where data sharing and collusion could be construed as price fixing and customer exploitation [72].

Like with the model proposed by [53], the model proposed by [183] did highlight some useful concepts. However, the approach and algorithm was not considered suitable for the grocery retail sector as its restrictive approach resulted in some vital information being lost, e.g. where frequent items are split across multiple retailers.

2.2.2.2 Cluster Analysis in MBA

The use of cluster analysis in MBA has been largely along one of two paths namely transaction clustering, as detailed in [100], [156], and [185], and product clustering, as detailed in [75], [169], and [179]. In transaction clustering, the focus has been on clustering similar transactions with the aim of profiling customers that engaged in these transactions, whilst in product clustering, the aim is to group similar or associated items into clusters.

Product Clustering

In general, product clustering to group associated items in grocery retail has had lim-

ited success, and is thus not widely used [77]. This is largely due to the simplicity and robustness of ARM which remains the preferred method [77]. Two notable attempts to create alternative models to ARM have been the *hypergraph* and *hyper-clique* models as detailed in [75] and [179]. Although the *hypergraph* product clustering approach was considered to be an alternative approach to ARM, a key element of this method was the generation of frequent itemsets using the "traditional ARM" Apriori algorithm, as detailed in [5]. Thus, product clustering using hypergraphs was not truly an alternative to ARM as it leveraged ARM techniques as part of its solution [75].

The *hyper-clique* product clustering approach taken in [179] was in essence an enhancement of the *hypergraph* approach and focussed on clustering products that were highly associated by addressing the generally accepted challenge around setting the value for minimum support ($minsup$) used in FIM. If $minsup$ is set too high, then algorithms eliminate highly associated rules that have low support, also referred to as rare rules in [99]. A novel contribution of the work in [179] was the introduction of the cross-support property, which eliminated clusters that contained itemsets with vastly differing support values. The rationale of the cross-support property is that if items x and y are items in a cluster, then for a predefined threshold value t, all patterns where $\text{supp}(x)/\text{supp}(y) < t$ can be eliminated as these patterns contain items with vastly differing supports. On closer examination, it is difficult to see how this is different from the notion of $minsup$ in ARM, as one still had to specify the value of t.

The seed generation product clustering approach was proposed as alternative to the *hyper-clique* in [99] and [100]. The seed generation technique grew strongly associated items by firstly generating frequent itemset seeds through the use of the Apriori algorithm and then through tight constraints combine seeds to create new frequent items that are highly associated. Here again, this method relied on ARM as part of

its processing.

As noted from the above, clustering to group associated items is generally more complex than ARM, with several algorithms leveraging ARM as part of the process anyway. It is thus not surprising that clustering is not seen as the approach of choice for FIM/ ARM [77]. Given this, product clustering as a technique to find associated items, was not considered a viable option for this study.

Transaction Clustering

In [170], [181], and [184], the focus was on grouping transactions into clusters based on the presence of large itemsets, commonly referred to as frequent itemsets. As noted in [181], large itemsets were an effective way of measuring the similarity of a cluster with an ideal cluster. An ideal cluster was defined as having a high ratio of large itemsets to small itemsets, and where these large itemsets were unique to one cluster. In these studies, transactions were allocated to existing clusters or assigned to new clusters based on a cost function that assessed the similarity of items within the transaction to existing clusters [181]. Analyses conducted in [100] found that in many cases this approach failed to find good representations of large itemset clusters as it "encouraged" the formation of smaller, higher support clusters as opposed to larger clusters that exceeded the minimum support threshold. Consequently, this made it difficult to implement from a practical perspective, as there were far too many clusters for targeting purposes [100].

Transaction clustering to find like-minded customers has been effective, with several algorithms developed over the years [156][181]. The OPOSSUM algorithm, developed in [156] to cluster like-minded customers, had one major drawback, in that it was focussed on finding like-minded customers and placed very little emphasis on the

quality, or interestingness, of the associations between items. This was partially addressed in [181] where clusters in caucuses was developed. This included a description of both the customer profile and their buying behaviour. Caucuses were created by defining starting customer profiles and transactions were then assigned to caucuses based on the matrix procedure similar to that detailed in [156]. These caucuses were then refined, i.e. their centroids were adjusted, through an iterative process. Whilst caucuses certainly added a valuable dimension to the clustering, it presented with similar specificity issues highlighted in the inter-transaction ARM approach where, a large itemset can belong to multiple caucuses and the true nature of the behaviour of the large itemset is difficult to ascertain [99]. The other issue with the caucus approach, as highlighted by [99], is that it required a demographic understanding of the customers to adequately describe a caucus. However, this may not necessarily be an issue for the UK grocery retail sector as most UK grocery retailers already have loyalty card systems in place where customer demographic data has been captured.

2.2.2.3 Hybrid Approach for MBA

Combining ARM and clustering has been an approach taken by researchers who created models that found associations between itemsets that have high confidence, but have low support, also referred to as rare rules [40][99][130]. Typically, such rules are not be detected through "traditional ARM" due to their low supports, and consequently the knowledge of their associations are usually lost. In this regard, clustering was used as a preprocessing step. The clustering step focussed on partitioning the dataset into local zones, either by transactions or by items, which resulted in these rare items being clustered together with the frequent items either not present, or split across these clusters. Upon isolating these local clusters, "traditional ARM" techniques were then used to mine associations within the clusters [99][130]. Whilst

this technique is very useful in finding rare rules, which could be used in identifying symptoms associated with rare diseases, or co-locating niche items that are frequently bought together in a store, it does not directly lend itself to FIM.

Several studies have been conducted that combined clustering and ARM in a attempt to enhance targeted promotions including [25], [106] and [140]. In this regard, clustering was used to form customer clusters by grouping together like-mined customers, and ARM was then used to find associations between items within each cluster. Whilst this is a formidable approach, and was used as a basis for comparison in several places throughout this study, it is believed that clustering customers before ARM can lead to increased "false negatives" as some customers who buy the targeted product may reside in a customer cluster that is not being targeted, and hence missed. This hypothesis was tested as part of this study and the results of the testing is discussed in Chapter 5.

2.2.2.4 Recommender Systems

Recommender Systems (RS) is largely being used for both predicting customer purchases and nudging customers towards certain products [154]. The "prediction" mechanism of RS has been best highlighted by Adomavicius and Tuzhilin, in [2], where the primary goal of RS is to calculate, based on several variables, the rating that a given customer will assign to an item that is unknown or new to them. If this calculated, or predicted, rating is sufficiently high, then this unknown item becomes a recommendation for this customer.

Much RS research in recent years has focussed on improving the ability of RS to contextualise transactions and deepen its understanding of the customer to avoid "false positives". "False positive", in this case, is defined as item that is wrongly

recommended to a customer, and the customer does not buying this item [108] [145]. Whilst "false positives" can sometimes result in wasted marketing spend for the retailer, it can have more serious impacts from a CRM perspective. In this regard, customers end up annoyed with the system, lose trust in the retailer, and may take their custom elsewhere [108][145]. Another issue that arises across RS and currently remains a major focus on RS research is the "new customer / new item" problem, commonly referred to as "cold start" problem [108]. "Cold start" is where a new item or customer enters the system and the RS is unable to provide recommendations within its accuracy threshold as it does not have the historical data [108]. Consequently, the RS does not offer or make incorrect recommendations to the new customer which may lead to "false positives" or "false negatives", and result in early customer defection [108]. Three popular approaches for RS highlighted in [2] remain popular today, namely: Content-Based Filtering (CBF), Collaboration Filtering (CF), and Hybrid RS, a combination of CBF and CF [96][154]. These are discussed further in sections that follow.

Content-based Filtering (CBF)

Despite being part of early research in RS, CBF is still widely researched today as it remains relevant to several fields including in web-based searches, where a customer viewing content related to a specific item is offered recommendations that are similar to, or associated with, the item under review [108]. Whilst RS based on CBF can be applied to grocery retail, where it can be used to provide consumers with recommendations for similar or associated products, its main disadvantage to retailers is over-specialisation [21][127]. Over-specialisation results when the recommendations are so similar to the original product that the customer does not try the new products, and the retailer is unable to expand the customer's spending in their store [21][127].

A recent, new angle for CBF has been the concept of enhanced contextualisation, driven largely by the prolific rise of social media usage over the last few years [108]. In this regard, organisations, including retailers, in conjunction with their own and third party social media sites, have seized the opportunity to gain knowledge of their customers by allowing customers to: (1) tag web pages, products and services with keywords that are meaningful to them, (2) follow others, (3) be followed by others, and (4) show preference, e.g. likes, dislikes [108]. It is through this process that retailers have gained greater contextualisation and are able to further customise content that is ideally "suited" to that customer [1][108]. In the grocery retailer context, one example may be where a retailer selling ready-meals updates its internal profile of a frequent, known customer based on a recent social media posting where the customer has tagged a dislike for mushrooms. Following this update, all future recommendations to this customer for ready meals could exclude mushrooms.

Collaborative Filtering (CF)

Unlike CBF, that uses similar items to make recommendations to customers, CF is based on the notion of recommending items to a customer given what other similar customers have purchased [1]. Several, well-researched CF systems including GroupLens, Video Recommender and Ringo are available but these are based on either one of two approaches namely: *memory-based* where the entire collection of rating data is used, or *model-based* where previous ratings data is used to build a model that offers predictions [2][28]. Although the *memory-based* approach is simplistic as it is the averaging of ratings from a group of similar users, it is computationally expensive and requires a substantial database of users in order to be effective. On the other hand, the *model-based* approach proposed by [28], although touted as a CF model, actually resembles a CBF model, as it is based on the probability of a user rating an item given how they rated other similar items previously.

Although CF is widely used, researchers in the field of RS dismiss CF in favour of CBF claiming that CBF has been shown to provide both better computational performance, and better predictive accuracy than CF [2][108][146].

Hybrid Recommender Systems

Hybrid RS remains an area of active research particularly as it looks to combine CBF and CF to both enrich the RS, and offset the individual disadvantages of CF and CBF [2][108]. Four general approaches are taken in formulating hybrid systems: (1) using CF and CBF separately and combining the output to form a single recommendation, (2) using CBF in CF systems, (3) using CF in CBF systems, and (4) combining both CBF and CF into a single unifying model that incorporates both item and user characteristics [2]. Several extensions have been proposed for enhancing hybrid systems with most of them focussing on contextualisation. These extensions include: (1) adding a knowledge-base that uses domain knowledge to make recommendations, e.g. "seafood is not vegetarian, and vegan is a stricter form of vegetarianism", hence do not recommend seafood to customers that identify as vegetarian or vegan [30], (2) adding a more comprehensive understanding of users and items [127], and (3) adding more dimensions to the RS, with the current dimensionality of the RS being two, i.e. *customer* and *item* [2].

Most RS research on contextualisation today relies on the RS being able to measure the preferences of customers using the customers' ratings of related or associated products. However, having customer ratings is not always possible nor practical for all applications. Hence, these models that are used for contextualisation may not always be relevant [2][108]. This is particularly true for applications in medical diagnosis, counter-terrorism and in physical-store retailing, e.g. grocery retailing, where customers, or patients, or potential criminals do not necessarily rate their preferences

for related items. Instead, preferences have to be ascertained from their previous transaction history. This becomes even more complex when such transaction history is spread across several databases and cannot be amalgamated for data privacy and anti-competitive reasons, e.g. a customer that shops at several different retailers, or a patient whose medical records are held by several different medical institutions [2][108].

Providing recommendations and enticing customers to act on recommendations within the UK grocery retail sector is complex as the market is highly competitive with different brands of the same product being very similar, in terms of utility, and consumers generally prefer their engrained shopping routines [141][154]. Thus, to entice a customer to act on a recommendation in the retail space, the retailer must establish if: (1) the customer prefers the item, (2) the customer obtains the item from its store or that of a competitor, (3) the customer's consumption pattern of the item, and (4) the appropriate price point to attract the customer and force them to break their routine and switch retailers [97][154][155]. The conclusions of Rhee and Bell, in [141], noted that customers are more likely to "switch", or in this case act upon a recommendation, if the store layouts are similar, and if the customer is a regular shopper, buying smaller baskets, as opposed to a periodic shopper doing their monthly shop. Whilst price is rapidly becoming a huge factor, it must be noted that customers are only likely to switch if the retailer has both lower current prices and that there is a strong indication that such prices will remain lower in the future [97][154][155]. Hence, it can be concluded that all else being equal, one-off vouchers on recommended products may not be effective in sustaining long term customer traction.

Summary remarks on Recommender Systems in the context of this study

Based on the above review, it can be clearly seen that RS is more of a downstream

data analytics process in that it leverages more fundamental data analytics methods including ARM, Clustering, Bayes Networks, and Markov Chains [2][108][142]. Indeed, this study may be seen as more fundamental than RS, and consequently will contribute to the body of knowledge in RS as it may likely serve as an input into broader RS processes. In this regard, this study produces a model, mt model, for not only targeting items, but also deciding on which customers are best to target for the selected item. Based on the current literature, this is not a current feature of RS [96][108][154]. Further, this study also proposes a novel algorithm that leverages the mt model for item targeting, clustering for customer targeting, and simulations for predicting the outcome of future interventions. Given that all of these steps are key features of hybrid RS, the algorithm and techniques proposed in this study may be used to both enhance current RS systems, and be included in future RS design.

2.2.3 Algorithms used in MBA

2.2.3.1 Algorithms used in ARM

The growth of ARM may be seen as the catalyst for widespread research on ARM-based algorithms, that were used as part of computer-based software programs to solve various real-life data mining challenges [24]. The Apriori algorithm introduced in [7] is still recognised as the benchmark for ARM-based data mining, as it is both efficient and robust [140][160]. However, it has one major drawback in that it is costly, from a time and computer-memory perspective, due to its breadth-first computational approach [68][76]. Consequently, there has been several attempts to enhance the efficiency of this algorithm most notably the Eclat and FP-Growth algorithms proposed by [187] and [78] respectively. The fundamental concept of these algorithms is the obedience of the downward closure property, also known as the monotonicity or Apriori principle [77]. Unlike the Apriori, the FP-Growth and Eclat algorithms perform a depth-first scan of the database to identify all frequent, 1-item sets, and

then uses this result and the downward closure property to generate larger frequent item sets [77]. In the Eclat algorithm, this is achieved by recording the transaction identities, tids, and support for all frequent 1-item sets, and then generating frequent 2-item sets through intersecting the frequent 1-item sets and comparing the support of the resulting set with *minsup*. Whilst the Eclat algorithm is faster than Apriori, as it scans the database only once, it is memory intensive as it initially requires a large part of the vertical database to fit into main memory [68]. This is particularly acute for large databases like those found in grocery retail. A further issue with the Eclat algorithm is that it compounds its memory demands by theoretically combining frequent 1-item sets to form larger, frequent itemsets, that may not necessarily exist in the actual database [68]. This shortcoming of a memory-intensive initial search was realised, and the dEclat algorithm was thus proposed as an update to the Eclat algorithm in [189].

The dEclat algorithm showed a significant improvement in memory usage when compared with the Eclat algorithm [189]. However on reflection, it was also shown to be more memory-intensive than the Eclat for sparse transaction databases with low minimum supports, which typically is the case grocery retail shopping transaction databases [24]. The dEclat algorithm is based on calculating the support of a set of items by considering the number of transactions in which this set is not present. This is represented by a diffset, where the diffset for item (X) is given by $d(X) = t - t_X$, where t is the total database and t_X is a cut of the total database with transactions that only contain (X) [189] . Given this, it can be clearly seen that if the database is sparse and minimum support is low, then $d(X) > t_X$, thus making the dEclat more memory-intensive than the Eclat algorithm.

Although the FP-Growth algorithm has been shown to be more memory-intensive

than the Eclat algorithm, it can be significantly faster because of the novel way in which it exploits the downward closure property [68]. The FP-Growth algorithm commences similar to the Eclat by scanning the database for all frequent 1-item sets but goes a step further and ranks the results in descending order of frequency [77]. The database is then scanned, transaction by transaction, to build-up a tree structure with items that are frequent in multiple transactions, frequent item antecedents, forming the branches and lower frequency consequents, forming the leaves. The support of the itemset, root to leaf, is then given by the support of the leaf, using the downward closure property, as the branch will always have at least the same support as the leaf. Database pruning is done concurrently and a leaf will only propagate further if it is frequent [78].

The several attempts to enhance the compactness of the Apriori, Eclat and FP-Growth algorithms have yielded good results, particularly those algorithms that use the downward closure property to prune computations that generate subsets of already existing larger sets [68]. However, several of these modifications are based on the depth-first search principle and requires the entire database to be scanned into main memory. The Genmax algorithm proposed in [71] enhances the speed and memory utilisation of frequent itemset mining by leveraging the Eclat algorithm to build an initial list of the largest frequent itemsets possible, commonly referred to as maximal frequent itemsets, MFIs, and thereafter prunes all subsequent searches if it results in itemsets that are subsets of the known MFIs or that are not frequent. The rationale behind Genmax is that once all MFIs are found then by definition all frequent itemsets have also been found, because any subset of an MFI is also frequent. Similarly, the CHARM algorithm proposed in [190] uses the closed itemset property to rapidly prune all subsets and co-occurring sets to reduce the mining exercise to only closed frequent itemsets which in practice is substantially less than the list of all frequent

itemsets [68][190]. Since, by definition, a closed itemset has no subsets with the same or lower support and no supersets with the same or greater support, all non-closed subsets can be pruned and replaced by the closed itemset. Further, if a closed itemset has support = *minsup*, then its is also an MFI and all supersets of this itemset can immediately be pruned as they are not frequent.

Another popular compression technique is the use of the non-derivable itemset, NDI, technique proposed in [31]. However, despite several enhancements to this technique, it was, in general, found to be ineffective in sparse datasets [117]. NDI mining is based on finding all itemsets whose support cannot be determined from its subsets, defined as non-derivable itemsets, and uses this to determine all frequent itemsets. By definition, all frequent 1-item itemsets are frequent NDIs, hence the initial steps are similar to Eclat, FP-Growth etcetera. Attempts were made to combine closed itemset mining with NDI mining to generate Closed NDIs [117]. Whilst this technique was found to reduce the number of NDIs required to generate all frequent itemsets, its performance in sparse databases was more complex but no better than other techniques [117].

The issues around sparse databases and high main memory requirements were to some extent addressed by the RElim (Recursive Elimination), and SaM (Split and Merge) algorithms proposed in [23]. However, comparison tests with the Eclat and FP-Growth algorithms showed that whilst both the SaM and RElim algorithms have simpler structures, SaM can be slower on sparse databases due to the additional calculations included in the "merge" step whilst RElim was slower on dense databases [23][24]. SaM and RElim are based on sorting itemsets and transactions with an ascending order of frequency using the prefix item. The frequency of the itemset is checked and if it is frequent, it is stored in a separate list. The prefix item is removed and the process recurs until all items have been removed. However, SaM and RElim default

to a divide and conquer method, similar to Eclat, and can be memory intensive [24].

2.2.3.2 Algorithms used in clustering

Clustering remains a fundamental process in data mining today and is widely used in several applications [33][94][119]. However, based on the discussion in Section 2.2.2.2, algorithms used for item or product clustering were not considered further in this study, hence the focus of this section is on algorithms for customer or transaction clustering. Several studies, including [70], [119], [162], and [175], noted that the K-Means (KM) and Fuzzy C-Means (FCM) clustering algorithms remain the most popular and widely used approaches today. Hence KM and FCM will be the focus of this section.

K-Means Clustering

K-Means clustering is well-documented in [160]. The underlying principle of KM is as follows: a given number of clusters, with initial values for the cluster centroids is defined. This is followed by all data points being assigned to its closest cluster, and an iterative process begins where the centroid is re-calculated after each data point assignment step. The process stops when there is no change in data point re-assignments [160][175].

While implementing a KM-based algorithm may be relatively straightforward, several previous works, including in [33], [140], [160], and [175] have noted that the method itself has three main drawbacks, which sometimes can be resolved:

1. Randomly choosing the initial centroids: choosing the initial centroids randomly generally produces poor results and this can be exacerbated by performing multiple runs with the same data [140][160]. Whilst there are several techniques to overcome this issue, the two best techniques were: incremental updates of centroids, where the centroid is updated after each additional point rather than

once all points are added, and K-Means bisection to produce the initial centroids, where the initial data set is divided into two clusters and then each cluster is bisected further [160].

2. Empty clusters: it is possible to have empty clusters as a result of the initial choice of centroids. This may unlikely resolve itself during the iteration process. One way of overcoming this is to manually remove that choice of centroid and select another centroid from that cluster which has the largest data spread, thus splitting this cluster and compacting the overall clustering process [160].

3. Outliers: outliers can influence the effectiveness of the clustering process and the typical approach to this problem has been to find outlying data points in advance and eliminate them [160]. However, outliers in some applications may have significance and care should be taken not to eliminate these [160]. Some everyday examples include: financial analysis where highly profitable customers or fraudulent transactions may show-up as outliers, or in the case of potential criminal activity where large purchases of an item, e.g. nails, screws etcetera, that is otherwise purchased in smaller quantities, may be considered an anomaly that should be eliminated, when in reality it is a signal of a major threat in progress.

Fuzzy C-Means Clustering

FCM, first introduced in [18], and based on the work in [44], was created to overcome some of the problems commonly associated with the crisp clustering approach of KM. These include those noted earlier, and the need for multiple passes to improve clustering accuracy [33][70][94]. Unlike KM, FCM uses a soft clustering approach in which data points on the boundaries are not forced into a single cluster but rather they are allowed to be members of multiple clusters with varying degrees of membership, such that the total membership of a data point across all clusters equals to one. This

approach, not only improves clustering accuracy, but also closely resembles everyday life [33][119]. There are several other fuzzy clustering algorithms that exist, but FCM remains popular as its relatively stable, reliable and fast [119]. However there are three main, well-known problems with FCM:

1. CPU usage as a result of speed: the speed benefits of FCM was noted to be computationally expensive, in particular for large data sets, and there have been several variations of FCM to improve on this over the years [119]. One approach taken in [32] to optimise CPU usage proved to be effective in reducing the CPU usage by a sixth. This was achieved by using a look-up table to determine an approximate value for the Euclidean distance calculations as opposed to computing the exact value.

2. Too many iterations as a result of sub-optimally selecting the fuzzifier "m" [119][162][174]. The generally used value of 2 for the fuzzifier "m" is not optimal for all applications and a sub-optimal "m" can be time consuming due to the increased number of iterations required to reach convergence. However a large number of applications use "m" = 2 and this will be used in this study as well [33][162].

3. Choosing too many initial clusters: Winkler et al. in [174], noted that the performance of FCM was weak when the number of initial clusters were high, at approximately 100. To combat this, a polynomial was introduced to FCM which was shown to reduce the impact of choosing a large number of initial clusters [174]. It should be note that for this study, the total number of clusters in not expected to exceed twelve, hence the issue of too many clusters is not going to pose a problem.

Based on the above, it was concluded that whilst FCM is not without its problems, its accuracy is superior to KM, and hence formed the basis for clustering in this study

[33].

2.3 Decision-Making models in MBA

Research into combining basic parameters of itemsets to compare the quality of association rules has typically been part of studies on interestingness measures [104]. The concept of interestingness, first introduced in [129], centred on the principle that statistically independent rules are generally not interesting as they obey the laws of probabilistic chance, i.e. if the observed value of $P(X, Y)$ is equal to the theoretical value of $P(X, Y)$, given by $P(X, Y) = P(X) \cdot P(Y)$. In this regard, such rules do not add to knowledge and may be disregarded [160]. However, rules where $P(X, Y) \neq P(X) \cdot P(Y)$, are of interest and merit further study [129]. Thus, it is on this basis that research into interestingness measures have gained popularity [160].

2.3.1 Interestingness Measures in MBA

The notion of combining variables to form interestingness measures has been well-studied over the years [10][85][129][161]. Research into interesting measures have generally been focussed on either creating new measures to determine the strength of associations, or map existing measures to specific real-life applications [168]. In this regard, thirty eight such measures were noted in [65] and [104], with twenty one being compared in [161], across a variety of real-life applications, to test their suitability. A similar exercise was conducted in [10].

One key reason for the plethora of measures is that the context plays a significant role in the performance of such measures and while one measure may be best suited to one context, it may to be totally unsuitable to another [10][168][161]. A common thread across most measures was that they largely centred on Piatetsky and Shaprio's

principles for rules interestingness which were outlined in Section 1.3 [65][104][161]. In light of this, and in the context of this study, Tan at al., in [161], found that the cosine measure was best suited to retail settings. However, in research conducted as part of this study using real-life grocery shopping data, Moodley et al. in [115], showed that the cosine measure performed poorly against their proposed uninorm measure. In this regard, tests conducted to assess the adherence to the monotonicity property, which forms part of Piatetsky and Shaprio's principles, showed that the cosine measure was generally poor, while the uninorm performed well in all scenarios [115].

It was surprising to note that while considerable research on interestingness measures has been focussed on comparing the strength of associations, there has been no evidence of studies that focussed on comparing rules to determine which rule will most likely result in the combination attaining frequency first. Some studies including in [168], focussed on proposing new measures to quantify the strength of associations while others focussed on combining measures to do the same [144]. It must be noted that the strength of associations alone do not guarantee frequency. Indeed some rare rules may be very strong, e.g. caviar and champagne, but these items will by no means be targeted to become frequent itemsets at most UK grocery retailers.

2.3.2 Identifying the best rules for targeted promotions

As noted earlier, there has been considerable research conducted on ARM algorithms, however the mathematical underpinning of ARM has remained fairly consistent with surprisingly little done on determining which rule will attain frequency first [140][175][189]. The task of leveraging ARM for practical retail applications has been widespread, with interest from both academia and retailers, who have seen the potential for gaining a competitive advantage [43][109]. A key consideration is the identification of the best items to target that fulfils retailers' objectives. In this

regard, FIM remains the typical starting point in identify items to target, and consequently up-sell and cross-sell [43][56][140][167]. However, shortlisting the set of all frequent items to find the best itemsets to promote can be tricky and remains context-dependent, particularly for large databases, like those typically found in the UK grocery retail [77][189]. The shortlisting of frequent itemsets to find the best itemsets to promote cannot be achieved by FIM algorithms alone, including Apriori, FP-Growth, Eclat, and this prompted a need to include further analytical elements to achieve this task [115][140]. In [115] the uninorm was shown to be effective in selecting the best rule for a given antecedent (e.g. $A \rightarrow C$ or $A \rightarrow D$) to aid shortlisting of frequent itemsets. However, the authors cited the more generic case of $(A \rightarrow C)$ or $(B \rightarrow D)$ as future work. Reutterer et al. in [140], detailed an approach that achieves shortlisting, however it is believed that this approach does have drawbacks with regards to the elimination or reduction of "false positives" and "false negatives".

In view of the above, it is clear that a gap exists in the current body of knowledge, in that whilst it is important to understand the strength of associations and compare rules in this manner, it is also important to understand which rules are likely to become frequent sooner. This extends beyond just theory and has real-life applications, like in grocery retail, where being able to determine which rule is likely to become frequent sooner can enhance sales and profitability of companies whilst minimising marketing spend. This gap in knowledge has been addressed as part of this study.

2.4 MBA and the Grocery Retail Sector

The impetus for MBA was originally driven by the grocery retail sector [5], and over twenty fives years on, despite its widespread use across a variety of fields, MBA remains grounded in grocery retail [140][165]. Given that grocery retail forms the

main thrust of this study, it was felt important to survey some of the main areas where MBA plays an important role within the UK grocery retail sector .

2.4.1 MBA in enhancing Customer Development

It is well-established that the battle for market share using price as the main weapon is ill-founded [134][154]. As a result, major UK grocery retailers are now heeding this advice and are instead focussing on CRM to win consumers and gain market share [72][134]. Having a deeper understanding of customers has been highlighted by several analysts as a key capability for retailers that want to standout amongst its peers [22][46][122]. The study conducted in [22] in the United States (U.S.) showed that a better understanding of customers is a key capability that grocery retailers should have, in order to better respond to the four forces that has changed grocery retail permanently. These four forces are: (1) value seeking, (2) technology enabled, (3) online encroaching, and (4) innovation [22]. Similarly, Big Data and Continuous Computing was identified as one of ten key trends that will shape the future of UK grocery retailing [46]. Further, the big challenge around data for retailers is not in the collection of data, but in the processing of large volumes of data, created by consumers and operational systems, to generate insights on consumer behaviour, particularly across the multiple retail channels, that has now become common place [46]. The conclusions in [122] highlighted similar shifts in retailing across the globe, noting that in order to win in the new retail environment, it is not only important for retailers to know their customers, by understanding their needs and wants, but by also to being in a position to measure the impact that they are making on the consumers, and adjust operations accordingly to drive maximum business value. Based on this, it can be concluded that the need for effective data analytics of customer behaviour, particularly MBA, is now more essential than ever before.

A well-researched article by The Register, in [138], highlighted several important considerations around data analytics. The first consideration was that whilst data collection and analysis has played an integral role in the market share ambitions of the major UK grocery retailers, through price setting based on the consumer dynamics for a given region or store, any differentiation in this regard has been rapidly neutralised, as all major UK grocery retailers now have this capability. This is in line with the conclusions in [134], that price should not be a key differentiator as it is inevitably a race to the bottom. In line with this, the market has now shifted to the second consideration as highlighted in [138], i.e. integrated big data analytics, which is currently being deployed by most major UK grocery retailers. The notion of integrated big data analytics is where UK grocery retailers are building consumer trust by having a holistic understanding of their needs and wants, and providing them with consistent, tailored offerings [138]. In this regard, UK grocery retailers are focussing on integrating multiple external data sources like data from third party providers, e.g. Kantar, Neilson etcetera, with their internal back-office operational data, e.g. purchasing, staffing, cost of operations, and customer sales activity across all sales channels to obtain a single view of the customer [43][60]. This will thus allow them to better tailor offers and promotions to each group or individual customers [138][43][60]. Having this single customer view, considered the "holy grail" of retailing, has been elusive for at least the last two decades largely because of technology limitations [154]. These have now been somewhat eliminated thanks to the advances in mobile technology and the proliferation of social media [138][154]. However, whilst technology has advanced, customers are no longer loyal to brands and UK grocery retailers like they have been in the past and consequently are spending their money across multiple grocery retailers and brands [86]. Thus, the single view of customers has been increasingly difficult to obtain because data is split across multiple systems, many of which are not accessible to individual retailers, for legal and proprietary reasons [86].

2.4.2 MBA in enhancing Product Development

In a study of Tesco's data analytics capabilities, Forbes, in [60], showed that predicting product sales behaviour is one key area where Tesco uses data analytics. This is achieved by focussing on products and how they are bought, as opposed to customers and how they shop. In this regard, Tesco used MBA, more specifically clustering, to ensure that product sales are predictable and "behave in the right way", so that it can be supplied in the right way, i.e. always be in stock, and not go to waste [60]. Similar studies have also been conducted by Waitrose, where MBA has been used to understand product placement and maintain optimum stock levels [109]. This approach, not only ensured that shelves were kept full, but also provided insight on the fluctuations in purchasing throughout the day, week or month. As a result, perishables were well-stocked, fresh, and waste is minimised [109]. The findings in [109] has also showed that MBA has moved-on from coupons at tills, with retailers now tailoring coupons to ensure that repeat business still continues to be effective. In the U.S., grocery retailer, Target, used MBA to identify when mothers were pregnant and tailor offerings accordingly. This resulted in a fifty percent increase in mother-baby product sales in the eight years to 2010, whilst Kroger used MBA to offer customers coupon-based discounts on products they have actually bought, as opposed to items that they would like consumers to buy [109]. As a result, Kroger recorded a fifty percent coupon-redemption rate amongst its regular customers, which was substantially higher than the industry average [109].

2.4.3 MBA in enhancing Supplier Cooperation

Given the success enjoyed by UK grocery retailers from MBA, it is not surprising that MBA-related data continues to be shared with its suppliers which has now yielded mutual benefits [8][109]. Three distinct areas have been identified where interventions as a result of MBA, has resulted in increased sales, and profitability for both retailer

and manufacturer alike. The three areas include: (1) perfecting in-store fill rates to ensure shelves are always stocked, (2) leaner, fresher stocks, using demand data to adjust manufacturing thereby producing enough stock to meet demand but not too much that ties up capital and/or goes to waste, and (3) cross-selling, where suppliers have worked with their retail clients to place their products across the store to enhance shopper convenience and boost sales of related products [8][109]. A good example of cross-selling is where global food manufacturer, Kraft, placed its salad dressing, pre-cooked bacon and grated cheese in the salad section of a grocery retailer, in addition to its traditional isle locations, to ensure that customers who typically bought these salad-related products as part of their baskets, also bought Kraft's products as opposed to that of its rivals [109]. In this case, the synergistic benefits for both supplier and grocery retailer clear. Customers sometimes need "memory jogs" or buy add-on products on impulse, and are hence more likely to either remember or include these add-on items in their baskets when they are co-located, as opposed to when they are in separate isles [141][109]. Further as a result of co-location, customers are also more likely to include the co-located brand of the add-on item, as opposed to another brand, located in a separate isle, thus enhancing brand traction for that specific supplier [109].

2.4.4 MBA in enhancing Targeted Marketing

Reutterer et al. in [140], noted that there has been an increase in research on the effectiveness of targeted promotions, particularly on its impact in increasing take-up compared to conventional promotions. Similar conclusions were made in [166] and [56]. In [166], Venkatesan and Farris, concluded that "customized coupon campaigns are more effective if they provide more discounts, are unexpected, and are positioned as specially selected for and customized to consumer preferences". In line with this, Fong et al. in [56], concluded that "targeted promotions based on individual purchase

histories are known to increase promotional response". Whilst targeted promotions are impactful, the issues of "false positives" and "false negatives" always remains relevant and should be avoided as much as possible [116][145]. In particular, "false positives" can be more damaging than "false negatives" as it angers customers and may result in a loss of trust of the retailer's ability to tailor offerings to customers' individual preferences [116][145].

Grouping customers for targeted marketing is very common, done by most major retailers and usually employs some form of clustering [12][109][157]. Clustering using K-Means (KM) was used to create target customer groups in [140]; however given the common challenges with KM, as noted in Section 2.2.3.2, it was decided to use FCM instead. The operational methods for targeting customers is beyond the scope of the present study, however it is worth noting that recent studies have demonstrated that the psychological targeting of customers has proved to be effective [110]. Whilst the typical method for targeted promotion has been "price", through some form of coupon redemption, it is possible, and may become increasingly used in the future, to target customers with products using their psychological preferences as the key criteria [110][140].

2.4.5 How this study supports these key UK Grocery Retailer focus areas

Given that this study focusses on understanding consumers' purchasing across multiple UK grocery retailers, and uses real-life shopping data that spans across multiple grocery retailers, it thus naturally supports the ambition of UK grocery retailers to develop a better understanding of customers' behaviours, and edges closer towards creating a single of view of the customer [90]. It also contributes towards enhancing product development and supplier collaboration as the core of this study depends on

generating associations between products. More importantly, because the data used spans across the UK grocery retail sector, it is likely to generate more associations between products than in a single grocery retailer's database alone. Consequently, this study enhances the opportunities for UK grocery retailers to develop more products and to increase collaboration with suppliers across a broader range of products, thereby enhancing business value.

2.5 Summary

This literature review highlighted the following key points which provided both a sound justification and a solid foundation for this study:

- MBA is a source of knowledge, with ARM and FIM being useful techniques in uncovering valuable patterns, and hence facilitating knowledge discovery.
- ARM remains the *de facto* method for MBA, particularly in grocery retail.
- Although FP-Growth and Eclat are faster, some of their implementations can be complex, and memory intensive especially for grocery retail. Hence Apriori is a good compromise choice for research-based MBA as it is less-memory intensive, albeit slower, while the quality of rules remains equally high.
- Decision-making between choices in MBA is surprisingly understudied and this study, through the development of the market target model and algorithm, makes a unique contribution to both MBA and KDD.
- ARM/ FIM alone is not be enough for targeted promotions in retail, indeed, a combination of ARM and Cluster Analysis is required to identify target items and target customers.
- Recommender Systems, although a current "hot topic", are not systems on their own. Instead, they rely on more fundamental processes, like ARM and Cluster

Analysis, to provide recommendations. In line of this, the market target model, proposed in this study, will contribute to the body of knowledge on RS.

- The grocery retail sector in the UK is undergoing major shifts and the demand and use of MBA related techniques, including the market target model, is now higher than ever before.

Chapter 3

Research Methodology

3.1 Introduction

The primary goal of this research was to provide a framework to assist stakeholders in predicting the outcomes of future decision making based on incomplete, historical data. Indeed, within the UK grocery retail context, and as noted in Chapter 2, having a single view of a customer is considered the "holy grail", and remains elusive, partly as a result of competition law and corporate ethics. Given this, being able to accurately predict customer decision making remains a challenge, and this study provided a novel framework for addressing this challenge by leveraging well-known research methods, data analytical techniques, and real-life data. The merits of the proposed framework are discussed in detail in the chapters that follow. It should be noted that this discussion on research methodology pertains to the UK grocery retail application that formed the main thrust of this research. Indeed, other applications have been studied as part of this research, and application-specific research methodologies are detailed in the respective sections in Chapter 6.

This chapter commences by providing an overview of the research philosophy that

was adopted as part of this study, which is followed by a detailed review of the research methods, data collection mechanisms, and simulation techniques that are typically used for data analytics research. The chapter then provides a justification for the choice of the approach that was taken for the various pieces of this study. The chapter also details approaches that were adopted for testing the effectiveness of key elements of the proposed framework, including theoretical tests, model comparison testing, and testing of the model with other datasets. Finally the chapter concludes with a summary of key points.

3.2 Overview of the Research Philosophy adopted as part of this Study

The philosophical approach taken by Saunders and Thornhill, in [147], was found to be most appropriate for this study given the highly commercial context of the UK grocery retail sector. As detailed in [147], the research paradigm that best suited this study was the "functionalist" paradigm as it compromised of both "objectivist" and "regulation" dimensions. From an "objectivist" perspective, this research sought to find practical solutions to practical problems under the basis that the UK grocery retailers are rational organisations whose management make rational decisions to advance their businesses. In this regard, the management of the UK grocery retailers adopt an "objectivist" approach, in that they view their organisations from a scientific perspective. Simultaneously, management of the UK grocery retailers are bound by strict internal corporate controls, in some cases by securities exchanges, like the London Stock Exchange, and by the laws of the country [131][138]. Hence all actions taken by management in both acquiring knowledge, and acting upon this knowledge are governed by the "regulation" dimension [147].

Table 3.1 provides a summary, based on the outline in [147], of the various research

philosophical positions that were adopted as part of this study, together with their relevant justification.

Philosophy	Stance Adopted	Justification
Ontology (Assumptions about the UK grocery retail sector)	**Objectivism** (Independence of management of the UK grocery retail sector)	UK grocery retailers are large, regulated organisations whose management act based on facts that are available to them. They are required to act independently of their personal stakes.
Epistemology (The acceptable knowledge base on the UK grocery retail sector)	**Positivism** (Using fact-based data to test the proposed model and hypothesis)	This research was conducted using a fact-based approach to test a mathematical model and algorithm. The data used was obtained from actual transactions within the UK grocery retail sector.
Axiology (The values of the researcher in this study)	**Value-Free**	The values of the researcher have not influenced the findings and conclusions of this research.

Table 3.1: Adopted Research Philosophical Positions based on [147]

3.3 Methods used in Research

This section provides a critical review of the typical research methods and experimental approaches that have been used in ARM, Cluster Analysis and CRM. The pros and cons of methods and approaches are discussed together with a justification of the approach that was taken in this study.

3.3.1 Experimental Process

The well-known Knowledge Discovery in Databases (KDD) process, see Figure 3.1, was first outlined in [51] and remains the primary methodology adopted in data mining research [160]. Tan, in [160], noted that at the heart of the KDD process is the

data mining phase which leverages models and algorithms to process data into information. In this regard, the Apriori and FCM algorithms form part of the data mining phase while the proposed market target model and target promotion simulations form part of the post processing phase, where patterns are interpreted to select the best information that contributes to overall knowledge [160].

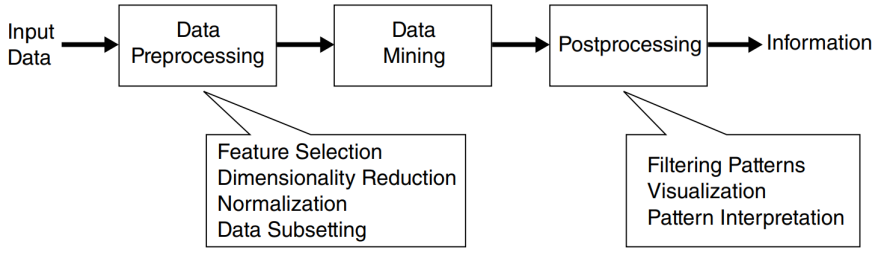


Figure 3.1: KDD Process as outlined in [160]

The system development framework, SDF, first proposed by Nunamaker et al. in [123], is also widely used today and remains one of foundational frameworks for systems development projects, which are now common place in data mining [128]. The framework, as shown in Figure 3.2, adopts a cyclic/ iterative approach between the three key pillars of research, namely: (1) theory building, (2) experimentation and observation, and (3) the formal development of a system, which in most cases is a technology/ computer-based.

From a grocery retail perspective, the classification framework for data mining in CRM proposed by Ngai et al. in [120], is a well-recognised model that is widely used in CRM research [9]. Further, the model is also in line with the theoretical foundations of retail as noted in several studies including in [101], [134], and [141]. A schematic representation of the framework is provided in Figure 3.3. As detailed in [120], CRM is at the heart of the process and is initiated by one, or a combination of four customer interaction stages, namely: (1) the customer identification stage, where the grocery retailer decides on which customers it wants to target, (2) customer

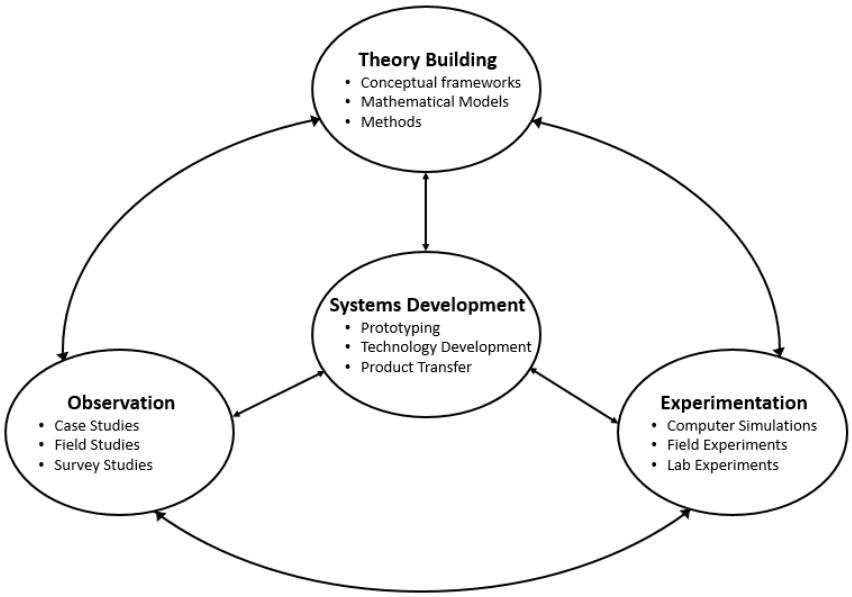


Figure 3.2: Systems Development Framework as outlined in [123]

attraction stage, deciding on how to attract these target customers, (3) customer retention stage, deciding on how to retain customers, and (4) customer development stage, deciding on how to grow customers by increasing their spending and loyalty. These four stages are then supported by the various data mining tools which are shown in the outer ring in Figure 3.3.

3.3.2 Methodology adopted in this study

All three frameworks detailed above have strong relevance for this study and have been amalgamated to form a coherent process that is widely used across this study. Indeed, in the KDD process, the data mining phase corresponds to development of the mathematical approach of this study and fits in with the theory building phase of the Systems Development Framework (SDF) shown in Figure 3.2. Similarly, the post processing phase in the KDD process corresponds to the development of the algorithm, the UK grocery retail sector case study, and the simulation of interventions. These all relate to the experimentation and observation phases detailed in the SDF.

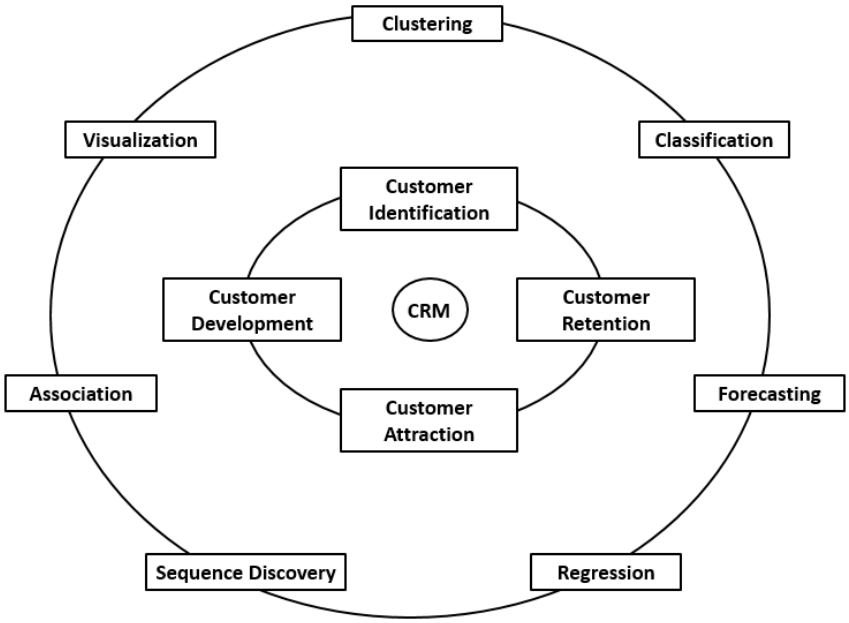


Figure 3.3: CRM process as outlined in [120]

The proposed algorithm has been used in the development of a computer-based software program which can be used in grocery retail, as well as in other applications, as shown in Chapter 6. The CRM process detailed in Figure 3.3 is seen as the contextual framework, as it provides a purpose for this study and the KDD and SDF models.

Whilst the CRM model is retail-driven, the process itself can be easily adapted for other domains, e.g. in medicine, customers can be replaced with patients and patient healthcare management can replace CRM. Similarly, customers can be replaced by citizens for government applications or criminals in crime prevention or pupils in educational initiatives.

3.3.3 Typical data sources used for MBA

At the granular level, its is good quality data that is always at the heart of any KDD process, and for MBA, both synthetic market basket data and sales data from actual retail transactions have been used since inception [6][160]. In the case of actual retail

transaction data, the data has been typically sourced from either a single retailer, or multiple retailers, or through the use of third party data collectors, who typically use consumer panels [5][103][183]. Whilst consumer panel data across multiple retailers is a very effective tool, in that it readily provides for the analysis of data from a large number of panellists transacting across multiple retailers, it does have drawbacks. Yang et al. in [183], found that the individual products between retailers were not "identical" in terms of their utility that they provided to the end user, hence comparisons were difficult. However, the authors used product categories rather than individual products which was sufficient to demonstrate their methodology [183]. This will not be an issue for this study as the data used has been categorised by Kantar (the data owner) to ensure product-equivalence across a variety of brands [90]. Further, product categories will also be used in this study to demonstrate the algorithm and methodology as opposed to individual products. For completeness, a detailed survey on the use of consumer scanner panels is provided in a subsequent section.

Research conducted on shopping data that has been synthetically generated has also produced successful results, notably in [6][23][183]. Synthetic data has typically been generated using a data generator algorithm, the most popular being the IBM synthetic data generator [183]. The IBM synthetic data generator uses a Poisson distribution to pick the size of transactions and includes a "corruption" factor to model the real-life situation, where not all frequent items are purchased together in the same transaction [139]. Variations of the IBM synthetic generator has been used over the years including the data generator created in [83]. Synthetic data has been used as part of this study, as well as the real-life scanner panel data available from Kantar [90]. In this regard, the generator proposed in [83] has been used as it was found to be very effective.

It should be noted that ARM studies are not limited to the retail sector only and

diverse data sources have been used, leveraging the underlying ARM principles, to demonstrate its applicability across multiple fields. Some of diverse applications and data sources include census, weather pattern, cancer research, and online news websites [23][52][160].

3.3.3.1 Consumer Scanner Panels in Grocery Retail

Consumer panels remain a key data collection method within the UK grocery retail sector as it provides a cross-sectional view including, market share, impact of promotions, and the impact of PESTEL (political, economic, social, technological, environmental and legal) factors [60][105]. In line with this, Leicester, in [105], noted that the consumer panels in the UK provide a good basis for a representative sample of the UK grocery consumer market, given that the eligibility criteria to become a panellist in the UK are relatively straightforward, and most households in the UK qualify.

The several advantages of consumer scanner panels were highlighted in [159] including: (1) offering deep insight into the longitudinal shopping patterns of consumer groups, (2) offering remarkably better insight than consumer surveys where the data is largely qualitative and lacks real-life context, (3) insight on the consumption habits of individual families, or their demographic, (4) insight on the impacts of marketing campaigns, and (5) insight on the impact of macro-economic changes, including the PESTEL factors.

Despite the above advantages, consumer scanner panels have some drawbacks that may be relevant for some data analysis applications. Both Leicester, in [105], and Swait et al. in [159], noted that consumer scanner panel data generally lacks the rationale behind purchases. This could be supplemented with consumer choice data, e.g. obtained from surveys, to add context to purchases, e.g. in the case of impulse

purchases, shopping for comfort, etcetera. Further, as the onus lies on the consumer to capture their purchases onto the system, some shoppers may not capture all of their purchases leading to an underestimation of goods purchased. To combat this, consumer scanner panel management companies have incentivised consumers with coupons and vouchers for increased spending, hence the more the consumer spends, and captures the data onto the system, the greater the rewards [159]. The final shortcoming for grocery retailers is the inability to perform scenario analysis on panel data, e.g. assessing the impact of a targeted promotion [159]. Hence, grocery retailers have to rely on supplementary panels, e.g. consumer test panels to provide input into the success of future campaigns [105][159].

Given the above, scanner panel data is well-suited to this study in that the data sets are substantially large, and represent real-life scenarios. This implies that the associations between items purchased are indeed a realistic reflection of what is occurring in the marketplace [105]. Further, this study will partly address the last shortcoming outlined above, in that it will enable some scenario analyses being performed on the panel data to gauge the impact of targeted promotions, e.g. performing varying "take-up" scenario modelling on a group of consumers whose baskets exclude a frequent item and are undergoing a targeted promotion campaign to entice them to buy this product.

3.3.4 MBA-related Simulation Techniques

Grocery shopping is a discrete activity, and thus naturally lends itself to discrete processing. In [92], it was noted that simulating discrete activities has typically been done using one of three approaches namely: (1) Markov models, (2) Discrete Event Simulation (DES), and (3) decision trees. In [77] and [160], it was noted that decision trees are typically good for simulating decision making between several options, where

each option has a probabilistic outcome. Given that the simulation requirements in this study was assessing the impact of a sustained marketing campaign, it was decided that decision trees, although can be made to be suitable for this application, was not the best approach.

In [54], DES is defined as simulating the effects on a system where one or more variables change at discrete points in time as opposed to continuously. Hence, in this study, simulating the impact of promotions, e.g. offering vouchers to targeted customers, fits this definition well, and may be considered as an application of DES. Karnon, in [92], noted that in general, Markov and DES models typically perform a similar function, but unlike Markov models, implementing DES models can be more complex as it requires more computer coding, more parameters or attributes, and longer time horizons. On the other hand, Markov models, provide a more simplistic view and relies on the state at the previous event to make predictions [92]. In line with this, Simpson et al. in [151], also noted that DES models are more complex than Markov models, and whilst both models typically provide similar simulations at a high level, DES models consistently outperforms Markov models on finer detail [151]. Whilst finer detail may not be so important for this study, and grocery retail in general, the finer detail is indeed extremely important in some applications including medical diagnosis and treatment, which has been the focus of the study in [151].

Given the need for a simple simulation approach as part of this study, it was decided that Markov models will be adequate for simulating the long term impacts of a sustained marketing campaign. Thus, Markov models are used in simulations as part of this study, with a detailed view of the approach provided in Section 4.4.

3.3.5 Theoretical testing of the proposed algorithm

The suitability of any proposed framework must be tested to ensure that it conforms to the primary objectives of the research, and indeed the theoretical underpinnings that govern the scope of this framework [147]. As discussed in [7], [77], and [115], models that target itemsets must obey the Apriori and probability principles. Further, from a business perspective, the model must prioritise the target item that will theoretically reach frequency first. These principles are described theoretically as follows, with the first three relating to Apriori, and the fourth to business principles, where $\mathcal{M}$ represents the output of the proposed model:

P1: $\mathcal{M}|_{(A,C)}$ preferred over $\mathcal{M}|_{(B,D)}$ $\quad \forall (A,C),(B,D);\ \text{supp}(A,C) > \text{supp}(B,D)$ and $\text{conf}(A,C) > \text{conf}(B,D)$

P2: $\mathcal{M}|_{(A,C)}$ preferred over $\mathcal{M}|_{(A,D)}$ $\quad \forall (A,C),(A,D);\ \text{supp}(A,C) > \text{supp}(A,D)$

P3: $\text{supp}(A) \geq \text{supp}(A,C) \quad \forall A, C$

P4: If $\text{supp}(A,C) > \text{supp}(B,D)$ and $\text{conf}(A,C) < \text{conf}(B,D) \quad \forall (A,C),(B,D)$, then the proposed model must be used to select the combination that will theoretically reach frequency first.

3.3.6 Comparative Testing of the proposed algorithm with other currently published algorithms

Comparative testing was conducted against a current published model to test its performance. Indeed, for the proposed algorithm to be unique, valuable, and additive to the body of knowledge in data mining, the proposed algorithm not only had to be novel, but comparatively better than what is currently available [147]. In this regard, the proposed algorithm was compared with the approach detailed in [140]. The two approaches were compared using four tests: (1) ability to target customers

who will buy the product and eliminate "false positives", (2) ability to target most, or all customers who want to buy the product and reduce "false negatives", (3) ability to offer customised treatment and avoid targeting loyal customers, as this drives down price, and (4) ability to enhance the frequency of the target itemset, by increasing the purchasing of the target itemset by the target customers. Further detail, together with the results and discussion of the performance testing of the algorithms are provided in Section 5.5.

3.3.7 Comparative Testing of the proposed algorithm with "reality"

The benefit of using third party consumer scanner panel data, like the 2012 and 2013 datasets from Kantar, was that it enabled a comparison of the proposed algorithm with what actually transpired. Consumer scanner panel data provides a holistic view of every panelist's, or customer's, purchasing behaviour across all stores. Thus, the assumption that customers who are classified as disloyal at one grocery retailer, because they purchase lower volumes of an item and/or in general spend less, are indeed purchasers of this item in other stores can be validated. A similar construct can be used for loyal customers. Being able to validate these assumptions are important for this study as it justifies the clustering mechanism and the treatment approach taken for each of the proposed clusters. Naturally, grocery retailers do not have this information about all of their customers, and consequently rely on intelligence from grocery retail analysts like Kantar, Dunhumby and Nielson, who obtain this information, from amongst other research, consumer scanner panels [43][90][122].

To quantify the effectiveness of the testing, the mean number of purchases of each group, for each itemset was computed for both the store in question, internal transactions, as well as at all others stores, external transactions. The percentage difference

of mean purchases given by Equation (3.1) was calculated for each group.

$$\% \text{ mean difference } = \frac{(\text{internal mean} - \text{external mean})}{\text{internal mean}} \times 100\% \qquad (3.1)$$

From Equation (3.1), it can be seen that in general, for the proposed algorithm to be effective, the % mean difference will be expected to decrease as customers move from a state of loyalty to disloyalty.

3.3.8 Tests with Other Datasets

The Apriori algorithm, which is used as part of this study, has long been criticized for its slowness, particularly on large, dense datasets [24][76]. Grocery transaction datasets are typically very large, but sparse. For example, the 2012 dataset used in this study has 2.6 million transactions, with 286 items, and an average density of 11 items per transaction. This is equivalent to a density of 4% (i.e. $11/286 \times 100\%$).

For completeness, tests were conducted on synthetic transaction data, created using the data generator described in [83], to compare processing times on datasets of differing densities. The synthetic datasets were processed using the same approach detailed in Section 4.5, with the output file from the frequent itemset mining for both the medium and large dataset processed further. In all cases, the computational approach, and the applicability of the proposed model was the same, that is, the best itemsets to target where those itemsets that had the best model output.

3.4 How does this Research Methodology facilitate the development of the Unique Contributions of this Study and adds to the existing Body of Knowledge?

The concept of knowledge in the data mining context, as discussed in Chapter 2 and detailed in [160], is the generation of patterns that benefit the user and task, with the value of such patterns being its interestingness. In this regard, the aim of this study was to adopt a research methodology that leveraged well-recognised research frameworks, data analytical techniques, and mathematical principles to generate patterns that are interesting, and thus produce knowledge. Further, given that the research paradigm is "functionalist", and that the researcher has taken a "value-free, observer of scientific fact" approach, the patterns and knowledge that result from this research approach are reproducible, scientifically coherent, and will be the same irrespective of who or what conducts the analysis [147].

Given this, the proposed research methodology has led to the development of the proposed targeted promotions algorithm, and this linkage is detailed throughout this study. The unique contributions of this study are also detailed in several sections throughout this study with the premise that: (1) it is unique, because it has not been done before, (2) it is effective, because it achieves the primary business objectives of the study, and (3) it is additive to knowledge, because it is comparatively better than other existing approaches. Based on this, it can be concluded that the proposed targeted promotions algorithm adds to the existing body of knowledge, and thus the proposed research methodology has been effective in supporting this objective.

3.5 Summary

The discussion on the development and justification of the research methodology adopted as part of this study highlighted several key points:

- The "functionalist" paradigm adopted as part of this study is well-supported in data mining research in that it is scientific and produces results that are reproducible and generally free from researcher bias.
- Amalgamating research methodologies to incorporate the development of data mining theory, computer-based algorithms, and the business context, is a an effective way to develop data mining solutions that address business challenges and contribute to knowledge.
- Testing the proposed algorithm against a variety of conditions is an effective way to demonstrate the uniqueness of the research, the contribution of the research to the existing body of knowledge, and to underscore the validity of the research approach that was adopted.

Chapter 4

Mathematical Model and Algorithm

This chapter details the mathematical model and computer-based algorithm that formed the foundation of this study. It commences by mathematically formalising several definitions that have been used throughout this study, followed by the development of the mathematical model, which is based on two parts namely: identifying target items/itemsets, and identifying target customers. The marketing simulation approach is then discussed, before the chapter concludes with an outline of the computer-based algorithm, and a summary of key points.

The unique contributions of this chapter, developed using the research methodology detailed in Chapter 3, are as follows:

- The creation of a novel, yet simple, mathematical model to support generalised decision making, that can not only significantly improve MBA, but can be applied to other fields that have similar decision-making constructs.
- The contribution to the overall body of knowledge on MBA, in particular customer and itemset targeting.

- The creation of a new algorithmic approach to improve targeted promotions within the retail sector, and which can be applied to other fields that have similar decision-making constructs.

4.1 Definitions

The following definitions are used throughout this study:

Items: Items are defined as per the original definition given in [5]. Let $I = \{I_1, I_2, \ldots, I_m\}$, be a set of all items, with the assumption that quality and quantity of items ($I = 1, \ldots, m$) remain constant across all stores, and customers do not stockpile. These assumptions were necessary to ensure consistency of items across all stores.

Customers: Customers represent households, with all members within the household viewed as one customer. This approach is consistent with practice, in that loyalty programs and retail analysts like Kantar, typically consider all members of a single household as one customer [42][90]. Customers are denoted by U, and represent a household with size $f; f > 0$, with U purchasing subsets of I, referred to as transactions, or baskets.

Transactions: All purchases are made in the form of transactions, or baskets, and contain a subset of I, for example $T_S = I_2,\ I_9,\ I_{11},\ \ldots,\ I_x$ is a single transaction from store S. Itemsets are defined as subsets of all items within a transaction. Customers make one transaction per time period, W, per store, and this study uses $W = 1$ week. This assumption takes into consideration the practical aspects of shopping, where the generally accepted length of a shopping period is one week, as it is in line with how most people plan their household activity [47][90][141].

4.1.1 Support and Confidence

Both support and confidence are central to MBA, and the standard definitions of support and confidence as outlined in [5], were used throughout this study, and is detailed in Equations (4.1) and (4.2).

$$\text{support of item, } \mathrm{I_i}|_S = \text{supp}(I_i)|_S = \left.\frac{\text{Number of transactions containing } \mathrm{I_i}}{\text{Total number of transactions}}\right|_S \tag{4.1}$$

$$\begin{aligned}\text{confidence of item } \mathrm{I_i} \text{ leading to items } \mathrm{I_iI_j}|_S &= \text{conf}(I_i \rightarrow I_iI_j) \\ &= \left.\frac{\text{Number of transactions containing } \mathrm{I_i} \text{ and } \mathrm{I_j}}{\text{Number of transactions containing } \mathrm{I_i}}\right|_S\end{aligned} \tag{4.2}$$

In addition to the above, two user-defined parameters are defined as follows:

- Minimum support, *minsup*, is defined as the minimum support required for an item or itemset to be frequent. Like $\text{supp}(I_i)$, *minsup* is a probability-based parameter, and hence $0 < \ minsup \ \leq 1$. Note that $minsup \ \neq 0$ is a practical constraint as it is nonsensical to speak of transactions or databases with zero items.

- Minimum confidence, $minconf$, is defined as the minimum confidence required for two or more items to be associated. Like with *minsup*, the constraints, $0 < minconf \ \leq 1$, and $minconf \ \neq 0$ are practical constraints, as the underlying principle of ARM is to look at items that are associated, that is where $\text{conf}(I_i \rightarrow I_iI_j) > 0$.

4.1.2 The Apriori Principle

The Apriori principle, first detailed in [7], remains fundamental to the study of MBA, and is a direct result of probability theory. The Apriori principle is defined as follows: For a given set of transactions, $\text{supp}(I_i) \geq \text{supp}(I_i, I_j)$ where I_i and I_j are two items contained within these transactions. From a probability theory perspective, Apriori may be written as $P(A) \geq P(A \cap B)$. In practical, retail terms it may be seen as: the transactions that contain *butter* are always greater than or equal to the transactions that contain both *butter* and *cheese*. Given that support and probability are interchangeable, this study uses $\text{supp}(I_i)$ and $P(I_i)$ interchangeably to denote support.

4.2 The Market Target Model for Identifying Target Itemsets

Consider an itemset, $J_1 = I_{11}, I_{12}, I_{13}, ..., I_{1x}$, that is not frequent in store S, over a time period t, but its subsets $(J_1 - I_{1x})$ and I_{1x} are frequent over the same time period, with $\text{conf}((J_1 - I_{1x}) \rightarrow J_1) \geq \ minconf$. A practical example of this could be that the itemset {*cheese, butter, milk*} is not frequent in S but its subsets, {*butter, milk*} and {*cheese*}, are frequent. Indeed, there may be several such itemsets that fit this criteria and given that stores have limited budgets for marketing, one key objective of their marketing departments will be to find the best J, or several such Js in the case of large stores, that should be targeted to minimise marketing spend, and maximum customer uptake.

Intuitively, the best target should be that which has the largest $\text{supp}(J)$ as well as the largest $\text{conf}((J - I_x) \rightarrow J)$ in S. These combinations are considered obvious, and not interesting by data scientists [160]. However, cases do exists where

$\text{supp}(J_c) > \text{supp}(J_k)$ but $\text{conf}((J_c - I_{cx}) \rightarrow J_c) < \text{conf}((J_k - I_{kx}) \rightarrow J_k)$, and the choice between J_k and J_c is not obvious. This type of scenario is not unique to retail, often requires further analysis before a decision can be made, and is considered very interesting from a KDD perspective [160].

Despite its regularity in decision-making, it was surprising to note that a generalised model for decision-making involving two independent alternatives; of the form: $(A \rightarrow C)$ or $(B \rightarrow D)$, was not readily available or well-publicised [160][167]. It is on this basis that the generalised model (market target model) was developed, to both speed up, and provide clarity to the decision making process. It should be noted that the model has been developed for the more generalised case of itemsets. However, an item is an itemset with cardinality equal to 1, hence the model applies to single items as well.

4.2.1 Using the Uninorm as a measure for decision making

The development of the generalised decision making model commenced with exploring the use of the uninorm as a starting point. The uninorm has been widely used for multi-criteria decision making, including in [177], and it was against this backdrop that it was felt appropriate to consider applying the uninorm to MBA decision making. A uninorm-like equation, Equation (4.3), was set-up incorporating the two key parameters within ARM, namely: support and confidence.

$$U(A,C) = \frac{P(A,C) \cdot \text{conf}(A \rightarrow C) \cdot (1-e)}{P(A,C) \cdot \text{conf}(A \rightarrow C) \cdot (1-e) + (1 - P(A,C)) \cdot (1 - \text{conf}(A \rightarrow C)) \cdot e} \tag{4.3}$$

where $U(A,C)$ is the uninorm with $0 \leq U(A,C) \leq 1$ and e is the user-defined neutral element, with $0 \leq e \leq 1$. Since $U(A,C) = 1$ when $P(A,C) = 1$ and $\text{conf}(A \rightarrow C) = 1$, choices with higher uninorms are preferred over choices with lower uninorms.

4.2.1.1 Testing the suitability of the Uninorm as a measure for decision making

The suitability of Equation (4.3) was tested using the principles outlined in Section 3.3.5 for generalised decision making between two independent alternatives. These tests were based on the criteria set out in [129] and [161] for model suitability and addresses the core requirements that underpin MBA; whilst ensuring that the proposed model meets the business objectives and practical constraints of commerce as outlined in Chapter 2.

Tests showed that the uninorm obeyed the Apriori condition as it was strictly monotic with respect to support, and was ideally suited to $(A \rightarrow C)$ or $(A \rightarrow D)$. This is the type (1) scenario outlined in Section 1.3 and principle P2 outlined Section 3.3.5. Indeed, the results of this test formed the basis of the study in [115], where the uninorm was shown to outperform all other popular measures, including the Jaccard and cosine coefficients, in decision making involving a fixed antecedent. However, tests conducted to determine the impact of "High Support or High Confidence Dominance", or principle P4 outlined Section 3.3.5, on the uninorm measure proved unsuccessful, in that the uninorm failed to adequately compensate for "High Support" or "High Confidence" dominance, and in some cases, alternatives with very high confidence were prioritised, as they had higher uninorm values, over more viable alternatives. Consider an example where there are 100 transactions, $minsup = 0.3$, $P(A, C) = 0.29$ and $\text{conf}(A \rightarrow C) = 0.05$, while $P(B, D) = 0.05$ and $\text{conf}(B \rightarrow D) = 0.8$. The value of $e = 0.3$ was selected to correspond with *minsup*. It is clear that the sales volume of (B, D) will have to increase by 25 units for (B, D) to become frequent given that *minsup* is equivalent to 30 units. This implies that 32 customers will have to be targeted for (B, D) to become frequent, while the required sales increase for (A, C) is 1 unit, implying that 20 customers will have to be targeted. Note that the implied number of customers to be targeted is the quotient of the required number of units

divided by the respective confidence. On calculating the uninorms, it was noted that $U(B, D) = 0.329$ while $U(A, C) = 0.0478$. This incorrectly suggests that (B, D) is a better target than (A, C) even though first principles calculations show that targeting (A, C) will require less effort than targeting (B, D), i.e. 20 customers to be targeted instead of 32. Based on this, it is clear that the uninorm approach fails principle P4 outlined Section 3.3.5.

4.2.1.2 Conclusions from modelling with Uninorms for decision making

Mathematical modelling with the uninorm showed that combining the support and confidence using mathematical operators and empirical parameters was unlikely to produce a simple and effective model to support generalised decision making as it failed to meet all the principles outlined in Section 3.3.5. However, modelling with the uninorm did shed light on the challenge at hand, and supported the need for a holistic approach, in which the "end-point", "start-point" and "speed" had to be considered simultaneously. In this regard, the "end-point" can be viewed as *minsup*, the point where an itemset is frequent. Similarly the "start-point" and "speed" may be viewed as the support, e.g. $P(A, C)$, and confidence, e.g. $\text{conf}(A \rightarrow C)$, respectively.

4.2.2 The Market Target Model for Generalised (A → C) and (B → D) Decision Making

Let items A, B, C and D be frequent items, or itemsets, in store S. Hence, $P(A), P(B)$, $P(C), P(D) \geq minsup$. Further, let $P(A, C)$ and $P(B, D) < minsup$, with (A, B) and (C, D) combinations being ignored for now. Four confidence equations for the above frequent items are created:

$$\text{Confidence of } (A \rightarrow C) \text{ in S} = \text{conf}(A \rightarrow \text{C}) = \frac{P(A, C)}{P(A)} \tag{4.4}$$

$$\text{Confidence of } (C \rightarrow A) \text{ in S} = \text{conf}(C \rightarrow \text{A}) = \frac{P(A,C)}{P(C)} \tag{4.5}$$

$$\text{Confidence of } (B \rightarrow D) \text{ in S} = \text{conf}(B \rightarrow \text{D}) = \frac{P(B,D)}{P(B)} \tag{4.6}$$

$$\text{Confidence of } (D \rightarrow B) \text{ in S} = \text{conf}(D \rightarrow \text{B}) = \frac{P(B,D)}{P(D)} \tag{4.7}$$

The number of physical transactions required for an itemset to be frequent is given by the number of physical transactions that correspond to minimum support. $P(A,C)$ is always initially less than $minsup$, hence the number of physical transactions required to make (A,C) frequent is given by:

$$\text{Transactions to make (A,C) frequent} = (minsup - P(A,C)) \cdot \text{total transactions} \tag{4.8}$$

The probability of a customer having (A,C) in their basket after initially purchasing (A) is predicated by purchasing (C) and is by given by the conditional probability equation:

$$P(A,C) = P(C|A) \cdot P(A) \tag{4.9}$$

Hence the probability of purchasing (C) after purchasing (A) is given by:

$$P(C|A) = \frac{P(A,C)}{P(A)} \tag{4.10}$$

From equation (4.10), it is possible that not everyone that purchases (A) will go on to purchase (C). Further, the right side of equation (4.10) is the same as the right side of equation (4.4), implying that $P(C|A) = \text{conf}(A \rightarrow C)$. Hence to achieve the required number of transactions to make (A,C) frequent will require a higher number of transactions, that contain the antecedent, to be targeted. In this study, this number

of transactions is referred to as the market target. Hence:

$$\text{Transactions to make (A,C) frequent} = \text{market target} \cdot \frac{P(A,C)}{P(A)} \tag{4.11}$$

Re-arranging equation (4.11) and by combining it with equations (4.8) and (4.4), equation (4.12) results:

$$\text{market target} \cdot \text{conf}(A \rightarrow C) = (minsup - P(A,C)) \cdot \text{total transactions} \tag{4.12}$$

One objective of a marketing department within a store is to minimise its marketing spend whilst increasing the sales of a product, or item, to make a product combination, or itemset, frequent. This may be achieved by minimising the number of transactions targeted, or market target. Equation (4.12) may be re-arranged to form equation (4.13). Further, it is noted that the market target is minimised when $P(A,C)$ and/or $P(A)$ is close to *minsup*, because $P(A) \geq minsup$ whilst $P(A,C) < minsup$ and $P(A,C) \neq 0$.

$$\text{market target} = P(A) \cdot \left(\frac{\text{minsup}}{P(A,C)} - 1 \right) \cdot \text{total transactions} \tag{4.13}$$

Equation (4.13) may be very useful in deciding on combinations within a store where the minimum support is the same, but could this notion be extended to compare combinations across stores or across minimum support thresholds? The significance of this "normalised" market target is that it will allow for combinations that are at different scales, be it frequency, monetary value, or shop floor space, to be compared. For example: is it better to target selling *cheese* to people that buy *butter* where $minsup = 0.2$ or target selling *nappies* to people that buy *beer* where $minsup = 0.3$? This question is best answered by a normalised market target. The market target described in equation (4.13) is an absolute value, however this is normalised, where

mt is the normalised market target, by expressing it as a fraction of the physical transaction value for minimum support, referred to as min support.

$$\frac{\text{market target}}{\text{min support}} = P(A) \cdot \left(\frac{\text{minsup}}{P(A,C)} - 1 \right) \cdot \frac{\text{total transactions}}{\text{min support}} \tag{4.14}$$

Equation (4.14) may be rearranged as follows, defining mt = market target/min support and noting that $minsup$ = min support/total transactions):

$$\text{mt} = \frac{P(A)}{P(A,C)} - \frac{P(A)}{\text{minsup}} \tag{4.15}$$

Note that $P(A,C) < minsup$, hence mt ≥ 0. Hence, the combination which minimises mt in equation (4.15), is always the best combination to target. The absolute value, the number of physical transactions, of the market target however is given by equation (4.13).

The value of $minsup$ may be adjusted to any value such that $P(A) \geq minsup$ because $minsup$ sets the threshold for frequency and item A was assumed to be a frequent item in our reasoning above. Further, $P(A,C)$ was assumed to be infrequent, hence $P(A,C)$ is initially lower than $minsup$ and $P(A)$. For some applications it becomes necessary to target a market such that $P(A,C)$ must become equal to $P(A)$, hence $minsup$ is set to $P(A)$, while $P(A,C)$ is initially lower than both $minsup$ and $P(A)$. Note that in these applications, $P(A,C)$ increases as the treatment of the target market progresses, while the mt, as described in equation (4.15), decreases. Eventually $P(A,C) = P(A)$ at which point treatment stops, as mt = 0. For example, in public health, authorities may want all people with compromised immune systems, A, to be vaccinated with some drug, C, eventually leading to $P(A) = P(A,C)$. In the UK, this is quite common in September each year, where the elderly or very young are encouraged to have the influenza vaccine to prevent illness during winter, which could

result in death [121]. Similarly in public safety, authorities may want all dangerous criminals, A, to be tracked using a wearable tracking device, C, eventually leading to $P(A) = P(A, C)$.

4.2.2.1 Testing the suitability of the mt model for decision making

The principles outlined Section 3.3.5 were used to evaluate the suitability of the mt model.

1. **Apriori:** It should be noted that $P(A, C) < minsup < P(A)$. Hence, in Equation (4.15), as $P(A, C) \rightarrow minsup$, mt $\rightarrow 0$ for a fixed $P(A)$. Consequently, mt monotonically decreases as $P(A, C)$ increases, and thus the mt model satisfies the Apriori condition or principle P2 outlined Section 3.3.5.

2. **High Support or Confidence Dominance:** It should be noted that $P(A, C) < minsup < P(A)$. Hence, in Equation (4.15), as $P(A, C) \rightarrow P(A)$, as is the case for very high confidence, then $minsup \rightarrow P(A)$ and mt $\rightarrow 0$ for a fixed $P(A)$. Given that *minsup* is always "sandwiched" between $P(A, C)$ and $P(A)$, mt is always "measuring the distance" between *minsup* and $P(A, C)$ and consequently "prevents" confidence dominance effects. If conf$(A \rightarrow C)$ is high, then the $P(A, C) < minsup < P(A)$ "sandwich" is compressed and consequently support is high, that is $P(A, C)$ is close to *minsup* and hence mt is low. If conf$(A \rightarrow C)$ is low, then the $P(A, C) < minsup < P(A)$ "sandwich" is expanded in either one of three ways and the output from the mt model is used to finalise the decision-making process, based on the "distance away from *minsup*":

   i. $P(A, C) << minsup < P(A)$: Likely that $P(A, C)$ is a poor target and other targets closer to *minsup* will be better.

   ii. $P(A, C) < minsup < P(A)$: Equidistant, and likely to be a poor target with other targets closer to *minsup* being better.

iii. $P(A, C) < minsup << P(A)$: Likely to be very close to *minsup* and may be a preferred target, with the output from the mt model being used to finalise the decision.

As a result, the mt model also obeys principles P1, P3 and P4 in addition to P2, as outlined Section 3.3.5.

4.2.2.2 Is targeting customers that buy item C with offers for A better than targeting customers that buy item A with offers for C ?

Making (A, C) frequent can be addressed by targeting customers that buy A with offers for C or vice versa. However which is easier? Whilst this could be easily answered using the mt equation, Equation (4.15), it is possible to make this decision by mere inspection of the values for $P(A)$ and $P(C)$. The antecedent should always be that item which has the lower support. Hence, target those customers who buy C with offers for A, if $P(C) < P(A)$. For completeness, a lemma is detailed below to prove this assertion.

Lemma 1: *If an itemset (A, C) exists but is not frequent in store S whilst both A and C are frequent with $P(A) = P(C)$ then both A and C are equally attractive products for marketing to target, in order to make (A, C) frequent.*

Proof:

$$\text{conf}(A \rightarrow C) = \frac{P(A, C)}{P(A)} \tag{4.16}$$

$$\text{conf}(C \rightarrow A) = \frac{P(A, C)}{P(C)} \tag{4.17}$$

Re-arranging the above:

$$\frac{P(A)}{P(C)} = \frac{\text{conf}(C \rightarrow A)}{\text{conf}(A \rightarrow C)} = 1 \tag{4.18}$$

Hence if both the support and confidence are equal, targets become equally attractive for marketing. ■

The extension of Lemma 1 may be stated as follows: If $P(A) > P(C)$, then conf$(C \rightarrow A) >$ conf$(A \rightarrow C)$, thus it is more attractive to target customers that buy item C with offers for item A as it has a smaller customer base with a higher probability of take-up, i.e. a higher confidence.

4.2.2.3 Conclusions from mathematical modelling with the mt model for decision making

Mathematical modelling with the mt model has showed it to be robust. It conforms to both the Apriori principle, and is not influenced by high support or high confidence dominance. Given this, the mt model was selected as the model of choice for item targeting, in that it is best able to support effective decision-making between two alternatives of the form $(A \rightarrow C)$ and $(B \rightarrow D)$.

4.3 Identifying target customers

The well-known RFM (Recency, Frequency and Monetary) framework for customer targeting in a retail setting was used as the starting point [50]. "Recency" is considered to be an elimination variable with the assumption that customers must have made at least two purchases during the period under consideration, with at least one containing the target item, else they are eliminated. Indeed, it is difficult to target customers who have not bought the target item as this introduces several variables that are difficult to measure, including taste, allergy/intolerance, and cultural aversion, and

may result in increased "false positives" as detailed in [145][154]. "Frequency" is item specific and is based on the support, supp(C), that the customer has for the target item C. "Monetary" is defined as the average customer transaction size for store S divided by household size, $\overline{|T_{Us}|}/f_U$, where $\overline{|T_{Us}|}$ is the average transaction size for customer U, in store S, and f_U is the family size. Given this, customers can be clustered into one of nine "buckets" as shown in Table 4.1 based on "Frequency" and "Monetary", to be considered as two linguistic variables of granularity three (Low, Medium, High). "Low","Medium" and "High" are assigned to both "Frequency" and "Monetary" based on the FCM clustering approach, which generates three clusters for each linguistic variable. Customers are considered to be "Switchers" if their frequency is less than their monetary spend, except for cluster one where both frequency and monetary spend is "Low". Customers are considered "Loyal" when their frequency and monetary spend are similar and at least "Medium". Customers that have a "High" or "Medium" frequency with a Low monetary spend are potential "Drop Outs" as they appear to spend elsewhere, and could be enticed to take their custom for the target product to another store. It should be noted that customers in Cluster 6 are considered to be "Switcher" because their frequency is less than their monetary spend, implying that these customers choose to purchase the target item elsewhere, even though they are essentially loyal to the store for other purchases. Consequently, the right incentive may enable customers in Cluster 6 to switch purchases away from other stores to this store.

Uplift theory, first discussed in [137], can be used to identify the "treatment" for each cluster: the treatment varies from "Switcher" to "avoid drop-out" and "leave alone / light touch" for "Loyal".

Switcher

Customers in this category either conduct a large proportion of their shopping at other

Cluster	Frequency	Monetary	Approach
1	Low	Low	Switcher
2	Low	Medium	Switcher
3	Low	High	Switcher
4	Medium	Low	Drop Out
5	Medium	Medium	Loyal
6	Medium	High	Switcher
7	High	Low	Drop Out
8	High	Medium	Loyal
9	High	High	Loyal

Table 4.1: Target Clusters

stores, and/or have a low take-up of the target item despite having high transaction volumes at the chosen store. An aggressive marketing campaign may be the suggested treatment to "break the habit" and force the customer to switch stores [141]. From Table 4.1, clusters that are identified as "Switcher" fall into this category.

Avoid Drop Out

Customers in this category have a high affinity for the target product from the chosen store, but conduct a large proportion of their shopping at others stores. Given this, they can be easily enticed to try an alternative from another store and "drop out". A gentle marketing campaign may be enough to ward off any pricing comparisons. Clusters identified in Table 4.1 that belong to this treatment approach are in the "Drop Out" clusters.

Leave Alone / "Light Touch"

Customers that are highly loyal to both the store and target item fall into this category. These customers may respond to marketing initiatives to increase their take-up but are usually "set in their ways", and could stockpile to save money which will lead to downstream decline in purchases, or ignore the initiatives completely. In any event promotions to this group may drive down prices unnecessarily as these customers

are unlikely to defect to other stores and would in essence result in them purchasing the same items at a lower price at the same store. Consequently, a "Light Touch" approach is recommended for clusters 5 and 8 where uplifts may be possible due to the medium valuation in one, or both variables "Frequency" and "Monetary". Customers in this category are identified in Table 4.1 as "Loyal".

4.3.1 Creating Clusters for Treatment

Nine clusters were created from Table 4.1, using a two-step fuzzy clustering process to enable targeted treatment. The fuzzy clustering process used the Fuzzy C-means (FCM) iterative algorithm proposed in [44] and modified in [18] is as follows:

Step 1: For each target item A_i in store S

- Cluster $\frac{\overline{|T_U|}}{f_U}$ into 3 clusters, by minimising the objective function F_A
- The objective function, $F_A = \sum_{i=1}^{u}\sum_{j=1}^{3}\mu_{ij}^m d(\frac{\overline{|T_{U_i}|}}{f_{U_i}}, c_j)$ where μ_{ij} is the degree of membership of $\frac{\overline{|T_{U_i}|}}{f_{U_i}}$ and the cluster, c_j, with $1 < m < \infty$ and $d(\frac{\overline{|T_{U_i}|}}{f_{U_i}}, c_j)$ is the Euclidean distance between the object, $\frac{\overline{|T_{U_i}|}}{f_{U_i}}$ and the centre of the cluster, $c_j = \frac{\sum_{i=1}^{u}\mu_{ij}^m \frac{\overline{|T_{U_i}|}}{f_{U_i}}}{\sum_{i=1}^{u}\mu_{ij}^m}$ and u is the total number of users being clustered
- The typical value for $m = 2$ was used as noted in [18]

Step 2: For each target item A_i in store S and Cluster, c_j created in Step 1

- Cluster the elements into 3 further clusters, based on the value of $\text{supp}(A_i)$, by minimising the objective function E_A
- The objective function $E_A = \sum_{i=1}^{n}\sum_{j=1}^{3}\mu_{ij}^m d(\text{supp}(A_i), c_j)$ where μ_{ij} is the degree of membership of $\text{supp}(A_i)$, and the cluster, c_j; n is the cardinality of each $\frac{\overline{|T_{U_i}|}}{f_{u_i}}$

cluster , with $1 < m < \infty$ and $d(\text{supp}(A_i), c_j)$ is the Euclidean distance between the object, supp(A) and the centre of the cluster, $c_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m}.\text{supp}(A_i)}{\sum_{i=1}^{n} \mu_{ij}^{m}}$

- The typical value for $m = 2$ was used as noted in [18]

4.4 Simulating marketing initiatives

Markov chains, as discussed in [136], were used to simulate the time series impacts of marketing. Given a set of transactions at the start of time period, W_t, where items A and C have been purchased frequently in store S, then the four possible "treatment" scenarios that exist for the purchasing of item C, by customers that purchase A, to make (A, C) frequent, as outlined in Table 4.1, are: (1) buying C frequently with large basket sizes ("Leave Alone"), (2) buying C frequently but basket size is medium ("Light Touch"), (3) buying C frequently but basket size is small ("Avoid Drop Out"), and (4) buying C at rate lower than the basket ("Switcher").

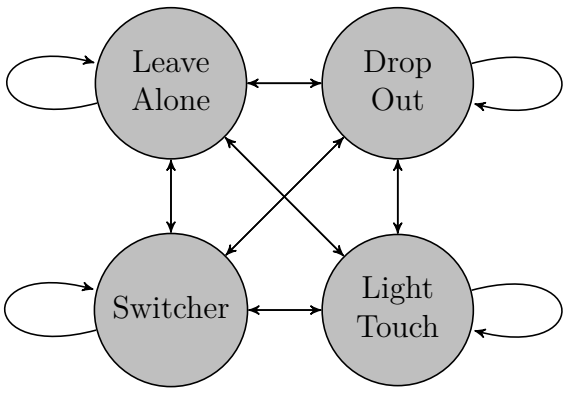


Figure 4.1: Markov Model for Simulations

This process was modelled using a Markov chain, see Equation (4.19), where P_{W_t} is the 1×4 proportion vector of customers buying A in one of the four states with regards to C. The process is ergodic, as shown in Figure 4.1, where customers can move from

one state to another based on their choice. Hence, the proportion vector at time W_{t+1} given by $P_{W_{t+1}}$ could be estimated using the Markov equation described in Equation (4.19), where M_{ACS} is the 4×4 Markov transition matrix that describes the probability of a customer buying A in store S, and moving from one state to another with regards to C over a time period. The values of entries in M_{ACS} are "user-defined" and can be adjusted to simulate a variety of scenarios, including conservative or aggressive marketing campaigns.

$$P_{W_{t+1}} = P_{W_t} \cdot M_{ACS} \tag{4.19}$$

4.5 Algorithm for enhancing the purchasing of target itemset (A, C) amongst frequent, known customers U in store S

The three models developed in Sections 4.2, 4.3 and 4.4 were combined together to form a coherent algorithm to enhance the purchasing of target itemsets. The steps of the proposed algorithm are detailed below:

Algorithm 1: Enhancing the purchasing of itemset (A, C) amongst frequent, known customers U, in store S

1. Create a set, L, containing all itemsets (A, C) that are not frequent in S but where its subsets A, C are frequent in S
2. Create a shortlist of L using Equation (4.15) and the extension of Lemma 1
3. For each A in the shortlist, create clusters as outlined in Section 4.3.1
4. Order the clusters based on Table 4.1
5. Run simulations as outlined in Section 4.4, re-prioritising the list based on the shortest time to frequency, noting that there may be some practical constraints as to why some target items cannot be aggressively promoted, e.g. supply shortages, and regulations

This algorithmic approach is different from that proposed in [140], in that it first identifies target items, and then targets customers who are most likely to purchase such items. The approach taken in [140] first groups customers, and then finds items

to target that will suit each of these groups. It was hypothesised that the approach taken in [140] enhances the potential for "false positives" and "false negatives" as it groups customers into clusters without verifying their purchase history for any affinity to the target product, which could lead to increased "false positives". Further, each group is targeted with a different itemset, hence there may be some customers within a group that would have purchased items that were offered to other customer groups only, thus increasing the likelihood of "false negatives". This hypothesis is tested in Chapter 5.

4.6 Summary

This chapter provided the foundation for this study, and is at the heart of this study's unique contribution to the body of knowledge on MBA, ARM, and targeted promotions. In this regard, a summary of the key aspects discussed in this chapter is provided below:

- A generalised model for target selection, or decision-making, between two choices, e.g. $(A \rightarrow C)$ and $(B \rightarrow D)$ within the MBA context, was not readily available from past research.

- Given this, and the potential usefulness of such a model, the market target (mt) model was developed as part of this study using the research methodology outlined in Chapter 3. This model allows grocery retailers to decide quickly, and effectively between two choices for targeting purposes, and thus benefit from savings in time, effort, costs, and in some applications, lives.

- The mt model was developed following an attempt to use the uninorm, which showed promise in other multi-criteria decision making applications, but failed in this instance, due to its inability to compensate for high support or high confidence dominance. In this regard, the mt model proved robust when tested

for both the Apriori and High Support/Confidence Dominance conditions detailed in Section 3.3.5, and was thus selected as the model of choice for itemset targeting, i.e. decision-making between two choices.

- Targeting customers was done using a FCM clustering algorithm that was developed based on the RFM framework, and taking into account both the customers' purchase history, and loyalty to the store. Customers within each cluster displayed similar behaviour, and which was different to other clusters, hence it was now possible to offer a tailored "treatment" plan for each cluster, thereby enhancing the purchasing of the targeted itemset without unnecessarily eroding the grocery retailer's revenue.

- A simulation model based on Markov chains, was also developed to simulate the impact of various marketing campaigns on overall sales of the target itemsets. This provided a mechanism to predict the impact on sales, and to make adjustments, in advance, to optimise sales campaigns, thus potentially saving costs and time.

- The three separate models for itemset targeting, customer targeting, and sales simulation were combined to form a coherent algorithm for enhancing the purchasing of target itemsets.

- Thus, the unique contributions of this chapter are:

  (1) The mt model for itemset targeting

  (2) The customer clustering approach to segregate customers for targeted treatment

  (3) The novel algorithm that combined the mt model, the customer clustering approach, and the simulation tool to enhancing the purchasing of the targeted itemset.

Chapter 5

Results and Discussion - Grocery Retail

5.1 Introduction

The effectiveness of the mathematical models, and algorithm proposed in Chapter 4, were tested by conducting experiments, as outlined in Chapter 3, using real-life shopping data, and for completeness, synthetic market basket data. The overarching objectives of the testing were to establish whether: (1) the proposed models and algorithm obey the mathematical underpinnings that govern MBA, as outlined in Chapter 4, and (2) the models and algorithm achieve the underlying business objectives, i.e. they enhance targeted promotions.

To achieve the above objectives, experiments were divided into four tasks: (1) identify target itemsets, using the mt model, (2) identify target customers, using the proposed clustering approach based on FCM, (3) simulating the impacts on sales of the targeted item, and (4) comparing the proposed algorithm against the approach detailed in [140], and against what actually happened using the cross-retailer, consumer scan-

ner panel data.

The chapter commences by detailing the experimental process and conditions that was used as part of this study. This is followed by a presentation and discussion of the results of the experiments conducted on itemset targeting, using the mt model, the clustering algorithm, and the targeted promotion algorithm. Results from the simulation testing, and model comparison experiments are then discussed, before the chapter concludes with a summary of key points.

5.2 Experimental Process

Experiments were conducted based on the Knowledge Discovery in Databases (KDD) process and Systems Development Framework (SDF) as detailed in Section 3.3. The heart of the KDD process is the data mining phase which leverages models and algorithms to process data into information, [160]. In this regard, the Apriori and FCM algorithms form part of the data mining phase while the proposed market target model and the simulation forms part of the post processing phase, where patterns are interpreted to select the best information that contributes to overall knowledge [160]. Similarly within the SDF, the testing of the model, the algorithm and performing simulations are all part of the experimentation and observation phases.

The key process steps of this experimental approach, i.e. first identifying target items, followed by identifying suitable customers for these items, followed by touting these customers with incentives for these items, and then finally retaining these customers for the long term is in line with the CRM model proposed in [120] and detailed in Section 3.3. Thus, it is through the combining of elements from the KDD, SDF and CRM models that this study enhances targeted promotions.

5.3 Experimental conditions

5.3.1 Data Sources

The 2012 and 2013 consumer scanner panel dataset obtained from [90] was used as the basis for the experiments. In general, the data sets contained over 32,000 unique customers and over 51 million individual scanned items, across 21 stores in the UK. Items were grouped at the product category level, and given an anonymised, unique identifier. For example, all milk products, including whole milk, goat's milk, etcetera were classified under the product category "milk". In total, there were 286 product categories, hence the anonymised datasets used in this study comprised of 286 unique items. All store formats of the same store were also combined, e.g. internet, express, garage shop etcetera. Three stores were chosen, a "Big Four" store, (Store 9), a hard discounter, (Store 13), and a high-end grocery retailer, (Store 21).

For completeness, the model was also tested on large, dense datasets to evaluate its processing performance. These datasets were created using a synthetic transaction database generator as detailed in [83], with the largest dataset consisting of 5 million transactions, 100 unique items, 5 frequent itemsets, frequency density of 0.5, and a maximum basket size of 50 items. It should be noted that this dataset is similar in transaction volume, but considerably denser, to that of the UK's largest grocery retailer's daily activity, and represents approximately 28% of the UK's grocery retail market share [131].

5.3.2 Identifying target itemsets

Frequent itemsets were mined using the Apriori algorithm as detailed in [7] for all itemsets with an initial minimum support of 0.05, and minimum confidence of 0.1. These "weak" constraints were placed on the dataset to prune highly infrequent, and

poorly associated itemsets. It should be noted that labelling these conditions as "weak" is justifiable as typical grocery retail datasets, like the datasets in this study, have highly associated items with confidence greater than 30%, and support of over 45%. Highly frequent single items typically have supports of greater than 60% [90].

Computer programs were written using R software to mine and analyse the data, whilst Microsoft Excel was used to compute the mt values and perform simulations using the techniques outlined in Chapter 4. Note that this study does not perform inter-store comparisons, hence the size of each selected store is not important. Indeed, intra-store comparisons are made, and whilst it is likely that some goods are purchased more frequently in-store or online than others, it is assumed that the pricing policy is the same across all formats and that customers have full choice in selecting a format that best suits them. These assumptions are considered fair given the prominence of multi-channel shopping, the emphasis that stores place on consistency across all channels, and the growing adoption of the "customer is king" mentality by UK retailers [47][157].

The data for each store, S, was analysed and target itemsets were identified based on the criteria that there exists two items, A and C, which are frequent, with $minsup = 0.1$, but their combination (A, C) is not frequent. Further (A, C) is the optimal combination to target, from all identified targets, based on the mt value obtained using Equation (4.15).

5.3.3 Identifying target customers

Target customers for each targeted itemset, (A, C), in each store, S, were identified based on the customer's purchase history and household size, using the FCM algorithm detailed in Chapter 4. Customer clusters were then classified based on the criteria outlined in Table 4.1. To eliminate "false positives", customers had to have visited

the store at least twice in the year, or the period under consideration, and purchased the antecedent, A, and the consequent, C, at least once during that period.

5.3.4 Simulating the impacts of the proposed model

An ergodic Markov model was created using Microsoft Excel to simulate the impact of marketing interventions on the shopping behaviour of the identified target customers. The model was based on the concepts outlined in Section 4.4. Two campaigns, one "conservative" and one "aggressive", were used, with their corresponding proportion vectors given in Tables 5.1 and 5.2, respectively. Note that the proportion vectors have a time period of 1 week. This is due to the assumption that all transactions made by a customer during a one week period may be considered to be one transaction, as detailed in Chapter 4. The following example is used to illustrate the concept of the proportion vector: from Table 5.1, 99% of all "switchers" in week 1 will remain "switchers" in week 2, while the other 1% will be elevated to "drop-outs". Then in week 2, 99% of all customers who are "switchers" will go on to remain as "switchers" in week 3, with the other 1% being elevated to "drop-outs". This process will recur whilst the treatment plan is in place.

		Future			
		Leave Alone	Light Touch	Drop Out	Switchers
Current	Leave Alone	1	0	0	0
	Light Touch	0.01	0.99	0	0
	Drop Out	0	0.01	0.99	0
	Switcher	0	0	0.01	0.99

Table 5.1: Conservative Marketing Campaign

It is clear that from Tables 5.1 and 5.2 that the general approach of the simulation campaigns, and indeed marketing, is to nudge customers towards increasing loyalty. Consequently, the proportion vectors in Tables 5.1 and 5.2 assume that the target customer's loyalty is always enhanced after each marketing campaign and not dimin-

		Future			
		Leave Alone	Light Touch	Drop Out	Switchers
Current	Leave Alone	1	0	0	0
	Light Touch	0.02	0.98	0	0
	Drop Out	0	0.02	0.98	0
	Switcher	0	0	0.02	0.98

Table 5.2: Aggressive Marketing Campaign

ished. This is a reasonable assumption to make, especially under the condition of zero "false positives", which this technique aims to achieve, and is demonstrated later on this chapter (see Section 5.4.2.1). In general, marketing initiatives provide incentives to customers to purchase more, and if those incentives are correctly targeted, e.g. providing incentives for goods that customers want to buy, then these incentives are most likely going to be acted upon [120][126][141]. This notion of "nudging customers towards loyalty" is discussed further in Section 5.4.3.

5.3.5 Comparative Testing of the proposed algorithm

The performance of the proposed algorithm was tested using the principles outlined in Chapter 3. The proposed model was compared with the approach detailed in [140] against the four tests detailed in Section 3.3.6. Further detail, together with the results and discussion of the performance testing of the algorithms are provided in Section 5.5. The proposed model was also compared against "reality", and tested using other data sets. The results and discussion of these tests are provided in Sections 5.5.2 and 5.6 respectively. Finally, and for completeness, the proposed model was also compared with alternative approaches, i.e. targeting "top sellers", and the results and discussion of this test is detailed in Section 5.5.3.

5.4 Results and Discussion

5.4.1 Identifying Target Items

The support, confidence, and mt value, given by Equation (4.15), were computed for all associated itemsets for Stores 9, 13, and 21, for both the 2012 and 2013 datasets. It should be noted that the data was initially pruned on the lower-side using $minsup = 0.05$ and $minconf = 0.1$. Following this, a $minsup = 0.1$ was used to determine whether an itemset is frequent. The total number of infrequent itemsets, where the support of the itemset is less than 0.1, but greater than 0.05, is significantly large, as is often the case in real life, with a breakdown provided in Table 5.3. It is clear from the large number of infrequent itemsets, as shown in Table 5.3, that choosing the best itemset(s) to target for marketing purposes can be complex. This is a well-known managerial challenge, particularly with large datasets, and is well-documented in the literature [77][175][189]. In this regard, the mt value, given by Equation (4.15), can be an effective tool in addressing this managerial challenge.

	2012 Dataset	2013 Dataset
Store 9:		
Frequent 1-itemsets	48	46
Frequent itemsets	445	434
Infrequent itemsets	3258	3218
Store 13:		
Frequent 1-itemsets	29	32
Frequent itemsets	102	151
Infrequent itemsets	446	751
Store 21:		
Frequent 1-itemsets	18	18
Frequent itemsets	57	67
Infrequent itemsets	245	276

Table 5.3: Results of Frequent Itemset Mining with $minsup = 0.1$ and $minconf = 0.1$

To illustrate the effectiveness of the mt model in supporting decision making, consider the selection of results presented in Tables 5.4 and 5.5. The results includes the sup-

port, confidence, and mt value for several infrequent itemsets, where $minsup = 0.1$, for the 2012 and 2013 datasets respectively. Clearly, the best itemset to target is that itemset which has both the highest support and confidence as it is the most popular and requires the least "effort", e.g. marketing investment in the form of: people, money, space, etcetera, to increase sales [56][165]. From the dataset extract presented in Table 5.4, the best itemsets to target are: (156,277) for Store 9, (284,277) for Store 13, and (135,163) for Store 21, as all three have the highest support and confidence for their store-related extracts. This is consistent with principle P1 as outlined in Section 3.3.5. Further, it can also be seen that these best itemsets have the lowest mt value for the respective stores, hence they will require the smallest number of customer transactions to be targeted for these itemsets to become frequent. Similarly itemsets (68,88) in Store 9, and (107,270) in Store 21 are considered the worst itemsets for targeting, from the given dataset extract in Table 5.4, as they have the lowest support and confidence. Yet again, this is consistent with principle P1 as outlined in Section 3.3.5. The market target model also confirms this result, as these two itemsets have the highest mt values for their respective stores in Table 5.4. Note that mt value can be computed from the support and confidence values, where $\text{supp}(A)$, which is required for the mt calculation, is given by $\text{supp}(A, C)/\text{conf}(A, C)$.

The highlighted rows are of particular interest, as the optimal choice cannot be easily made by merely inspecting both the support and confidence values. Hence, this choice is best determined by using the mt equation detailed in Equation (4.15). Consequently, (68,270) and (153,268) in Stores 9 and 13 respectively are the better choices from the highlighted rows as they have smaller mt values.

The scenario is significantly more complex in the dataset extract presented in Table 5.5, where every selected itemset within each store is not immediately superior, from

Itemset	supp(A, C)	conf(A, C)	supp(A)	mt	Target Priority
Store 9					
156 to 277	0.072	0.696	0.103	0.402	Priority 1
68 to 270	0.063	0.593	0.106	0.623	Priority 2
156 to 268	0.062	0.600	0.103	0.632	Priority 3
68 to 88	0.058	0.541	0.107	0.784	Priority 4
Store 13					
284 to 277	0.073	0.705	0.104	0.384	Priority 1
213 to 163	0.060	0.584	0.103	0.688	Priority 2
153 to 268	0.058	0.534	0.109	0.792	Priority 3
146 to 274	0.056	0.548	0.102	0.799	Priority 4
Store 21					
135 to 163	0.074	0.677	0.109	0.384	Priority 1
57 to 274	0.059	0.593	0.100	0.695	Priority 2
107 to 268	0.058	0.508	0.114	0.830	Priority 3
107 to 270	0.051	0.444	0.115	1.115	Priority 4

Table 5.4: Target Itemsets - Stores 9,13, and 21 for 2012 Dataset

Itemset	supp(A, C)	conf(A, C)	supp(A)	mt	Target Priority
Store 9					
57 to 88	0.077	0.514	0.150	0.443	Priority 1
36 to 270	0.065	0.549	0.118	0.642	Priority 2
33 to 268	0.061	0.594	0.103	0.664	Priority 3
68 to 88	0.056	0.553	0.101	0.798	Priority 4
Store 13					
126 to 88	0.060	0.491	0.122	0.822	Priority 1
146 to 268	0.054	0.521	0.104	0.882	Priority 2
35 to 110	0.069	0.287	0.240	1.086	Priority 3
78 to 251	0.068	0.243	0.280	1.309	Priority 4
Store 21					
107 to 270	0.056	0.469	0.119	0.965	Priority 1
135 to 268	0.051	0.466	0.109	1.048	Priority 2
78 to 209	0.058	0.351	0.165	1.190	Priority 3
88 to 285	0.059	0.336	0.176	1.226	Priority 4

Table 5.5: Target Itemsets - Stores 9,13, and 21 for 2013 Dataset

a targeting perspective, to the other items in that store. For example, in Store 9 in Table 5.5, (57,88) has a significantly higher support than itemset (68,88), but a lower confidence. It is thus clear that more detailed calculations are required before a

decision on the prioritisation of which itemset to target can be made. In this regard, the merits on the use of the mt value in supporting decision making are clearly evident.

The mt model, has been derived from first principles, hence by definition, it will always result in the optimal target itemset having the smallest mt value. However, for completeness, the calculations presented in Tables 5.6 and 5.7 illustrate the relationship of the first principles calculation for the prioritisation given in Table 5.4 and the mt value.

	Total Volume	Min Support
Store 9	328210	32821
Store 13	221330	22133
Store 21	47170	4717

Table 5.6: Summary of Transaction Volumes, $minsup = 0.1$, 2012 Dataset

Itemset	Volume (A, C)	conf(A, C)	Required Volume	Required Target	mt	Target Priority
Store 9						
156 to 277	23620	0.696	9200	13200	0.402	Priority 1
68 to 270	20700	0.593	12120	20400	0.623	Priority 2
156 to 268	20360	0.600	12460	20700	0.632	Priority 3
68 to 88	18900	0.541	13920	25700	0.784	Priority 4
Store 13						
284 to 277	16150	0.705	5980	8500	0.384	Priority 1
213 to 163	13230	0.584	8900	15200	0.688	Priority 2
153 to 268	12770	0.534	9360	17500	0.792	Priority 3
146 to 274	12450	0.548	9680	17700	0.799	Priority 4
Store 21						
135 to 163	3490	0.677	1230	1820	0.384	Priority 1
57 to 274	2800	0.529	1920	3630	0.695	Priority 2
107 to 268	2730	0.508	1990	3920	0.830	Priority 3
107 to 270	2380	0.444	2340	5260	1.115	Priority 4

Table 5.7: Target Itemsets - Stores 9, 13, and 21 for 2012 Dataset

The detailed calculations for the data in Table 5.7 is as follows and uses Store 9, itemset (68,270) as an example. The Volume (A, C) represents the actual transaction volume of the itemset, and is given by the product of the support of the itemset in Table 5.4, and the total volume in Table 5.6. Hence $0.063 \times 328210 = 20700$, rounded to three significant figures. The Required Volume is the volume required to make the itemset frequent and is computed by the difference between Volume (A, C) and Min Support, given in Table 5.6. Given that the itemset confidence is in essence the conditional probability of the itemset being present, given the antecedent, i.e. $P(C|A)$, the Required Target of transactions for itemset (A) thus becomes the required volume divided by the confidence of itemset (A, C). Clearly, these calculations are tedious, time consuming, and can hinder effective decision making when the numbers are large. However, the mt value achieves the same calculation in a single step, that is, the required target is equal to the product of the mt value and Min Support, given in Table 5.6.

Based on the above, it is thus clearly evident that the mt value simplifies itemset targeting. Further, given that the minimum support for all transactions within the store may be similar, merely comparing the mt value for each itemset will accurately establish target priority. The mt value is also a unit-less, normalised parameter, hence the "effort" to make any itemset frequent is relative to its minimum support. In light of this, the mt value may be used to compare the relative "effort" required to make itemsets frequent across stores and/or across minimum support thresholds. For example, the transaction volume in Store 9 is significantly higher than that of Store 21 (see Table 5.6) but the "effort" required by Store 9 to make itemset (68,88) frequent will be substantially lower than the "effort" required by Store 21 to make itemset (107,270) frequent, even though the required target volume is approximately 5 times greater in Store 9 compared to that of Store 21. Similarly, the mt values

of (284,277) in Store 13 and (135,163) in Store 21 are the same at 0.384, but the required transaction volume in Store 13 is 4.7 times greater than that of Store 21. This directly correlates to the ratio of their minimum supports, as given in Table 5.6. In view of this, the mt model obeys principle P4 as outlined in Section 3.3.5, in that it is effective in consistently selecting the optimal target itemset when the choice is not immediately obvious.

5.4.1.1 Itemset Monotonicity - Principle P2

The mt value was tested for monotonicity in line with the practical considerations outlined in [7] and the conclusions in Lemma 1, and Moodley et al. in [115]. The monotonicity property of comparing two itemsets with the same antecedent is detailed in P2 of Section 3.3.5. Frequent antecedents, A, were selected from the three stores (9, 13, 21) for the 2012 dataset, with the mt value calculated for a variety of (A, C) combinations, and presented in Figure 5.1. It can be seen that the mt value monotonically decreases for increasing $\text{supp}(A, C)$, which is consistent with Equation (4.15), Lemma 1, the practical considerations outlined in [7], and the conclusions in [115]. This implies that the higher the frequency of an infrequent itemset, the fewer the customer transactions required, for targeting, so that the itemset may become frequent.

5.4.1.2 Summary comments from Itemset Targeting

The results presented in this section confirms that the mt model obeys the key principles detailed in Section 3.3.5, which are underpinned by all previous studies on MBA and FIM including [5], [24], [76], and [175]. Indeed, the discussion in Section 5.4.1 has demonstrated that the mt model is both an effective and rapid way of deciding the optimal itemset for targeted promotions. A detailed review of previous work has showed that no similar model exists for itemset targeting, hence it can be concluded

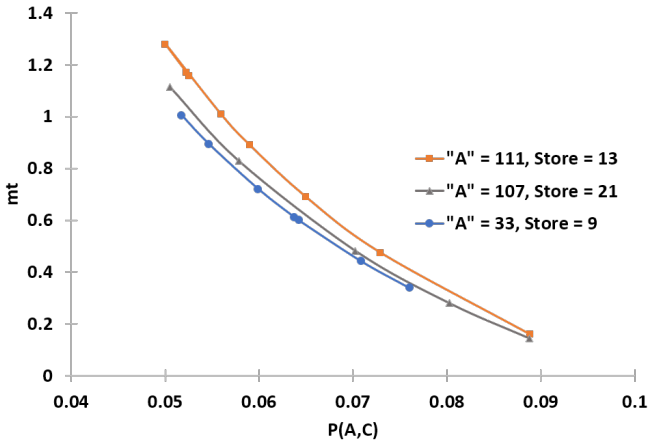


Figure 5.1: Monotonic Property of the Proposed Model

that the mt model is indeed a unique contribution to the study of MBA.

5.4.2 Identifying Target Customers

Target customers were identified using the principles outlined in Section 5.3.3 for all itemsets presented in Tables 5.4 and 5.5. The number of customers together with their "false positive" potential is presented in Tables 5.9 and 5.10 for the 2102 and 2013 datasets respectively. Note that the "false positive" potential for all customers is zero, as only customers that have purchased both the antecedent and consequent, whether together or separately, during the year have been selected. Hence, every target customer has the potential to purchase the itemset (A, C) if incentivized with offers for (C) or both (A) and (C). Thus the mt model satisfies the requirements of Test 1 in Section 3.3.6. "False positive" pose significant challenges in grocery retail, and was cited as a key challenge in [116] and [145]. The approach taken in this study ensures that this issue is eliminated altogether. Moreover, this approach reduces both the data mining effort, and unnecessary marketing spend, as it initially prunes all customers with "NULL" values for the target item before further processing is undertaken. This saves computational effort in clustering and avoids artificially inflated customer target groups, who potentially could receive expensive marketing incentives

that are not taken up. Equally importantly, this approach prunes all customers that may have an aversion to the target item despite buying the antecedent, e.g. allergy, cultural, or dietary aversions. This prevents the store from losing or angering customers as a result of falsely targeting them, for a product that they are never going to buy.

	2012	2013
Total	32726	32625
Store 9	23169	22843
Store 13	17981	19501
Store 21	6157	6500

Table 5.8: Summary of Customers

It can be seen by comparing the target customers in Table 5.9 with the required volume in Table 5.7, that the required volume typically exceeds the target customers. The required volume is an annual metric, as the period under consideration is one year, and since one objective of marketing is to create a sustained increase in purchasing of the target item by the target customers, the required volume is likely to be reached over the year, if the marketing treatment is successful [141]. It is in this light that the marketing simulations, as outlined in Section 5.3.4, is performed on the target customer base to understand the long term impacts of marketing campaigns on the purchasing, and support, of the targeted item. It can also be seen by comparing the target customers in Tables 5.9 and 5.10 with the total number of customers per store in Table 5.8, that in general, only a fraction of the customers, up to a third, are targeted in any one campaign. Naturally, this is very beneficial as keeping target customer volumes low reduces marketing spend and does not force a price decrease across the entire customer base, which helps the grocery retailer sustain its revenue [120][134].

Target Itemset	Target Customers (Actual)	Target Customers (% of Total)	"False Positive" Potential
Store 9			
156 to 277	7472	32%	0%
68 to 270	7578	33%	0%
156 to 268	7249	31%	0%
68 to 88	7008	30%	0%
Store 13			
284 to 277	4494	25%	0%
213 to 163	5770	32%	0%
153 to 268	5070	32%	0%
146 to 274	5050	28%	0%
Store 21			
135 to 163	1158	19%	0%
57 to 274	1032	17%	0%
107 to 268	1050	17%	0%
107 to 270	1021	17%	0%

Table 5.9: Target Customers - Stores 9, 13 and 21 for 2012 Dataset

Target Itemset	Target Customers (Actual)	Target Customers (% of Total)	"False Positive" Potential
Store 9			
57 to 88	8326	36%	0%
36 to 270	7204	32%	0%
33 to 268	7259	32%	0%
68 to 88	6733	29%	0%
Store 13			
126 to 88	6309	32%	0%
146 to 268	5917	30%	0%
35 to 110	7004	36%	0%
78 to 251	8017	41%	0%
Store 21			
107 to 270	1189	18%	0%
135 to 268	1200	18%	0%
78 to 209	1642	25%	0%
88 to 285	1158	18%	0%

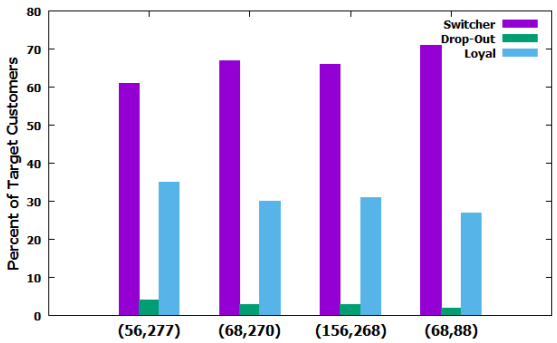
Table 5.10: Target Customers - Stores 9,13 and 21 for 2013 Dataset

5.4.2.1 Clustering Customers

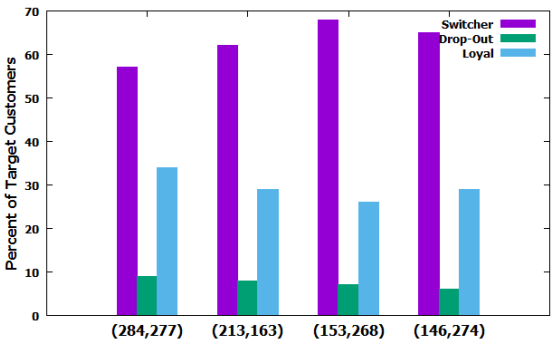
The data in Tables 5.9 and 5.10 were clustered using the FCM principles outlined in Sections 5.3.3 and 4.3.1, with the results of the clustering presented in Figures 5.2

and 5.3. Clearly, most customers fall into the "Switcher" cluster because it comprised of four of the nine clusters in Table 4.1, and the selected antecedents are marginally frequent, as shown in Tables 5.4 and 5.5, i.e. having frequencies just over the targeted *minsup* of 0.1. This implies that these items are not "top sellers" in these stores. In any event, if the antecedents were "top sellers", e.g. *bread* or *milk*, then it will be likely that the clustering mix will be swayed towards loyalty. Whilst it is not in the best interest of stores to promote "top sellers", as there is always strong demand and it can erode revenue, it should be noted that this model can effectively separate customers under these conditions as well, thereby allowing for the execution of different treatment measures to different clusters or no treatment at all [66]. This is an important consideration as it meets the requirements of Test 3 in Section 3.3.6. The notion of having antecedents as "top sellers" and targeting customers with promotions for items of lower support is discussed further in Section 5.5.3.

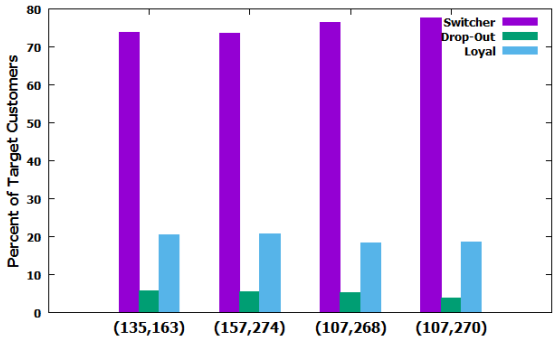
In general, having a larger "Switcher" group and smaller loyalty-based groups, i.e. "Drop Out", "Light Touch", "Leave Alone", is ideal from a marketing perspective as it allows grocery retailers the opportunity to spend most of its marketing budget on attracting customers who would otherwise spend their money elsewhere. This will help grocery retailers drive-up their regular, and/or large spending customer base, thus increasing their revenue. This is consistent with marketing studies that prioritise marketing spend on attracting new customers [154]. It is well-documented that spending marketing money on "Loyal" customers is not ideal, as it not only lowers the price expectations with the most frequent/high value customers, but it also erodes revenue as a large proportion of these customers would not necessarily switch to other stores in the absence of promotions [120][154][166]. In line with this, it can be concluded that the proposed mt model, and FCM clustering approach is an effective approach to optimising the targeting of customers for marketing promotions,



(a) Target Customer Splits - Store 9, 2012



(b) Target Customer Splits - Store 13, 2012



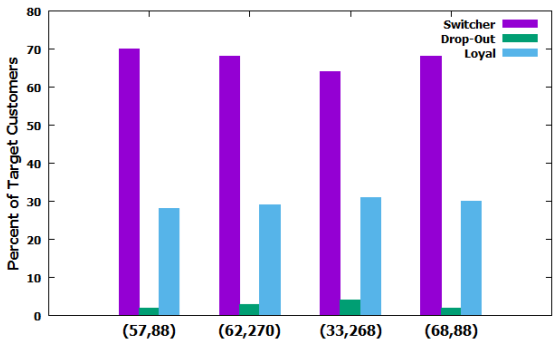
(c) Target Customer Splits - Store 21, 2012

Figure 5.2: Target Customer Splits - 2012 Dataset

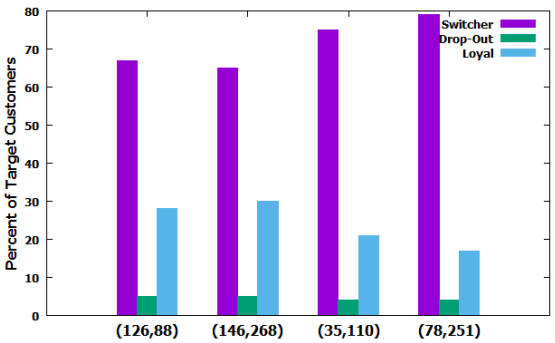
as it correctly identifies the best items to target and it correctly segments customers in a way to maximise the impact of marketing promotions by targeting a large population of customers who purchase the target item elsewhere.

5.4.3 Simulating the impacts of the proposed model

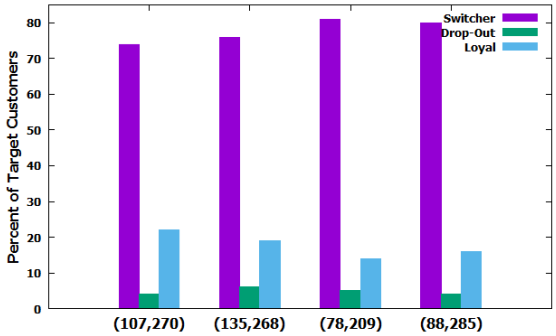
A Microsoft Excel-based Markov simulator, created as part of this study, was run using the conditions outlined in Section 5.3.4. The various cluster start sizes, for the selection of target itemsets, are given in Tables 5.11 and 5.12 for the 2012 and 2013 datasets respectively. The datasets were tested for both the "conservative" and "aggressive" campaigns. Note that Clusters 5 and 8, which are essentially "Loyal" clusters, have now been classified as "Light Touch". The rationale for this is that customers who fall into this category sit on the cusp of loyalty and if presented with an opportunity to switch to other stores, they will most likely take it. These clusters



(a) Target Customer Splits - Store 9, 2013



(b) Target Customer Splits - Store 13, 2013



(c) Target Customer Splits - Store 21, 2013

Figure 5.3: Target Customer Splits - 2013 Dataset

are in contrast to the highly loyal cluster, Cluster 9, where switching behaviour is unlikely.

Cluster	Approach	Store 9 (156,277)	Store 9 (68,88)	Store 13 (213,163)	Store 13 (153,268)	Store 21 (135,163)	Store 21 (107,268)
1	Switcher	14%	14%	21%	20%	31%	27%
2	Switcher	29%	37%	31%	35%	31%	34%
3	Switcher	11%	13%	6%	8%	6%	11%
4	Drop Out	3%	2%	7%	5%	5%	4%
5	Light Touch	18%	14%	18%	16%	11%	10%
6	Switcher	8%	8%	4%	4%	6%	4%
7	Drop Out	0%	0%	2%	1%	1%	1%
8	Light Touch	9%	6%	9%	9%	4%	4%
9	Loyal	8%	6%	2%	2%	5%	5%

Table 5.11: Target customer cluster sizes at start - 2102 Dataset

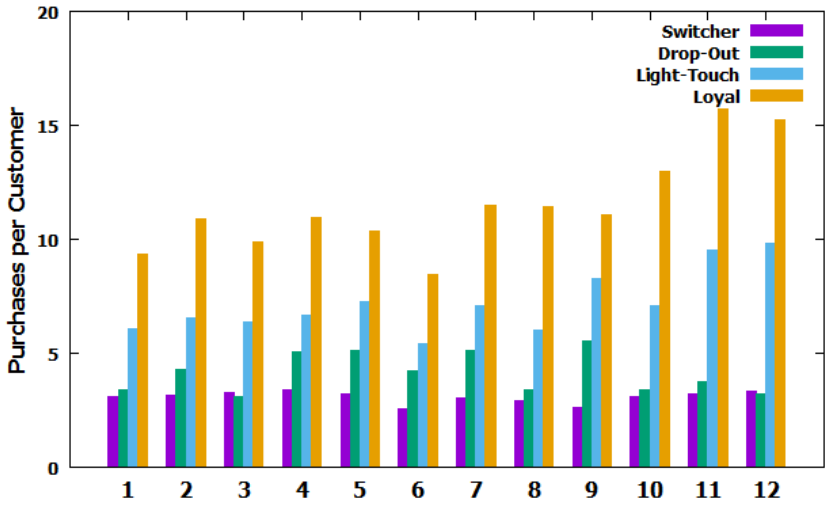
Cluster	Approach	Store 9 (57,88)	Store 9 (68,88)	Store 13 (146,268)	Store 13 (35,110)	Store 21 (107,270)	Store 21 (78,209)
1	Switcher	15%	13%	15%	20%	27%	37%
2	Switcher	35%	33%	35%	41%	32%	31%
3	Switcher	11%	12%	9%	10%	9%	8%
4	Drop Out	2%	2%	4%	4%	4%	4%
5	Light Touch	16%	17%	20%	15%	13%	9%
6	Switcher	8%	9%	6%	4%	6%	6%
7	Drop Out	0%	0%	1%	1%	0%	0%
8	Light Touch	6%	6%	8%	4%	5%	3%
9	Loyal	6%	6%	3%	1%	4%	2%

Table 5.12: Target customer cluster sizes at start - 2103 Dataset

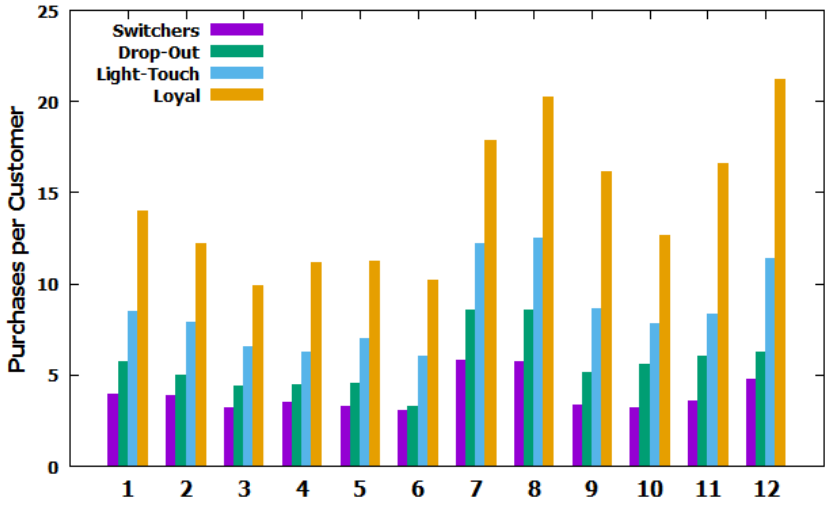
5.4.3.1 Nudging customers towards loyalty

The counter-argument to "nudging customers towards loyalty" may be stated as follows: do customers that are classified as loyal buy more of the targeted product than their counterparts who are classified as "switchers"? Intuitively, the answer to this question will be yes, as the clustering is done in a way to ensure that those that purchase larger volumes of an item are classified as loyal. Hence, the approach of nudging towards loyalty is justified. However, for completeness, this was tested empirically using the 2012 and 2013 datasets for the twelve itemsets given in Tables 5.9 and 5.10 respectively, with the results presented in Figures 5.4 and 5.5. For simplicity, the itemsets in Tables 5.9 and 5.10 were assigned numbers 1 to 12, which corresponds to the order in which they appear on the tables.

From Figures 5.4 and 5.5, it is evident that the number of times an itemset is purchased per customer increases with customer loyalty. Hence, the more loyal the customer, the greater their purchases of the targeted itemset. Thus, nudging customers towards loyalty, including through the use of marketing initiatives, is a good strategy for grocery retailers to pursue.



(a) Antecedent Item Purchases - 2012

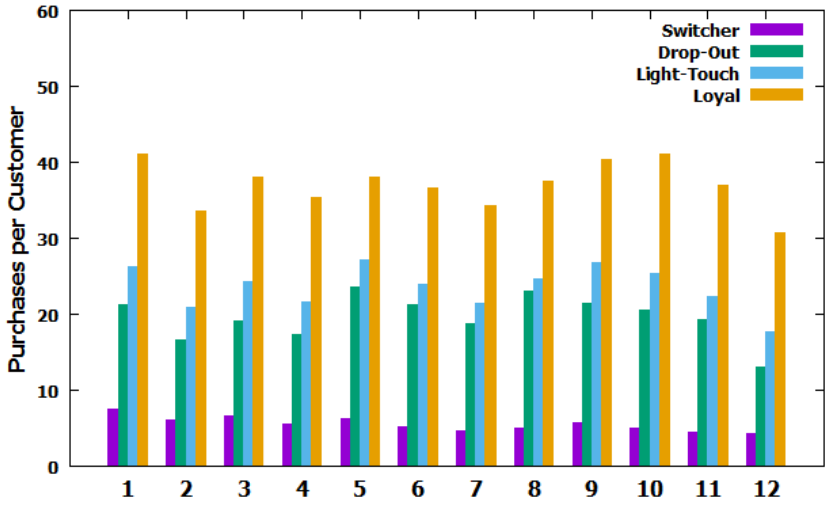


(b) Antecedent Item Purchases - 2013

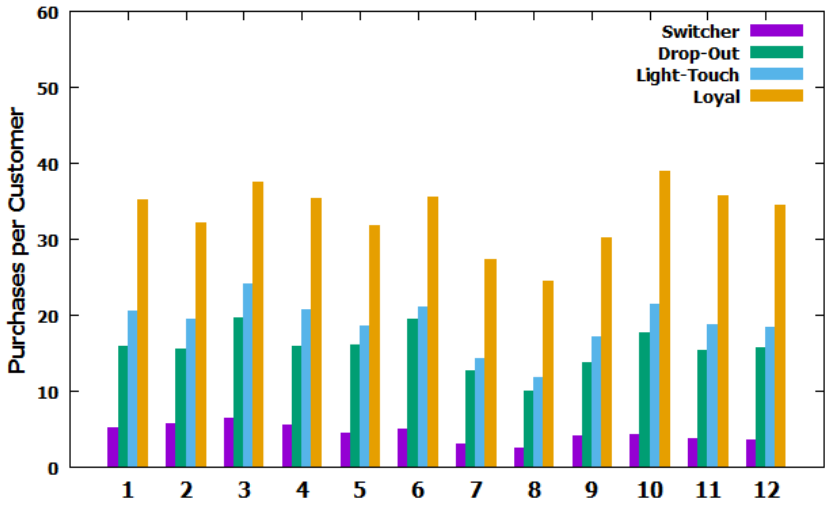
Figure 5.4: Antecedent Item Purchases per Customer

5.4.3.2 Simulating marketing impacts

Four itemsets, two each from Tables 5.11 and 5.12, were selected for simulation testing. The itemsets, together with their target customer, and "purchases of the antecedent per customer" splits are provided in Table 5.13. The effectiveness of the proposed approach can be clearly seen from the data in Table 5.13. Loyal customers, within the target customer set, purchase the antecedent far more than any other group, e.g. 48% for Store 21 (78,209) in the 2013 dataset, whilst "switchers" make up the lowest



(a) Consequent Item Purchases - 2012



(b) Consequent Item Purchases - 2013

Figure 5.5: Consequent Item Purchases per Customer

number of purchases, e.g. 10% for Store 21 (78,209) in the 2013 dataset. It does make sense, from a marketing perspective, that marketing initiatives should focus on predominantly targeting "switchers", and nudging these customers towards loyalty, bearing in mind that loyal customers are unlikely to defect to other stores, and marketing to loyal customers drives down overall pricing and erodes revenue [120][134]. This is precisely what the proposed algorithm does - it predominantly targets "switchers", e.g. 82% for Store 21 (78,209) in the 2013 dataset, and over time "nudges them towards loyalty", where purchases of the antecedent per customer are very high.

At a practical level, customers that are classified as "switchers" will get vouchers to buy the consequent target item, or combined offers where if they buy the antecedent, then they will receive a discount on the consequent target item. Similar, but smaller incentives may also be offered to "Drop-Out" and "Light-Touch" customers. For example: 82% for Store 21 (78,209) in the 2013 dataset are switchers and account for 10% of purchases of the antecedent. This group will be heavily incentivised to make purchases of the consequent, which will then see them move up the "loyalty chain" as they make more purchases. As time progresses, the percentage of switchers reduces while the percentage of loyal customers increases, resulting in increased volumes of the target itemset being purchased, and thus making it frequent.
The simulation output for each marketing scenario ("conservative" and "aggressive") for the datasets presented in Table 5.13 are shown in Figures 5.6, 5.7, 5.8 and 5.9. The modelling was done on a per-week basis. The simulation tests show that it is possible to achieve the required weekly target volume (denoted by the grey dotted, horizontal line on each chart) for the various itemsets with both "conservative" and "aggressive" campaigns. This is achieved by a conversion of customers from "switchers" to loyal customers, over time, using customised treatment for the various clusters. In Figure 5.7, it can be seen that the itemset (213,163) reached the required volume after six

	Store 9, (156,277) (2012)	Store 13, (213,163) (2012)	Store 9, (57,88) (2013)	Store 21, (78,209) (2013)
Target Customer Splits				
Switchers	62%	62%	69%	82%
Drop-Out	3%	9%	2%	4%
Light Touch	27%	27%	22%	12%
Loyal	8%	2%	7%	2%
Antecedent Purchase per Customer Splits				
Switchers	14%	12%	12%	10%
Drop-Out	16%	20%	18%	18%
Light Touch	28%	28%	26%	24%
Loyal	42%	40%	44%	48%

Table 5.13: Simulation Testing Data from 2012 and 2013 Datasets

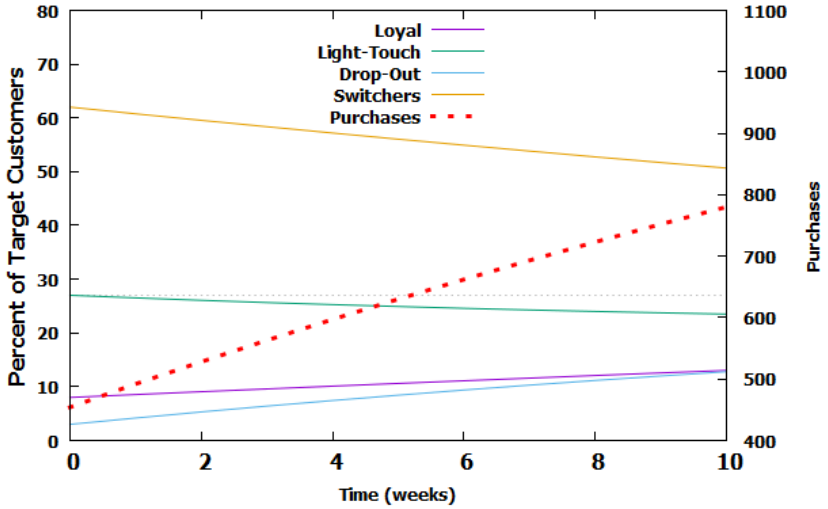
weeks of "conservative" treatment, and the volume will continue to increase as long as this treatment is in place. Note that over the ten week period, the percentage of "Loyal" customers increased from 2% to 7% whilst the number of customers classed as "switchers" declined from 62% to 50.6%. Thus, there is significant headroom to expand the number of loyal customers even further, although targets have been reached, and this increase could likely continue whilst this treatment is in place. In this regard, it is well-documented that customers are likely to switch and remain loyal to the new grocery retailer, if amongst other things, the price point is perceived to be permanent [97][126][141]. If however, the treatment were to stop, and the price point were to return to a higher value, then it is likely that the customers too would return to their previous ways [97].

On a contemporary and practical level, this "switching" behaviour by customers is consistent with the rise of the German discounters in the UK, and across Europe for that matter, where customers have switched to these discounters from traditional stores and remain loyal to these discounters as the price has remained low whilst the

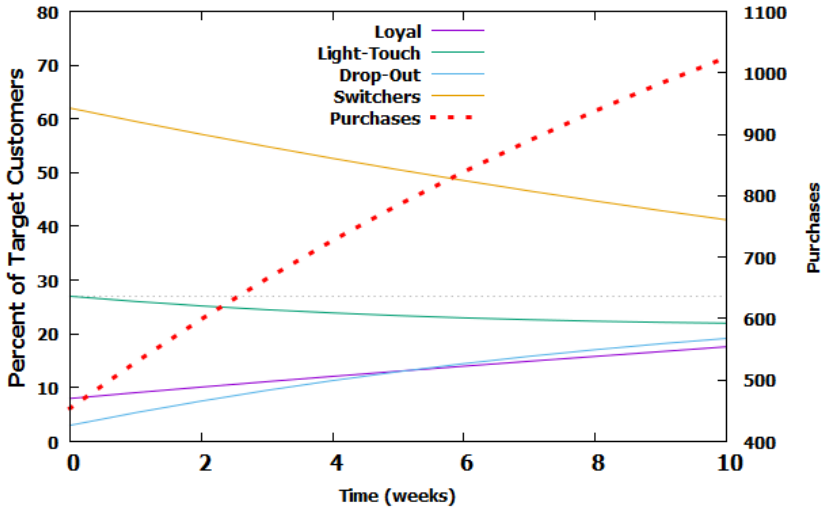
quality has either remained acceptable or has increased [66]. To further retain customers, the German discounters are rapidly expanding their footprint across Europe and enhancing their stores with value-added features like in-store bakeries, localised produce, and multiple payment methods, thus increasing localisation, and customer convenience whilst neutralising any differentiation that the traditional stores may have over them [66][73].

The "aggressive" simulation campaigns follow a similar mechanism to the "conservative" simulation campaigns, except that the incentives may be greater, and thus attract more customers to the grocery retailer quicker, see Figures 5.6 to 5.9. This results in a more rapid increase in purchases, with the target volume being achieved sooner. In addition, the conversion of customers from "switchers" to loyal also occurs at a faster pace. Given this, it can be seen why new store openings often have very attractive, low priced store opening sales, i.e. aggressive campaigns, as these are designed to rapidly attract customers, usually within the first two weeks of opening, and convert as many as possible to being loyal to the new store, and grocery retailer [66][73]. In the case of the German discounters, they have also ensured that their price points remained lower than that of their neighbouring traditional grocery retailers, thus retaining these converted customers after the opening sales have ended [66].

Based on the above simulation testing, its is clear that nudging customers towards loyalty, and treating customers with a tailored incentive plan that is based on their shopping patterns, can be very effective in increasing the frequency of targeted itemsets. Further, if the treatment plan is sustained, as is the case of a permanent price reduction, then the frequency of the target itemset is likely to increase above the require level with the store converting more customers from "switcher" to loyal.

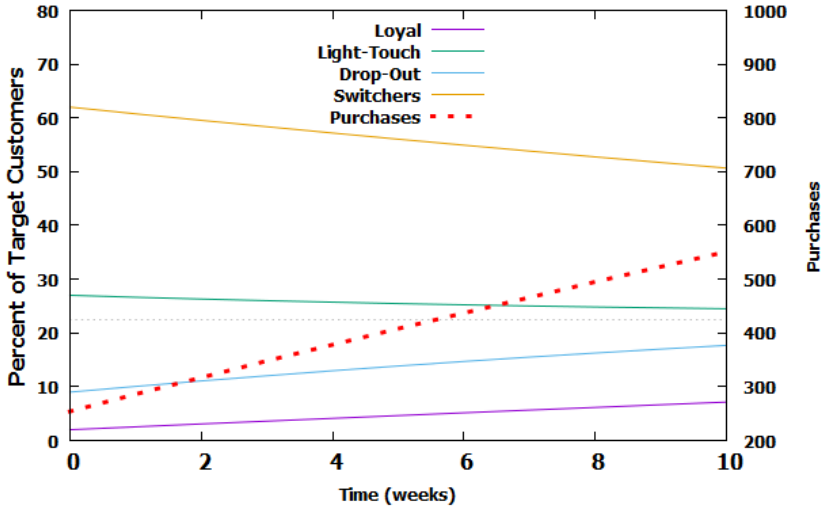


(a) Conservative - Store 9, (156,277)

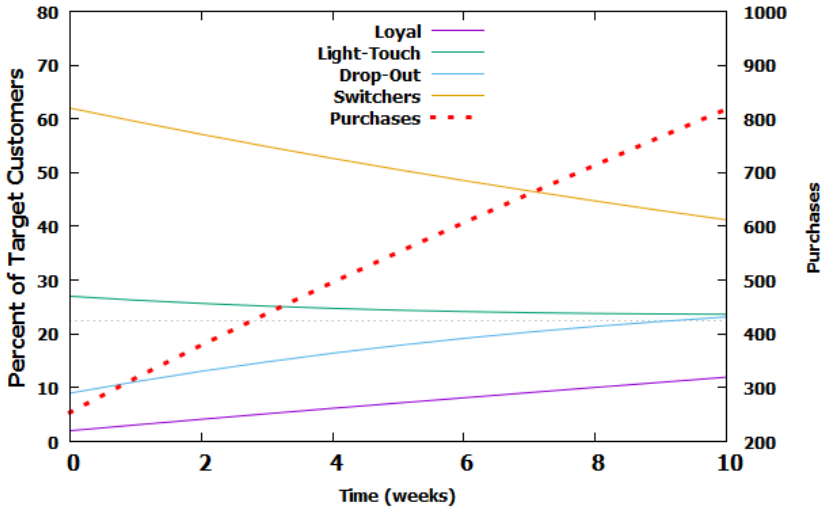


(b) Aggressive - Store 9, (156,277)

Figure 5.6: Simulation Results - Store 9, (156,277) - 2012 Dataset

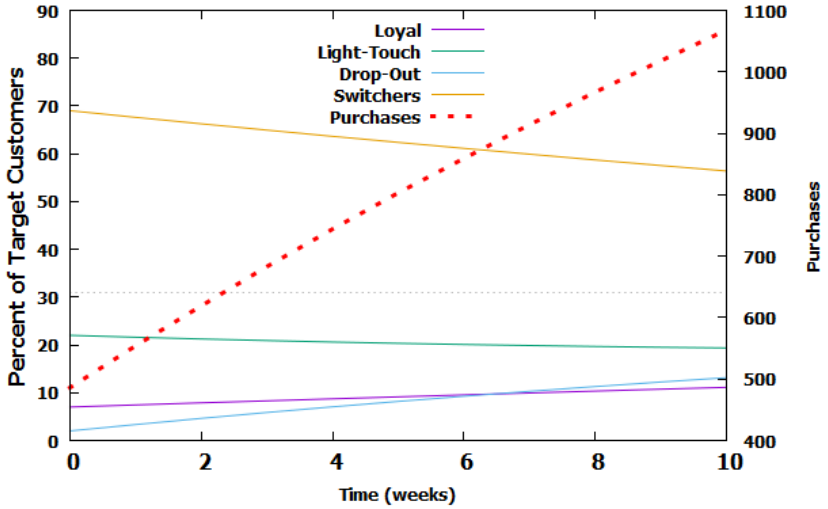


(a) Conservative - Store 13, (213,163)

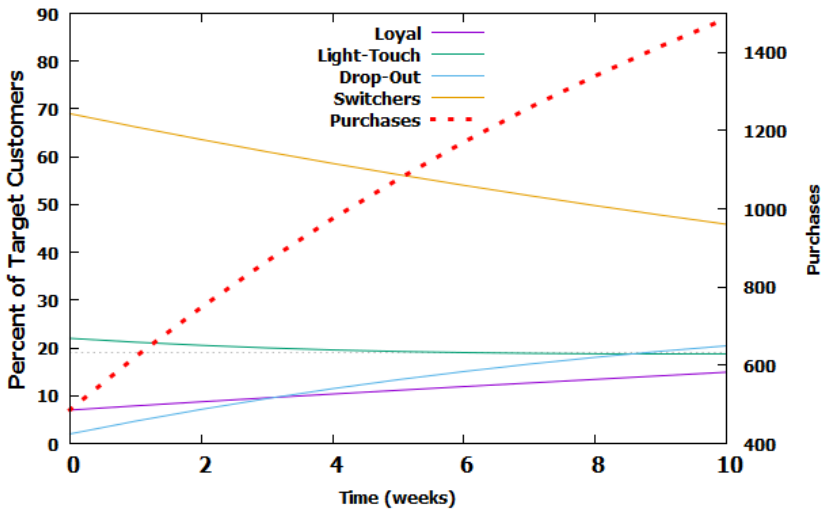


(b) Aggressive - Store 13, (213,163)

Figure 5.7: Simulation Results - Store 13, (213,163) - 2012 Dataset

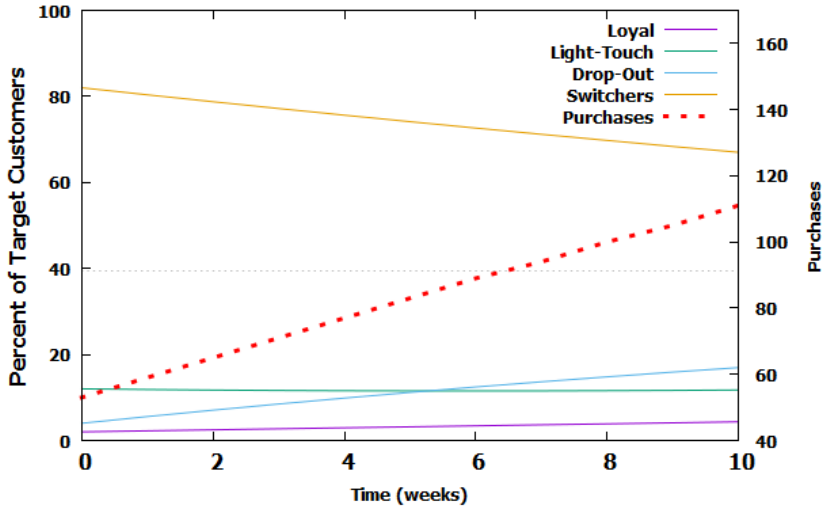


(a) Conservative - Store 9, (57,88)

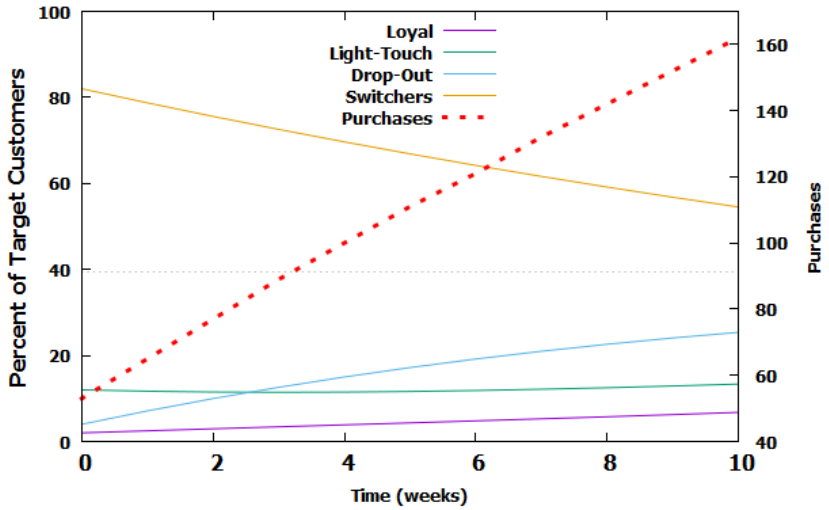


(b) Aggressive - Store 9, (57,88)

Figure 5.8: Simulation Results - Store 9, (57,88) - 2013 Dataset



(a) Conservative - Store 21, (78,209)



(b) Aggressive - Store 21, (78,209)

Figure 5.9: Simulation Results - Store 21, (78,209) - 2013 Dataset

5.5 Model Comparison Testing

5.5.1 Comparing the proposed algorithm with the approach detailed by Reutterer et al. in [140]

The approach detailed in [140] was run on the same itemsets as provided in Table 5.13. The experimental process was similar to the methodology detailed in [140], where the top 5% and the bottom 20% of customers, based on transaction size, were initially pruned. This was followed by an exclusion of high frequency items, the twelve most frequent items for each store, which as per [140], results in skewed clustering. Following this, eleven clusters were created for each of the four itemsets detailed in Table 5.13, with the most popular cluster for each antecedent being selected for treatment, i.e. marketing promotions. The comparison testing was conducted in two parts. The first part compared the effectiveness of the clustering approach, in particular its ability to reduce "false positives" and "false negatives", whilst the second part compared the model's ability to offer differentiated treatment to customers, with the focus on attracting new customers, and preventing revenue erosion.

5.5.1.1 Comparison of the effectiveness of the clustering approach

From Table 5.14 it can be seen that following the approach detailed in [140] results in a much smaller customer base being selected, compared with the relevant cluster sizes in Tables 5.9 and 5.10, which in itself is not a problem, provided that every customer is a highly likely candidate to act upon the treatment. However, in most cases only a fraction of the total target customers are known purchasers of the antecedent, and in some cases this fraction is as low as 42%, resulting in a very large "false positive" potential which is considered bad for marketing as it angers customers [116][145]. Further, a very large proportion of potential, good, target customers are left untreated as they fall in other clusters, as evidenced by the difference in the

number of target customers between Tables 5.9 and 5.10, and Table 5.14). This is a missed opportunity, and results in an increase in "false negatives". For example, in Store 13, itemset (213,163), 1521 customers will be targeted of which 58% will be falsely targeted as they have no history of purchasing the antecedent. At the same time, 5124 customers (approximately 89% of known purchasers of the antecedent, as identified by the algorithm proposed in this study) are placed into other, non-targeted clusters. These customers will not be targeted for the itemset, and this will result in a missed opportunity for the grocery retailer. Consequently, the model proposed in this study has a more effective clustering approach than the approach outlined in [140], as it is more effective both in terms of Test 1, minimising "false positives", and Test 2, minimising "false negatives" as detailed in Section 3.3.6 .

Target Itemset	Target Customers	Known Purchasers	"False Positive" Potential	"False Negative" Potential
Store 9, (156,277) (2012)	2331	1008	57%	6464
Store 13, (213,163) (2012)	1521	646	58%	5124
Store 9, (57,88) (2013)	2394	1253	48%	7073
Store 21, (78,209) (2013)	632	299	53%	1343

Table 5.14: Target Clusters and Customers based on the approach proposed in [140]

5.5.1.2 Ability to attract new customers and offer a differentiated treatment approach

The "Known Purchasers" detailed in Table 5.14 were mapped into their original four loyalty groups, as detailed in Tables 5.11 and 5.12. The results of the mapping is presented in Table 5.15. It can be seen, from Table 5.15, that in general, the overall

trends of the loyalty splits from the approach taken in [140] are similar to the results obtained for the proposed algorithm in this study. Indeed, there are some subtle differences in some itemsets, and this is expected as the approach detailed in [140] prunes the lowest 20% and highest 5% of transactions, based on size, thereby reducing the pool of loyal customers who typically have higher transaction sizes. It should be noted that both approaches prune at the lower transaction level, hence no immediate comparisons can be drawn in this regard.

However, unlike the algorithm proposed in this study, the approach detailed in [140] does not segregate customers within these clusters any further. Consequently, all customers within a cluster receive the same treatment. This approach is likely to drive down prices. Being able to attract new customers, while retaining revenue spend from loyal customers is a challenge for many retailers as noted in [43], [66], [72], and [109]. Consequently the approach detailed in [140] has limitations in this regard as it could lead to unnecessary price reductions due to loyal customers being offered discounts that the store does not need to offer. Given this, and the fact that the algorithm proposed in this study offers customised treatment through a two-step clustering process, it can be concluded that the algorithm proposed in this study performs better than the approach proposed in [140] in terms of Test 3, ability to offer customised treatment, and avoid targeting customers that are loyal. It should be noted that both models are equally good at attracting new customers, given the high volume of "switchers" in the cluster pool.

Comparing the models for Test 4, ability to enhance the frequency of the target itemset could not be done given that there is a large proportion of customers that are potentially "false positive" in the results obtained from the approach detailed in [140]. Given that both models demonstrate benefits in increased frequency of purchasing, it

	Store 9, (156,277) (2012)	Store 13, (213,163) (2012)	Store 9, (57,88) (2013)	Store 21, (78,209) (2013)
Target Customer Splits This Study				
Switchers	62%	62%	69%	82%
Drop-Out	3%	9%	2%	4%
Light Touch	27%	27%	22%	12%
Loyal	8%	2%	7%	2%
Target Customer Splits Reutterer et al. [140]				
Switchers	61%	57%	71%	84%
Drop-Out	4%	9%	3%	5%
Light Touch	28%	30%	22%	10%
Loyal	7%	4%	4%	1%

Table 5.15: Model Comparison Testing - Loyalty Splits

is noted that both models pass Test 4.

5.5.2 Comparing the proposed customer clustering model with "reality"

The mean number of purchases of each group for each itemset in Table 5.15 was computed for both the store in question, internal transactions, as well as at all others stores, external transactions. The percentage difference of mean purchases given by Equation (3.1) was calculated for each group, with the results shown in Figure 5.10. There are two key points to note from Figure 5.10, which underpins the validation of the customer clustering approach taken in this study. Firstly, the loyal group has a positive % difference for all four itemsets. Based on this, it can be concluded that loyal customers purchase more of the antecedent internally than externally. Similarly, the "switchers" purchase more externally than internally. This is in line with what one would intuitively expect from loyal and "switcher" customers, and is consistent with previous studies [101][141]. Secondly, the trend as one moves from loyal to "switcher",

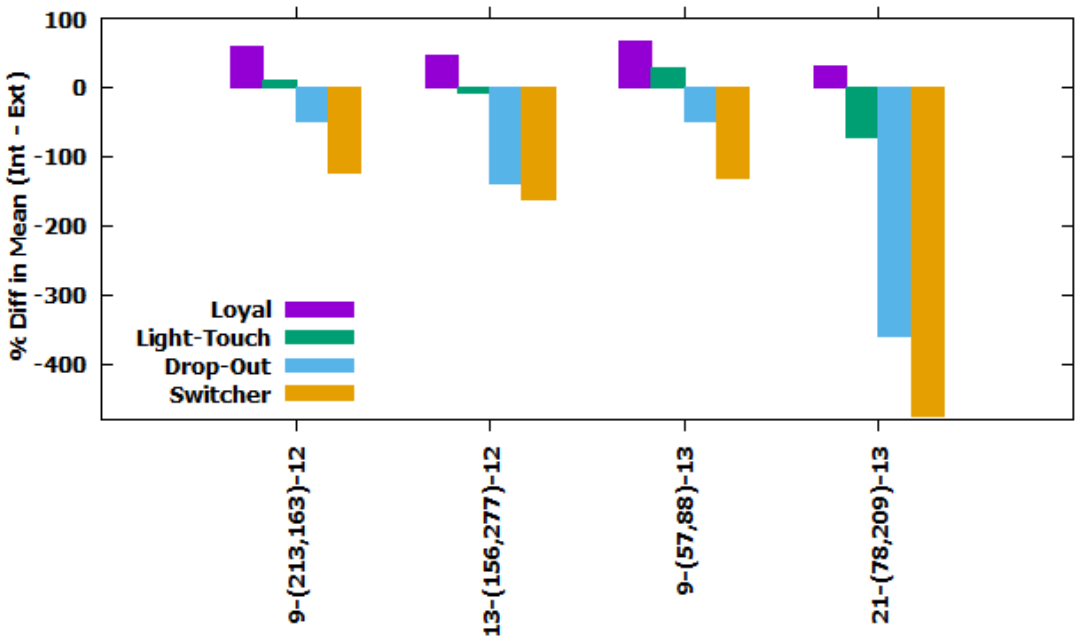


Figure 5.10: Percent mean difference between loyalty groups

i.e. descending order of loyalty, shows a consistent decrease in the % difference across all itemsets. Here again, this is consistent with intuition. This validates the clustering approach proposed in this study, as it conforms with "reality", i.e. what actually transpired. Given this, it can be concluded that the proposed clustering model, and the subsequent classification of customers into the four loyalty groups, which is a store-only clustering view, is an effective approach in segmenting customers for a differentiated marketing treatment plan.

5.5.3 The alternative approach - targeting customers who buy "top sellers"

The premise in this chapter thus far, as well as in Chapter 4, has been to target customers who purchase an antecedent that has a lower support than the consequent target item. The rationale behind this is that these antecedents will likely have a lower target customer base, as they have lower supports, and a higher probability to

purchase the consequent target item, as they have higher confidences. Hence, this approach will likely result in a lower marketing spend as fewer customers will need to be targeted to achieve the same result. The validation of this rationale is detailed as follows:

5.5.3.1 Using the marketing target model to select the optimal antecedent for targeting

As noted in Section 5.4.1, the itemset that has the lowest mt value is the optimal choice. This can also be used to justify whether stores should pursue a marketing target approach of $(A \rightarrow C)$ or $(C \rightarrow A)$. For example, the mt value for itemset (156,277) in Store 9, using the 2012 dataset and where item 156 is the antecedent is 0.402. The mt value for the same itemset is equal to 2.275 when item 277 is the antecedent. Clearly the mt value of the former is considerably lower, hence targeting customers who buy 156 with offers for itemset 277 is the optimal approach.

5.5.3.2 Targeting a sub-optimal antecedent can be costly

Whilst it is indeed acceptable to target customers of any of the items that make up an itemset, selecting a sub-optimal antecedent can prove to be costly. Again, using itemset (156,277) in Store 9, from the 2012 dataset as an example, the total number of customers buying item 277 at least once is 17314 compared with 7824 for item 156. From this, it is clear that targeting customers who buy 277 with offers for 156 may cost substantially more, given the larger number of target customers. One way of restricting the spend, which has been adopted in this study, is to target customers who have a history of purchasing both items at least once. In this regard, as can be seen from Table 5.11, that the total customers that will be targeted is 7472. By applying this constraint, it not only reduces the target customer base (and any associated costs), but it also increases the likelihood of positive customer outcomes,

which is always beneficial to the store. In this regard, customers will receive offers for items they have bought and likely will buy again in the future, and view the store's ability to tailor to their needs as a building block of a long and trusted relationship [120].

5.6 Tests with Other Datasets

Two synthetic datasets, a medium dataset, and a large dataset were created, using the approach detailed in Chapter 3 and in [83], with transaction volumes of 1 million and 5 million transactions respectively. The large dataset, had 100 different items with a maximum of 50 items per transaction, and an average density of 29 items per transaction. This was equivalent to a density of 29%, which is in essence seven times more dense than a typical grocery dataset. It should be noted that this dataset is similar in transaction volume, but considerably denser, to that of the UK's largest grocery retailer's daily activity, and represents approximately 28% of the UK's grocery retail market share [131].

Both synthetic datasets were processed using the algorithm detailed in Section 4.5, with the output file from the frequent itemset mining for both the medium and large dataset processed further. In all cases, the computational approach, and the applicability of the mt model was the same, that is, the best itemsets to target where those itemsets that had the lowest mt value.

In terms of computational speed, the large dataset did take longer to process with the runtime for the Apriori algorithm taking 748.8 seconds (approximately 12.5 minutes) to generate the list of frequent itemsets compared with 40.5 seconds for the 2012 scanner panel dataset. Tests were conducted using a personal computer that had

two 2.66 GHz Intel Xeon 5150 processors with 32GB of RAM. Similarly, the frequent itemset output file for the large dataset was larger than the 2012 scanner panel dataset which increased processing time, but yielded the same results, i.e. the best itemsets to target were those that had the lowest mt value. From the above, it can be seen that although the processing time is indeed increased when the density and volume of the dataset is increased, the processing time is still very manageable for batch processing, even with small, laboratory-size computer, as used in this study. Indeed, real-time processing by large grocery retailers will leverage considerably larger computers, and thus will significantly reduce the processing time.

5.7 Summary

The primary objective of this chapter was to validate the mathematical models and algorithm presented in Chapter 4, using the research methodology detailed in Chapter 3. This was achieved by conducting a series of experiments using real life data, synthetic data, and comparing the effectiveness of the model against other approaches. The experimental approach for testing combined the KDD, SDF and CRM models, as discussed in Chapter 3, to first identify target items, followed by identifying suitable target customers for these items. This was then followed by touting these customers with incentives for these items, and then finally retaining these customers for the long term. This approach is in line with the CRM model proposed in [120], and as detailed in Section 3.3. The key points of this chapter are summarised as follows:

- Tests conducted to validate the effectiveness of identifying target items, and indeed the proposed mt model, showed that the proposed approach obeys the principles as detailed in Section 1.3 and Chapter 3, and that the mt model is both an effective, and rapid way of deciding on the optimal itemset for targeted promotions.

- Tests on clustering customers using the proposed approach, showed that in general, the clustering approach successfully grouped like-minded customers, with the focus on eliminating both "false positives" and "false negatives", and effectively segregated loyalty groups to enable customised, targeted treatment to take place.

- Simulation testing conducted on the selected itemsets showed that it is possible to reach the targeted frequency with both the "conservative" and "aggressive" marketing campaigns. Indeed, "aggressive" campaigns do result in the targeted frequency being achieved sooner, albeit at a rate that is still practically possible.

- The detailed model comparison of the proposed model with the approach taken in [140] for targeted promotions showed that the proposed model outperformed on all fronts. It outperformed the Reutterer et al. model, detailed in [140], by accurately targeting customers, and offering differentiate treatment approaches, whilst minimising "false positives" and "false negatives".

- When comparing the model with "reality", the proposed approach showed that its customer clustering approach, and loyalty group splits behaved as one would expect, where "switchers" did indeed purchase more at other stores, and vice versa for loyal customers.

- The tests conducted where the proposed model was compared with an alternative approach of targeting "top sellers" showed that targeting customer who buy "top sellers" with offers for less-frequent items will likely cost the grocery retailers more, as it will require a larger target customer base. More importantly, targeting customers who buy "top sellers" would result in a lowering of prices across a broader customer base, and this could lead to revenue erosion in the long run.

- Testing the algorithm's processing speed, given that the Apriori algorithm has been noted to be slow on large, dense datasets, showed that whilst the processing time increased for larger, denser sets, it was still quite manageable, even with laboratory-scale computers. In this regard, tests were conducted on both the 2012 grocery dataset, and a considerably more dense synthetic datasets.

Overall, this chapter provided a sound validation of the research methodology presented in Chapter 3 and the models and algorithm proposed in Chapter 4, which formed the basis and unique contributions of this study.

Chapter 6

Other Applications

6.1 Introduction

The development, and application of the mt model to grocery retail has been detailed in Chapters 4 and 5 respectively. However, it was realised that the fundamental concept of the mt model, i.e. decision making between two choices ($A \rightarrow C$) and ($B \rightarrow D$), is not only an issue within grocery retail, but indeed likely to be a common problem in other fields as well. Given this, three applications in fields related to people's everyday lives were studied, with the aim of using the mt model to enhance the effectiveness of decision making by key stakeholders. It is believed that enhanced decision making will lead to enhanced outcomes, not only for the organisations involved, but indeed society at large.

The three applications were: (1) using the mt model to improve school attendance, (2) using the mt model to highlight the underlying relationships between Coronary Heart Disease (CHD) and the four metabolic syndrome conditions from large-scale public health data, and (3) using the mt model to evaluate the effectiveness of the use of the Stop and Search (S&S) power by police. On improving school attendance, a

detailed action-research based experiment was conducted, using a local UK primary school as a case study. The mt model was used to identify root causes for poor attendance and following this, a treatment approach was implemented with its impacts on pupil attendance evaluated. On the use of the mt model on CHD data, publicly available data from the British Heart Foundation (BHF) and the Institute for Health Metrics and Evaluation (IHME) were analysed using the mt model, with the results, and possible areas for the application of the mt model discussed with experts [124]. Similarly, on S&S, publicly available data from the Metropolitan Police was analysed using the mt model, with the results, and possible areas for the application of the mt model discussed with experts [26].

The chapter commences with a detailed account of using the mt model to improve school attendance, followed by the use of the mt model on CHD data, and then the work on S&S. The chapter concludes with a summary of the key points detailed in the various applications.

6.2 Improving school attendance - Action Research Experiment

6.2.1 Introduction

Given that pupil attendance has a strong, positive correlation with pupil attainment, pupil well-being and improved economic outcomes for pupils later in life, it is therefore not surprising that it remains a key focus for schools, local authorities and national governments across the world [14][35][81]. In the UK, both parents, which also includes guardians in this study, and schools have strict laws that must be followed in order to ensure school attendance is actively managed [57]. Parents are legally obliged

to send their children to school and ensure regular attendance, while schools have a legal duty to take the necessary steps, and have policies in place to effectively manage pupil attendance [57]. In this regard, there is a significant requirement from schools to be proactive on attendance management as they must: (1) accurately record attendance, (2) proactively follow-up with parents on all absences, and (3) put initiatives in place to manage and encourage good attendance [57].

The underlying reasons as to why pupils are absent from school have been well-studied and generally fall into one, or a combination of three categories: (1) unable to attend school due to other obligations, e.g. illness, carer duties, family instability, (2) avoiding school due to fear, embarrassment, boredom, e.g. being bullied, and (3) pupil/family do not place value in schooling, and/or have other activities that they would rather do, e.g. taking vacation, or have high levels of illiteracy within the family [14][143]. To this end, strategies for managing absenteeism, which are predominantly qualitative, have also been well-studied with models, frameworks and initiatives for improving school attendance being proposed and evaluated [14][35][81][143]. The quantitative approaches involving school attendance has primarily been seen as a task within the Educational Data Mining (EDM) branch of research, where the key objective is to improve pupil performance through the use of data mining and artificial intelligence (AI) techniques [158]. Indeed, there have been several models proposed to predict pupil outcomes, however attendance has typically been used as an input variable for these models rather than being a key focus area [39][158][178].

The case for increased use of data analytics and AI to improve attendance has been well-made, however very little use-cases and readily available analytics models exist, that can be easily adopted by school practitioners to improve school attendance [14][35]. Further, school practitioners are new to data analytics, with most not having

a data science background, and whilst there is ongoing training and certification being done, most schools are not yet in a position to conduct a deep analysis of data, let alone develop their own models and algorithms [3]. It is against this backdrop that this study aims to provide school practitioners with a simple, yet effective model to improve school attendance. This is achieved by identifying, and acting on attendance patterns that are not obvious without the use of data analytics. The mt model was applied, using an action research methodology, to a live setting using Willen Primary School (WPS) as a case study. WPS is a local authority-maintained primary school in Milton Keynes, UK. The approach, findings and recommendations from this study can be easily leveraged by other schools wanting to improve pupil attendance.

6.2.2 Literature Review related to Improving School Attendance

6.2.2.1 School Absence

There exists a myriad of terms used to describe school absence, which help focus diagnoses, so that targeted plans could be put in place to address their underlying causes [14][57][81][84]. While some absences may be seen as acceptable, in the UK, schools have become tough on all absences, irrespective of their reason, as they have been shown to be equally destructive to learning [14][57]. Authorised absence, defined as an acceptable absence approved by the school, e.g. illness or bereavement, is typically granted but, schools have become wary of its abuse, particularly close to end of term when parents want to capitalise on cheaper holidays without incurring fines [36][49][57]. On the other hand, unauthorised absence, i.e. absent without permission, has received widespread condemnation from both law makers, and education non-profit organisations, with several cases being trailed in court, or parents being fined in line with local authority and national government policy [36][49][57].

The concepts of school refusal (SR) and truancy form part of unauthorised absence and has been well-outlined in [81][84], with SR defined as non-attendance due to the expectance of strong negative emotions while at school, e.g. fear as a result of bullying, or embarrassment as a result of being teased, or separation anxiety [81]. Truancy has been defined as absences related to anti-schooling sentiments, i.e. absent without parental consent, and is a consequence of several factors including finding school boring, or finding activities outside of school more attractive, e.g. going to the cinema during school time [81]. School withdrawal, e.g. taking time off to go on holiday, is similar to truancy, but with parental consent, and is generally very difficult to address once it becomes excessive as it usually requires multi-agency involvement that focusses on both the pupil and their family as a whole [143]. The notion of persistent absence or chronic absence has also been well-studied, with the definition in the UK being: where a pupil is absent from school for 10% or more, irrespective of the reason [14][36]. Persistent absenteeism is being well-tracked by schools and local authorities in the UK with initiatives and policies put in place to deal with the problem as it arises [36][57]. However, the situation is not the same in other developed countries, including in the U.S., and is often overlooked by stakeholders [14]. As a result, persistent absenteeism unfortunately wreaks havoc, in these countries, long before the problem is diagnosed [14].

Separation anxiety, or in its more severe form, Separation Anxiety Disorder (SAD), is a type of school refusal and has been well-documented in [79]. SAD is common among young children, up to 1 in 20 children suffer from SAD, and is defined as the fear leaving the safety of parents or caregivers [79]. Children experiencing SAD often present with tantrums, panic attacks or bad behaviour and can have a significant negative impact on the child's academic, social, and physiological development

[29][79]. Indeed, separation anxiety is most common after children have spent longer spells with their parents or caregivers, and is common after weekends or holidays, and may also present every morning in some children, after they have spent the previous afternoon and night with parents or caregivers [29][79].

6.2.2.2 Why are pupils absent?

There is broad consensus by researchers as to why pupils do not attend school, and the underlying causes for absence falls into three categories, which are indeed very large by themselves: (1) unable to attend school due to other obligations, (2) avoiding school (school refusal), and (3) pupil/family do not place value in schooling, and/or have other activities that they could rather be doing [143][14] . This notion of not placing value in schooling has been further separated into: truancy, pupils that stay away from school without parental knowledge, and school withdrawal, also sometimes referred to as parental condoned absence, and is where parents condone the absence as it proves beneficial to them or the family at large [84]. Given the vast array of underlying causes, researchers have tended to become more specific in examining the problem of absence. Havik et al. in [81], focussed on school refusal and truancy, with peer relationships and classroom management by teachers as underlying causes. In this regard, Havik et al. found that both good peer relationships, and effective classroom management had strong positive correlations with good attendance [81]. Similarly, tackling truancy and parental beliefs, as part of school withdrawal, were the key focus areas of studies in [35] and [143] respectively. Both studies showed that there is a strong positive correlation with good attendance and effective, regular communication between school and home [35][143].

6.2.2.3 Impacts of absence

Balfanz et al. in [14], were firm in their conclusions that "missing school matters", noting that in the US, missing school impacted academic achievement irrespective of age, and that those who were from low-income backgrounds were more impacted by absence, as they were less likely to have provisions at home to make-up for lost time. In the UK, similar sentiments were echoed in [36] and [57] with respect to absence, including more long term impacts on the pupil such as social anxiety, and lack of self-confidence, both of which are known pre-cursors to interrupted employment and consequently lower economic attainment in adulthood [14][29][84]. Whilst these are all significant impacts in their own right, the key impact of absence, which was noted across several studies [14][35][36][57][143], was the long term disengagement with education which not only impacted the pupil in adulthood, but also created the foundation for a vicious cycle when these pupils become parents, and project their negative attitudes towards education onto their children.

6.2.2.4 Improving attendance

The conceptual framework proposed in [93] for designing interventions to improve attendance was found to be both relevant and very useful. The proposed three tier framework targeted all pupils along the absenteeism spectrum with tier 1 strategies focussed on pupils with emerging attendance problems, whilst tier 2 focussed on pupils that are at risk of being persistently absent, and tier 3 on those that are already persistently absent. The overall approach of the framework in [93] emphasized early identification and treatment, rather than a sole focus on those that are already persistently absent.

This approach is well-recognised and several studies have operationalised this framework in varying depths including in [14], [35], [36], and [57]. In [36] and [57], which is

relevant to the UK context, the guidelines have suggested that all absenteeism should be tackled with context-specific approaches that include using data analytics, working with parents, using incentives, and enforcing fines. Similarly, The Early Truancy Prevention Program (ETPP) introduced in [35] proposed a five step approach to tackle absenteeism, all of which required the teacher to be proactive, and work actively with parents to drive-up attendance. Pilot tests using the ETTP did show a significant improvement in attendance [35], however most initiatives were time intensive, and required teachers and school administrators to spend a large amount of time working with parents on an ongoing basis. This is not practical in the UK, given that teachers are already stretched, and where school budgets are being continuously squeezed [150]. Efforts to improve attendance in [14] and [35], were underpinned by offering both short-term and long term rewards through local and national/ state campaigns. At a local level, schools offered rewards for pupils who attended regularly that were more meaningful to pupils, and included fun activities like dance, and diplomas for completing short courses. While at a national level, school attendance was stressed by senior political figures, and "success mentors" who were largely celebrities that attributed their success to regular school attendance [35].

6.2.2.5 Educational Data Mining (EDM)

The definition of EDM in [3], accurately surmises the approach of using what was once commercial data mining techniques to improving outcomes in education, specifically government sponsored education. EDM, according to [3], "seeks to analyse educational data repositories to better understand learners and learning and to develop computational approaches that combine theory and data to transform practice to benefit learners". In this regard, EDM may be compared to commercial techniques like Market Basket Analysis (MBA), which is in essence, a technique that leverages data analytics on customer transaction data to enhance customer engagement and

transaction intensity within the retail sector [37][114][158].

Clustering and ARM have been widely used in EDM in a variety of contexts. Daniel, in [37], and Merceron et al. in [114], noted that ARM has been very useful in educational applications such as: (1) finding mistakes that are commonly made together by students, (2) making recommendations to students on e-learning course choices, and (3) finding associations in behavioural patterns of students. Similarly in [113], ARM was used to find factors that influenced student performance in courses, with the study concluding that student performance was directly correlated to paying attention in class which includes: attendance, completing assignments, and good note-taking.

Clustering has also been widely used in education with good success. In their review of clustering within EDM, Dutt et al. in [45], discussed the various educational contexts in which clustering was used including: using K-Means clustering to improve learning by grouping students with similar learning styles, and clustering brain scans of students who showed similar responses to learning into groups and targeting each group differently to improve learning. Similarly, clustering was also used to understand student behaviour in online learning environments by comparing sequential student data and leveraging a clustering algorithm to group like-minded students [98]. It should be noted that while clustering does have its place, it needs to be done carefully within the government schooling sector as it may be perceived by some parents as unfairly "targeting" groups of pupils, which is generally not the case [176].

6.2.3 Problem Statement for improving School Attendance

It is well-documented that providing pupils with the right incentives to attend school results in improved attendance and consequently improved pupil attainment and progress [14][57]. Given this, the problem being addressed by this study may be

stated as follows: Let S be a school with all its pupils, U and the school week, J, be divided into k distinct sessions, J_i, such that $J_i \in J = \{J_1, J_2, \ldots, J_k\}$. Further, let T be a database in S, that contains the attendance records of all pupils across all sessions for a period, W. Hence, there may exist a dataset, T_t, where $T_t \subseteq T$, that contains the attendance records of pupils U_t, where $U_t \subseteq U$, who have below the required attendance in at least one school session and/or the overall average attendance, but where attendance in all other sessions are above or equal to the requirement. Given that the leadership and staff of S are intent on maximising pupil attendance, with the focus on driving up the overall average pupil attendance, O, through incentives and interventions, while minimising effort and associated costs, largely incentives and staff costs, it becomes necessary to optimise the targeting of J_i. Thus, the aim of this investigation is to provide a framework and useful tool for schools, based on the mt model, for targeting the right school session(s) with incentives and interventions that maximises the impact on improved overall school attendance.

6.2.4 Analytical Model

6.2.4.1 Identifying Target Sessions

Pupils that have above or equal to the required attendance in each and every session are generally considered to have very good attendance and in essence help the school boost its overall average attendance. Let T_p be a dataset containing the attendance records of all pupils that are persistently absent, hence $T_p \subseteq T_t$. Persistent absenteeism in the UK is defined as having an overall average attendance of less than 90% [36]. Given that schools take severe action once attendance drops below 85%, including removing a pupil from the school roll, T_t thus represents a significant portion of T for a school that has overall-below-the-required-average attendance [57]. Hence, improving pupil attendance in T_t will enhance overall attendance and as most schools have limited resources, the question of which J_i should be targeted often arises. To

facilitate easy processing, T_t is converted to a dataset with binary attributes, with sessions and/or the overall average attendance being assigned a "1" when attendance drops below the required levels. Intuitively, the best target should be that session which has both the highest absence and highest association with poor overall average attendance, that is, $\text{supp}(J_i, O)$ and $\text{conf}(J_i \rightarrow O)$ is the largest. However, scenarios do exist where $\text{supp}(J_c, O) > \text{supp}(J_k, O)$ but $\text{conf}(J_c \rightarrow O) < \text{conf}(J_k \rightarrow O)$ in which case, the choice between J_k and J_c is not obvious. This problem may be easily addressed using the mt model, as was done for grocery retail in Chapter 5.

6.2.4.2 Applying the mt model to identify target sessions

Applying the mt model, as detailed in Chapter 4, to identify the best sessions to target was relatively straightforward. The mt value was computed for each $(J_i \rightarrow O)$ combination and that which had the lowest mt value was considered to be the best session to target. Note that the database T_t was in essence an absenteeism database, and $P(J_i)$ in database T_t was high for sessions that had high absenteeism. Similarly, $P(J_i, O)$ in T_t was high for sessions that had high absenteeism, and where such absenteeism leads to high overall absenteeism. Consequently, the "ideal" state in T_t was targeting sessions, J_i, that maximized overall absenteeism. The rationale for this was derived from the practical constraints of managing a school. Initiatives must target all or the majority of school pupils to ensure fairness, and given that most initiatives comprise of largely fixed costs, e.g. the effort in planning an activity is similar whether the audience is 100 or 250, it makes sense to target the session which impacts overall absenteeism the most [176]. The most impactful session was that session which has the lowest mt value as it requires the least "effort" to reach its "ideal" state.

6.2.5 Experimental Process

Experiments were conducted based on the well-known Action Research process as detailed in [34] and outlined in Figure 6.1. As per [34], the process begins by defining the context and purpose by asking the question: why is this project required or desirable? However, it is the diagnosing phase that usually proves to be the most challenging as it involves identifying the possible issues or the most impactful issue, which is sometimes not obvious [160]. Consequently, data analytics is often leveraged to simplify this task through the use of models and algorithms to process data into information [160]. In this regard, the mt model forms part of the diagnosing phase. Note that the Action Research process is cyclical and actions taken have to be regularly evaluated against the context and purpose, which could also change over time [34].

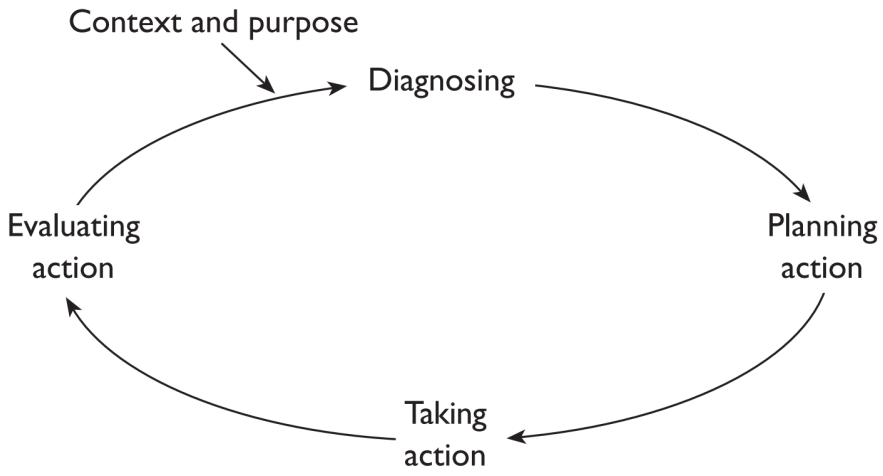


Figure 6.1: Action Research Process as outlined in [34]

Research was conducted at WPS with the context and purpose of improving overall school attendance to be above or in line with the national requirement, which is currently set at 96% in the UK [36][57]. The mt model, part of the diagnosing phase, was then used to identify that session which was most impactful to overall school absence. Options for possible action were brainstormed with school leadership and

evaluated in the planning action stage. Following this, selected actions were carried out at the school over several months, the "taking action" stage, with the impact on overall school attendance then assessed in the evaluating action stage.

6.2.6 Experimental conditions

6.2.6.1 Willen Primary School (WPS)

WPS is a mixed, 2-form entry primary school on the north-eastern side of Milton Keynes, catering for 4 to 11 year old children. The school has a capacity of 420 pupils and had 366 pupils on its roll at the end of July 2019, with approximately 35% of its pupils coming from outside the school's catchment area [148][176]. The school was rated "Good" by the UK's Office for Standards in Education, Children's Services and Skill (Ofsted) in its last inspection, which was conducted in November 2017 [125]. Whilst the inspector cited very good attendance management practices by the school leadership, he did note that further improvement should be made [125]. Given this, the school has continued to fervently promote the importance of good attendance and explored the use of novel approaches to address the issue of absenteeism, giving rise to this study [176].

6.2.6.2 Diagnosing

School attendance data for the previous three academic years (2015/16 to 2017/18) were used as the basis for improving school attendance in 2018/2019. The datasets were scrubbed to remove pupils that either joined the school after the start of the academic year, or left the school during the academic year, to produce the datasets T for each academic year, W, as discussed in Sections 6.2.4.1. The data was further filtered to produce T_t by selecting those pupils, U_t, who either had an overall attendance of less than 96%, the required national average, and/or who were absent at least three times per session during the academic year. Given that the cardinality of sessions

were generally between 34 and 39 per year, it was not practical to filter these sessions at the 96% level as it was too restrictive, i.e. equivalent to less than two absences per session per year. The restriction on analysing pupils that were present for the entire academic year was placed to ensure that the data analysis process was not unfairly skewed. For example, consider a scenario that occurs fairly regularly: some pupils may enrol at WPS after the start of the academic year, have 100% attendance for two weeks and then transfer to another school, possibly one that is closer to their home [176]. In this case, these pupils will have 100% attendance and be treated analytically as the same as pupils who had 100% attendance for the entire academic year.

T_t was then analysed using a FIM algorithm in R, with a $minsup$ of 0.3 and a $minconf$ of 0.3 to prune all rare rules. The output from the FIM stage was further analysed, using Microsoft Excel, to compute the mt value for each frequent itemset from which the best target session was identified.

6.2.6.3 Planning Action

Given that the school has strict obligations, guidelines, and its strategic agenda that it must adhere to, it was realized that a multi-prong approach had to be undertaken with regards to planned actions that would improve attendance, and validate the targeting approach proposed in this study. These planned actions were over and above what the school was currently doing to monitor and promote attendance. Hence a two-pronged approach was adopted with: (1) session-targeting focussed on demonstrating that session (and overall) attendance can be improved by targeting identified session(s), and (2) overall attendance improvement initiatives focussed on improving attendance in line with the strategic and statutory obligations of the school.

Several alternatives were considered by school leadership and based on their expe-

rience, the best two selected were: (1) focus on shorter periods for full attendance, with prize-based rewards, and (2) create more exciting initiatives for targeted sessions [176]. The selected initiatives were consistent with the tiered approach described in [93].

6.2.6.4 Taking Action

Apart from continuing to fulfil its statutory and strategic objectives with regards to attendance, including dealing with persistent absenteeism, promoting and fostering a good environment for improved attendance, and dealing with truancy; the school implemented the two initiatives outlined in Section 6.2.6.3.

Initiative One (I1) focussed on increasing the frequency and perceived meaningfulness of the rewards for full attendance, so that pupils could both feel tangibly rewarded for full attendance, and know that they can always be eligible for rewards in the next reward period should they not win in the current or previous period. I1 commenced at the start of the Spring term in January 2019, with all pupils that had full attendance for the month placed in a draw to win one of eight tickets to a popular, local trampoline park. The reward was meaningful to the pupils as it was something that they enjoyed, and it was something that was not always available to them due to cost constraints [176]. Given this, there was considerable excitement from pupils when the initiative was introduced. Initiative Two (I2) was geared towards targeting the sessions that had the largest impact on poor attendance. Exciting activities were conducted during the most impactful session throughout the Summer term starting at the end of April 2019. These activities centred on a theme, involved the entire school, was in line with the learning objectives, and included elements that the pupils would consider exciting [176]. Further details on I2 are provided in Section 6.2.8.2.

6.2.6.5 Evaluating Action

Following the implementation of the initiatives, the pupil attendance records for the 2018/19 academic year was analysed using Microsoft Excel, and compared with previous years to quantify the impact of I1 and I2. This then fed into school planning operations for the 2019/20 academic year.

6.2.7 Results and Discussion

6.2.7.1 Identifying Target Sessions

The average attendance for all pupils who were on the school roll for the entire academic year was calculated using Microsoft Excel, with the results presented in Table 6.1.

Session (J_i)	2015/16	2016/17	2017/18
Mon-AM	93.7%	93.7%	93.8%
Tues-AM	94.1%	94.9%	94.6%
Wed-AM	95.1%	95.0%	95.3%
Thur-AM	94.9%	95.3%	94.8%
Fri-AM	94.6%	94.7%	94.1%
Mon-PM	94.1%	94.5%	94.1%
Tues-PM	94.7%	95.6%	95.1%
Wed-PM	95.5%	95.6%	95.8%
Thur-PM	95.5%	95.7%	95.4%
Fri-PM	94.9%	94.9%	94.5%
Average AM	94.5%	94.8%	94.5%
Average PM	95.0%	95.3%	95.1%
Overall (O)	94.7%	95.0%	94.8%

Table 6.1: Average Attendance for 2015/16, 2016/17 and 2017/18

From Table 6.1 it can been that the school generally did not achieve the required overall average attendance of 96% in any of the previous three academic years. Further, attendance in the morning (AM) sessions were lower than the afternoon (PM)

sessions, with Monday AM being consistently the most poorly attended session across the years. This is consistent with theories on separation anxiety where young children often dislike going back to school after spending long periods away from school with their parents and family, and school withdrawal [29][36]. Separation anxiety may be exacerbated when parental collusion occurs (school withdrawal) and parents keep pupils at home for fear that they may become distressed further [14][36][84]. However, despite Monday AM being the most absent session based on averages, the question of which session is most impactful to overall attendance arises as it is possible that the most frequently absent session is not the most impactful, as children could return to school in the next session and have perfect attendance for the rest of the week. For example, consider this scenario of school withdrawal, which occurs fairly regularly [176], where: a child that is often absent on a Monday AM because they are dropped off at school by a parent that lives far away from the school (e.g. in a divorce), after spending the weekend at this parent's home. The child then returns to their regular home at the end of the day and has perfect attendance for the rest of the week. In this case, the child is unlikely to have overall attendance below the average but may be significantly below the required attendance for the Monday AM session.

6.2.7.2 Targeting the Most Impactful Session

The mt value for each session was calculated on each T_t for the previous three academic years, as per the process outlined in Section 6.2.3, with the results detailed in Tables 6.2, 6.3 and 6.4. Some sessions were automatically eliminated, consistent with Lemma 1 detailed in Chapter 4, as both their support and confidence were less than other sessions in the same year. In Table 6.2, Fri-AM had both higher support and confidence than every other session except Mon-AM, hence there was no need to compute the mt value for all other sessions except Fri-AM and Mon-AM. The mt model in Equation (4.15) was used to decide the better target session between Fri-AM

and Mon-AM, with a *minsup* value of 0.550 (the lower support between Fri-AM and Mon-AM) being used. Mon-AM had the lower mt value and hence was selected to be the best session to target.

Session (J_i)	$P(J_i)$	$P(J_i, O)$	Conf($J_i \rightarrow O$)	mt
Mon-AM	0.594	0.561	0.944	-0.021
Mon-PM	0.539	0.517	0.959	-
Tues-AM	0.494	0.472	0.955	-
Tues-PM	0.483	0.472	0.977	-
Wed-AM	0.456	0.433	0.951	-
Wed-PM	0.378	0.356	0.941	-
Thur-AM	0.494	0.467	0.944	-
Thur-PM	0.394	0.383	0.972	-
Fri-AM	0.550	0.539	0.980	0.020
Fri-PM	0.506	0.472	0.934	-
Overall (O)	0.856	-	-	-

Table 6.2: Identifying Target Sessions - 2017/2018

The negative value for mt was also interesting to note. In practical terms, it implied that there were more records in T_t that contained both Mon-AM and O that were below the required levels, than records that contained Fri-AM being below the required level. Thus any initiative to resolve absenteeism on Fri-AM will always be less impactful than absenteeism on Mon-AM. Hence all other sessions except Mon-AM were considered to be rare rules as there exists a (J_i, O) combination that is under consideration with $P(J_i, O) > minsup$. This is not always the case and the scenarios were quite different for the 2015/16 and 2016/17 academic years.

From Table 6.3 all rules except Mon-AM, Mon-PM and Fri-PM were shortlisted as the others were determined to be rare. The mt values were computed for each of the shortlisted sessions, with *minsup* set at 0.549, the lowest support between Fri-PM, Mon-AM and Mon-PM. Mon-AM was found to be the most impactful session to

Session (J_i)	$P(J_i)$	$P(J_i, O)$	Conf($J_i \rightarrow O$)	mt
Mon-AM	0.609	0.549	0.903	0
Mon-PM	0.559	0.505	0.904	0.089
Tues-AM	0.505	0.480	0.951	-
Tues-PM	0.436	0.417	0.955	-
Wed-AM	0.490	0.480	0.980	-
Wed-PM	0.456	0.446	0.978	-
Thur-AM	0.480	0.480	1	-
Thur-PM	0.436	0.431	0.989	-
Fri-AM	0.539	0.510	0.945	-
Fri-PM	0.549	0.515	0.938	0.066
Overall (O)	0.848	-	-	-

Table 6.3: Identifying Target Sessions - 2016/2017

overall below average attendance. Similarly from Table 6.4, Mon-AM was found to be the most impactful session with *minsup* set at 0.589. Given that Monday AM was found to be the most frequent and the most impactful session, it can be concluded that the poor attendance on Monday AM may be attributed to a combination of school refusal, e.g. due to separation anxiety, and school withdrawal/truancy where the return to school may not be seen as exciting as the weekend that just passed [14]. Clearly an easier, more exciting start to the school week, which is initiated by the school, may prove successful in addressing this issue.

Session (J_i)	$P(J_i)$	$P(J_i, O)$	Conf($J_i \rightarrow O$)	mt
Mon-AM	0.624	0.584	0.935	0.009
Mon-PM	0.589	0.558	0.948	0.056
Tues-AM	0.609	0.563	0.925	0.048
Tues-PM	0.569	0.548	0.964	-
Wed-AM	0.492	0.482	0.979	-
Wed-PM	0.467	0.457	0.978	-
Thur-AM	0.558	0.543	0.973	-
Thur-PM	0.487	0.487	1	-
Fri-AM	0.508	0.487	0.960	-
Fri-PM	0.503	0.472	0.939	-
Overall (O)	0.857	-	-	-

Table 6.4: Identifying Target Sessions - 2015/2016

6.2.7.3 Early Warning System

The very high confidence values (>0.9 and in some cases = 1) was also of significant note as it suggested that any pupil that was absent for at least three times in any one session was very likely to have below overall required attendance. This could be a good tool for the school to use in tackling absenteeism as it may be used to identify pupils that are at risk of falling below the requirement, consistent with the recommendation in [93]. Further, it could be used as part of conversations with parents and pupils in addressing their beliefs and misconceptions about attendance, and is consistent with the recommendations in [14], [35], and [57] for improving attendance through leveraging analytics. This fact-based approach is more likely to resonate well with parents and may negate any possible insinuations by parents that their families are being victimised or treated unfairly by teachers and school leadership [14][35].

6.2.8 Evaluating the Impacts of Initiatives I1 and I2

I1 and I2 were conducted as detailed in Section 6.2.6.3. Following the results of the analysis conducted as part of Section 6.2.7.1, the school decided to target Mondays with the emphasis on the Monday AM session as part of I2. The Monday Matters initiative was launched in the Summer term of 2019 and consisted of a "m-themed" program for five of the ten Mondays during the term. The initiatives were selected by the school staff as it represented themes that would resonate well with the pupils. The five themed Mondays were: Move-It Monday, Muffin Monday, Mindfulness Monday, Mask Monday and Movie Monday. For each themed Monday, pupils were allowed to come to school appropriately dressed, for example sports kits on Move-It Monday, and participate in a range of planned activities related to that theme which were also linked to the work that was being done in the classroom.

6.2.8.1 I1: Frequent Rewards for Full Attendance

Draws were held every month during the Spring and Summer terms of 2018/19, with the exception of April, for all pupils that had full attendance during the month. The April draw was omitted given that April had fewer than 10 school days in that month.

Session (J_i)	2015/16	2016/17	2017/18	2018/19
Mon-AM	94.1%	93.9%	94.3%	95.8%
Tues-AM	94.2%	95.2%	95.0%	96.2%
Wed-AM	95.1%	95.5%	95.4%	96.2%
Thur-AM	94.9%	95.5%	95.1%	96.3%
Fri-AM	94.4%	95.0%	94.2%	95.6%
Mon-PM	94.7%	94.6%	94.8%	96.5%
Tues-PM	94.9%	95.8%	95.5%	96.8%
Wed-PM	95.6%	95.8%	95.8%	96.6%
Thur-PM	95.6%	95.8%	95.6%	96.8%
Fri-PM	94.9%	95.1%	94.1%	96.0%
Average AM	94.6%	95.1%	94.8%	96.0%
Average PM	95.2%	95.4%	95.2%	96.5%
Overall (O)	94.9%	95.2%	95.0%	96.2%

Table 6.5: Average Attendance for Spring and Summer Terms: 2015/16, 2016/17, 2017/18 and 2018/19

From Table 6.5, it is clearly evident that the shorter, more meaningful rewards for full attendance have contributed to a significant improvement in overall attendance for the Spring and Summer terms in 2018/19 with the attendance for every session being considerably higher than the attendance in the previous three years. This result was consistent with the findings in [14]

6.2.8.2 I2: Monday Matters Initiative

Table 6.6 presents attendance data for Summer term Monday attendance for the 2015/16, 2016/17, 2017/18, and 2018/19 academic years. There is fluctuation in the number of Mondays from year to year due to the timing of Easter which influences

the half-term break as well, which is typically held towards the end of May.

	2015/16	2016/17	2017/18	2018/19
Average Attendance (%)	94.4%	94.5%	94.2%	96.5%
Range	6.2%	5.3%	4.5%	2.8%
Median	94.5%	95.0%	94.1%	96.4%
No. of Mondays	12	11	12	10

Table 6.6: Comparison of Monday Summer term attendance data for I2

From Table 6.6 it can be seen that the average attendance for Mondays in the Summer term of 2018/19 was significantly higher than the previous years. Further, not only was the 2018/19 attendance data higher, but it was also above the required 96% target and the first time that this was the case in four years. The range and median for the data also showed the strength of 2018/19 attendance data when compared to previous years. The range in 2018/19 was less than half that of 2015/16, and indicative of a consistently high Monday attendance throughout the term.

There were some concerns from school leadership on the "stickiness" of Monday Matters events, where having an event every other Monday fosters good attendance on other Mondays and indeed other days of the week, and whilst there were spikes in attendance on Monday Matters days, attendance during the other Mondays were quite good, as evidenced by the data in Table 6.6. These findings are consistent with other studies that noted that in general, pupils are "creatures of habit" who thrive on routine, and are thus likely to sustain good attendance once a routine has been established [14][35][57].

6.2.9 Evaluating the Overall Improvement in School Attendance

The full year attendance comparison is presented in Table 6.7. It can be seen that initiatives in the Spring and Summer terms of the 2018/19 have contributed to an improvement in the whole school attendance for the full academic year. Indeed, WPS achieved the required attendance target of 96% for the first time in four years in 2018/19.

The data in Table 6.7 also reveals the success of the Monday Matters initiative on the full year attendance data. Monday AM and PM sessions have seen the largest increase in attendance, with increases of 1.8 and 2.1 percentage points improvement respectively. As a result, Mondays no longer have the worst performing AM and PM sessions, and the shift in focus now moves towards Fridays, where the underlying reasons for poor attendance may be quite different. Unlike Monday absenteeism, which is influenced to some extent by separation anxiety, Friday absenteeism may be more influenced by school withdrawal where parents may: (1) want to extend the weekend or start holidays earlier to beat the rush and/or save on costs, and (2) sometimes assume that Fridays are typically low value school days in which limited learning takes place, and hence pursue other activities outside school [49][84]. Hence, the action plan to tackle Friday absenteeism must be geared more towards school withdrawal as opposed to the Monday Matters initiative which was focussed on tackling both school refusal and school withdrawal.

One argument that parents do make on Friday absence is that their child(ren) have excellent attendance on all other sessions and these occasional absences should not impact the child and the school. Whilst it is well-documented that all and every absence impacts pupil learning, the question of whether Friday sessions have now

Session (J_i)	2015/16	2016/17	2017/18	2018/19
Mon-AM	93.7%	93.7%	93.8%	95.6%
Tues-AM	94.1%	94.9%	94.6%	96.0%
Wed-AM	95.1%	95.0%	95.3%	96.1%
Thur-AM	94.9%	95.3%	94.8%	96.1%
Fri-AM	94.6%	94.7%	94.1%	95.3%
Mon-PM	94.1%	94.5%	94.1%	96.2%
Tues-PM	94.7%	95.6%	95.1%	96.6%
Wed-PM	95.5%	95.6%	95.8%	96.4%
Thur-PM	95.5%	95.7%	95.4%	96.4%
Fri-PM	94.9%	94.9%	94.5%	95.4%
Average AM	94.0%	94.8%	94.5%	95.8%
Average PM	94.5%	95.3%	95.1%	96.2%
Overall (O)	94.2%	95.0%	94.8%	96.0%

Table 6.7: Average Attendance for 2015/16, 2016/17, 2017/18 and 2018/19

become the most impactful session to overall absence arose [14][57][84]. In line with this, the analysis detailed in Sections 6.2.6.2 and 6.2.7.2 was conducted on the 2018/19 dataset. It was clear from the results in Table 6.8 that Friday is now the most impactful day to overall below the required attendance, with Fri-AM being the most impactful session. Mon-AM is no longer the most impactful session to overall below average attendance for the first time in the four academic years.

Session (J_i)	$P(J_i)$	$P(J_i, O)$	Conf($J_i \rightarrow O$)	mt
Mon-AM	0.500	0.467	0.933	-
Mon-PM	0.425	0.392	0.922	-
Tues-AM	0.483	0.467	0.966	-
Tues-PM	0.350	0.333	0.952	-
Wed-AM	0.508	0.492	0.967	-
Wed-PM	0.425	0.408	0.961	-
Thur-AM	0.458	0.433	0.945	-
Thur-PM	0.433	0.408	0.942	-
Fri-AM	0.575	0.525	0.913	0.050
Fri-PM	0.550	0.500	0.909	0.100
Overall (O)	0.867	-	-	-

Table 6.8: Identifying Target Sessions - 2018/19

6.2.9.1 Persistent Absenteeism

The impacts of initiatives I1 and I2 on persistent absenteeism (attendance <90%) were also analysed with the results presented in Table 6.9. Persistent absenteeism at WPS has been significantly higher than the national average for at least the last three years, this despite regular and close monitoring by the school's leadership team (including the school governors) and the school's attendance officer. However, the level of persistent absenteeism has significantly decreased in 2018/19, and was lower than the latest published national average for persistent absenteeism of 8.7%. Note that national average data for the 2018/19 academic year is expected to be published in March 2020, in line with previous years.

	2015/16	2016/17	2017/18	2018/19
WPS (% of total)	13.1%	11.3%	12.8%	5.8%
National (% of total)	8.2%	8.3%	8.7%	-

Table 6.9: Comparison of Persistent Absenteeism

This is a significant improvement and consistent with previous studies that sought to tackle the problem of chronic (persistent) absenteeism, in particular [14]. Indeed some of the approaches for tackling persistent absenteeism discussed in [14] have been leveraged in the development of I1 and I2 including the concept of making rewards more frequent and meaningful.

6.2.10 Summary Remarks for Improving School Attendance

The mt model, described in Equation (4.15), was used to improve school attendance at Willen Primary School. The algorithm detailed in Section 4.5, which included the mt model, was used to identify the school session which was most impactful to overall below the required average attendance. In line this, the previous three years' atten-

dance data from WPS was analysed and it was found that the Monday AM session was consistently the most impactful session to the overall below the required average attendance. Two initiatives were carried out at WPS and leveraged approaches in previous studies and the collective wisdom of WPS leadership and staff [14][35][176]. Initiative I1 provided more frequent and meaningful rewards for full attendance over a shorter time frame, whilst I2 focussed on improving Monday attendance through the use of themes that were known to be exciting for the pupils.

Both I1 and I2 resulted in a significant improvement of attendance at WPS, with attendance in 2018/19 being at its highest over the past four academic years. Overall average attendance for the 2018/19 academic year was at the required target of 96%, whilst the combined Spring and Summer term attendance was higher at 96.2%. Monday attendance during the Summer term also improved significantly from an average and range perspective. The average Summer term Monday attendance in 2018/19 was significantly higher than the three previous years at 96.5%, while its range was significantly lower 2.8%, implying that attendance on Mondays were consistently better throughout the term.

Analysis of the 2018/19 data using the mt model has revealed that Monday AM is no longer the most impactful session to overall below the required attendance, instead it is now Friday AM. The underlying dynamics as to why this is the case may also include a shift away from school refusal and more towards school withdrawal (parental condoned absence) which is underpinned by a variety of reasons including cheaper holidays [49][176]. Addressing this is considered to be part of the future work and is detailed in Chapter 7.

6.3 Using the mt model on Medical Data

6.3.1 Overview

Making sense of large-scale public health data is often a minefield for doctors, let alone the public. Indeed, the head of Public Health, England noted in [48], that good data can often be lost when stakeholders pursue different agendas. McCartney, in [112], criticised the role of the media in misrepresenting medical studies and/or for extrapolating the results of small-scale studies too widely to create hype and increased readership in their quest for profits. The recommendation in both [48] and [112] called for greater ethicality in medical research to create value and instil ongoing public trust. In this regard, the key conclusion drawn from [48] on achieving value from public health data analytics was considered apt: "to find truths from complexity of data, present it in an unambiguous way and explicitly counteract those perceptions which are not based on the facts."

At a global level, prioritising targets for public health campaigns is a very complex task as it not only straddles sectors within healthcare, but all other areas of governments including the economy, welfare, and defence [11]. Whilst there are several criteria for prioritising health care campaigns, especially as resources are scarce, a central theme across these criteria is that priority must be given to areas that will result in maximum public gain [11]. Defining the concept of maximum gain is in itself a complex undertaking, however one universally accepted measure is indeed minimising mortality, i.e. priority should be given to health issues that have the maximum impact on mortality [11].

With this in mind, the mt model was applied to public health data related to Coronary Heart Disease (CHD), and its relationship with metabolic conditions, with two

aims: (1) to support the notion of finding truth in complex data and presenting it in an unambiguous way, and (2) to establish whether the mt model can aid in setting public healthcare priorities. Metabolic conditions are four medical conditions namely: (1) high blood pressure (HBP), (2) high body mass index (BMI), (3) high cholesterol, and (4) diabetes [58]. These four conditions are considered to be primarily responsible for CHD in patients that do not have a congenital heart defect, and are part of a larger set of conditions, known as modifiable risk factors, that include physical inactivity, smoking, and alcohol consumption [58][63][173]. Given that they are modifiable, it implies that patients can often eradicate these conditions through lifestyle changes, hence they are also considered to be preventable conditions [63].

Why focus on CHD and metabolic conditions? The three reasons for this is as follows: (1) cardio vascular disease, including CHD which is the largest contributor, remains the top cause of premature death in the UK, and indeed globally [17][88], (2) there is an abundance of detailed data on both the prevalence, people living with CHD, and the mortality rates of CHD, thus naturally lending itself to data analytical studies [173], and (3) there is often misrepresentation of the impacts of these metabolic conditions and their treatment in the media, which creates mass-scale confusion amongst the population, who may not be well-equipped to make informed choices [112]. For example, the simple advice on egg consumption is contradictory, not only across well-respected information sites like the National Health Service and British Heart Foundation, but also within these sites as well, with one article highlighting its good for you, while another highlighting that it is bad for you [61][62][149].

Given this, the mt model was applied to CHD prevalence and mortality data for the UK, to better understand the role and impact of each metabolic condition on overall CHD-related deaths. The findings of this analysis were then tested with the

University of Leicester Cardiology Research Team [124].

6.3.2 Data Sources and Analysis Techniques

Data for CHD prevalence and mortality for 2017 was obtained from [58] and [63]. The data was then analysed using Microsoft Excel. The underlying premise of the analysis was that in general, the population as a whole within that age group is considered healthier, in terms of any of the four metabolic conditions, than the people that died as a result of one or a combination of the conditions. For example, the proportion of people that died at age 35 of CHD as a result of a high BMI, or where high BMI was one of the attributable causes, should be higher than the general population at age 35. If the population as a whole is taken as the "ideal state" and the people that died as the "undesirable" or current state, then the mt value can be used as a measure for the deviation from "ideal" for each condition within each age group. Thus, by comparing mt values, the "true" impact of each condition on the overall CHD mortality within that age group can be assessed.

The specific mt parameters are thus as follows: $P(A)$ will be the percent prevalence of the condition within the age group, *minsup* will also be the percent prevalence, as the compromise state is indeed the "ideal state" in this scenario, and $P(A, C)$ will be the percent of people that died of CHD with that condition. Consequently, the higher the mt value, the more severe the condition is in terms of CHD death. Similarly, a negative mt value will imply that people that die of CHD with that condition present may likely have died as a result of a direct impact by another cause, or that in general that metabolic condition within that age group, is not impactful.

6.3.3 Results and Discussion

The results of the data analysis conducted on the datasets described in Section 6.3.2 is detailed. The results were presented to the University of Leicester Cardiology Research Team and their comments are included as part of the discussion [124].

6.3.3.1 Deaths by CHD

From Figure 6.2, it can be seen that approximately 90% of all CHD deaths occur in people aged over 60 years, with 70% occurring over the age of 75 years. Although CHD deaths remains largely an older person issue, the contributory effect of CHD to the overall mortality in the over 75 age group has been reduced. In 2017 CHD contributed to 11% of all deaths in the over 75 age group in 2017 compared to 13% in 2012 [63]. Several reasons have been put forward for this, including improved treatment, and advances in medical interventions [124].

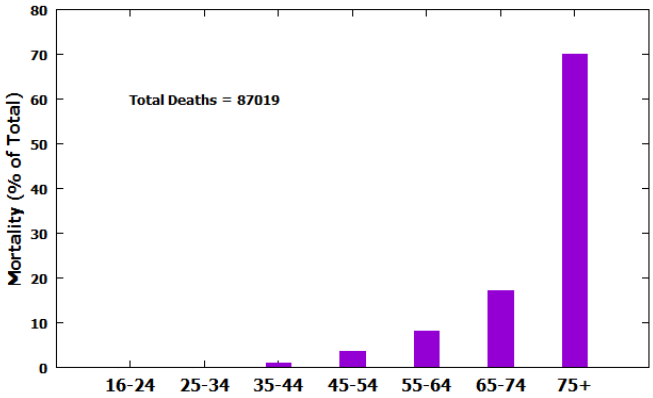


Figure 6.2: 2017 UK CHD Deaths by Age group

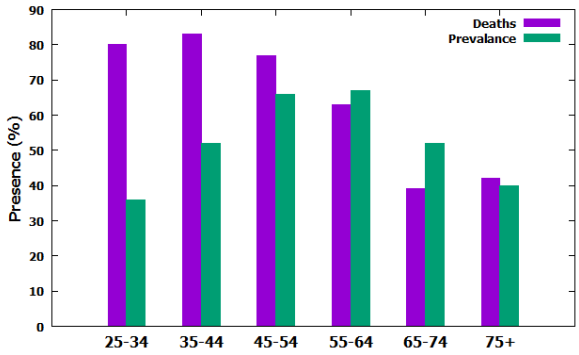
6.3.3.2 Prevalence versus Deaths

From Figure 6.3 it can be seen that the four metabolic conditions across the six age groups results in twenty four potential initiatives for CHD alone. Deciding on whether to prioritise one of the four conditions, or a selection of the twenty four potential initiatives in itself is a daunting task, let alone deciding across the healthcare spectrum

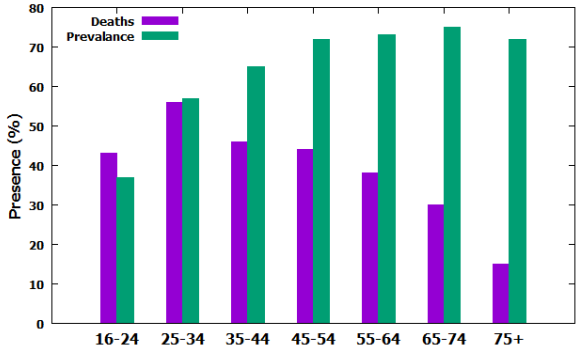
that includes other major diseases like cancer, mental health or liver disease [124]. For example within CHD, is a cholesterol lowering campaign in under 45s more important than a diabetes campaign in the over 55s? Inspecting the graphs in Figure 6.3, it is evident that for some conditions, e.g. BMI, which is a proxy for obesity, the prevalence of having a high BMI is larger than in those that died from CHD. In this regard, it could suggest that a high BMI is actually protective against CHD mortality, which is contradictory to the multitude of public health campaigns that profess a positive correlation between BMI and CHD [88]. Conversely, the prevalence of diabetes is significantly lower, in most cases less than half, than in those that die from CHD. It is clear that making such choices are not straightforward, even when looking at it from a purely data analytics perspective.

The mt values for each of the twenty four combinations, as shown in Table 6.10, were calculated using the values for the mt variables discussed in Section 6.3.2. As noted in Section 6.3.2, the combination with the highest, positive mt value is the most serious, as it shows a significant difference between those that die from CHD and the population at large. From Table 6.10, it can be seen that diabetes is a significant risk factor across all age groups, while BMI appears to be inversely correlated to CHD mortality for all age groups, especially in the over 75s. Similarly, high blood pressure and cholesterol appears to be an issue with younger age groups, but becomes less of an issue amongst the older age groups.

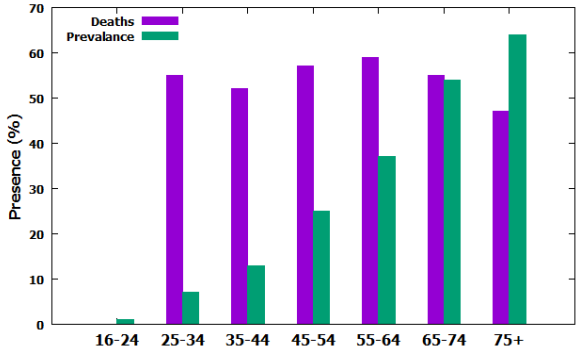
These results were discussed with the University of Leicester Cardiology Research Team, and it was noted that the BMI results correlate with the obesity paradox principle, where “fatter” people are more likely to survive CHD events than their “thinner” counterparts [102][124]. Similarly, the diabetes results also correlate with current observations, and indeed it is now becoming one of the top causes of premature



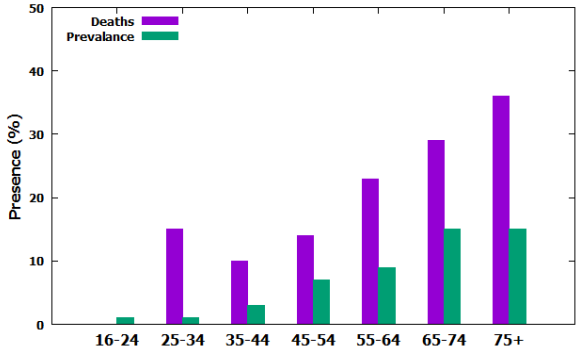
(a) Deaths vs. Prevalence - Cholesterol



(b) Deaths vs. Prevalence - BMI



(c) Deaths vs. Prevalence - HBP



(d) Deaths vs. Prevalence - Diabetes

Figure 6.3: Death versus Prevalence for the four Metabolic Risk Factors for CHD

deaths globally, not only as part of its contribution to CHD, but to other illnesses as well including cancer, organ failure, and circulatory diseases [17].

Age Group	Cholesterol	BMI	HBP	Diabetes
25-34	0.55	-0.02	0.87	0.96
35-44	0.38	-0.42	0.75	0.72
45-54	0.15	-0.65	0.56	0.50
55-64	-0.06	-0.95	0.38	0.62
65-74	-0.34	-1.53	0.01	0.49
75+	0.03	-3.78	-0.36	0.58

Table 6.10: mt values for the Metabolic Risk Factors by Age Group

The mt model does provide a quantitative mechanism to prioritise initiatives. For example, from Table 6.10, reducing the prevalence of diabetes and high blood pressure should be the focus across all age groups. This is particularly important for the under 34s where diabetes has the highest mt value of 0.96, and high blood pressure, the second highest value of 0.87.

6.3.4 Summary Remarks on the use of the mt model in healthcare

The mt model was used to support the prioritisation of public health initiatives, and to bring clarity to the data with the aim of presenting the findings in an unambiguous way. Experiments conducted on the UK CHD for 2017 showed that in general diabetes and high blood pressure should be prioritised for public health campaigns. This correlated well with current global mortality rates and trends where cardiovascular disease, including CHD, remains the number one killer globally, with diabetes and high blood pressure being key contributors [17]. The results also showed that contrary to popular belief, people with high BMIs are not necessarily at an increased risk for CHD mortality. Indeed, this is a well-recognised fact amongst the cardiovascular research community, with the obesity paradox, people with higher BMIs having

better CHD mortality outcomes than their lower BMI counterparts, being an active area of research [102][124].

An important limitation in the interpretation of the results is the role of confounding factors, including the impacts of other diseases that people may suffer from concurrently. For example, there is an argument that people aged 75+ may have lived for several years with CHD, and that they would have died earlier had their BMI been higher [124]. Hence 75+ CHD mortality has a high negative correlation with BMI [124]. Similar arguments may be used for high blood pressure and cholesterol. However, it should be noted that the mt model normalises these arguments in that it compares the age groups to their peers, based on prevalence. Thus it could be asked: if low BMI prevented the earlier onset of CHD mortality, and that high BMI generally enables other confounding conditions, then how is it possible that the general population of 75+ have higher BMIs? Surely, a larger proportion of these people should have died much earlier as a result of high BMI-related confounding conditions?

Clearly there is further investigation required to fully address these questions. However, it is evident that this analysis demonstrates the power of the mt model, in that it can be easily applied to global data and through this high-level, data driven approach, it can uncover subtleties, like the obesity paradox, and create a platform for further discussion. In this regard, the conclusion of University of Leicester Cardiology Research Team are apt: “the mt model and conclusions from its analysis could be used to quickly uncover some associations, and generate hypotheses for further studies” [124].

6.4 Use of the mt model in Crime Prevention

6.4.1 Overview

The power of police to stop citizens, demand an account of their movements and actions and the right to search their possessions is ubiquitous around the globe [27][74][163]. In England and Wales, this power is known as Stop and Search (S&S) and is underpinned by various pieces of legislation notably Section 1 of the Police and Crime Evident Act, 1984, commonly referred to as PACE, and Section 60 of Criminal Justice and Public Order Act, 1994 [27][111]. Unlike PACE, which requires the police to search a person under "reasonable suspicion", Section 60 does not have any restrictions and police can search people without justification [27][111]. Given this, S&S is highly charged both politically and socially. From a political perspective, governments want to portray a "zero tolerance" approach to crime prevention to sustain and garner further support from the electorate, while from a social perspective, community organisations want to ensure that people are free from undue discrimination and intimidation, which contributes towards the ongoing social decay [27][74][111]. There is also considerable academic research on S&S, however most of it, after some consideration of the pros and cons of S&S, conclude that it is negative, and its impact on crime reduction as a whole is low [27][111].

Irrespective of the positions that political and law enforcement leaders, and academic take, they all believe that minimising "false positives", that is searching people who are innocent, and "false negatives", missing a search on a prospective criminal are both damaging [26]. However, from a practical standpoint, "false positives" are considerably more damaging than "false negatives" from all sides [26]. Law enforcement see "false positives" as a waste of manpower. More importantly, "false positives" open the government to huge costs through potential litigation and to ongoing disharmony

between themselves and the communities that they serve [26]. On the other hand, communities also view "false positives" as extremely damaging as it unjustly labels people as criminals, demoralises them based on a negative perception of their appearance, smacks of discrimination based on race or religion, and sustains a sense of distrust of government [27][74][111].

The rise of knife crime in the UK in recent years and the recent change of government, in July 2019, has seen the spotlight being once again cast on more proactive policing of which S&S remains a key element [16]. Given this, the objective of this task was to evaluate the use of the mt model in S&S with the aim of helping enforcement agencies enhance the effectiveness of S&S. The findings of this analysis was tested with a leading UK academic who specialises in S&S [26].

6.4.2 Data Sources and Analysis Techniques

S&S data from the Metropolitan Police for the month of December 2018 was obtained from [38] and used as the basis of this analysis. There were a total of 15522 recorded S&S by the Metropolitan Police across eight offending categories including offensive weapons, stolen goods, and drugs. Other data categories included race, age group, and outcome of S&S. The outcome of S&S included arrest, community resolution, penalty notices and no further action, with the key efficiency measure being the number of arrests, or arrest rate percent [38].

The data was analysed using the proposed mt model and algorithm, outlined in Section 4.5, with the combination of race, age group and offending category being the antecedent, while outcome being the consequent item.

6.4.3 Results and Discussion

The results of the analysis of the S&S data as outlined in Section 6.4.2 is presented and discussed in detail. The discussion commences by a broad-based evaluation of the data in the context of widely-held beliefs and past trends. This then progresses into a discussion on S&S efficiency, and the potential scope of the mt model in this regard.

6.4.3.1 Broad-based Evaluation

The distribution of the outcomes of S&S for the December 2018 dataset is presented in Figure 6.4. From Figure 6.4, it is clear that arrests make-up a small proportion of the overall outcome, with most people that are stopped and searched being let go without any further action. While these figures are consistent with what the Metropolitan Police advertise on their website, a 16% efficiency rate by most standards is very low, while a "false positive" rate of at least 72% is very high [26][132].

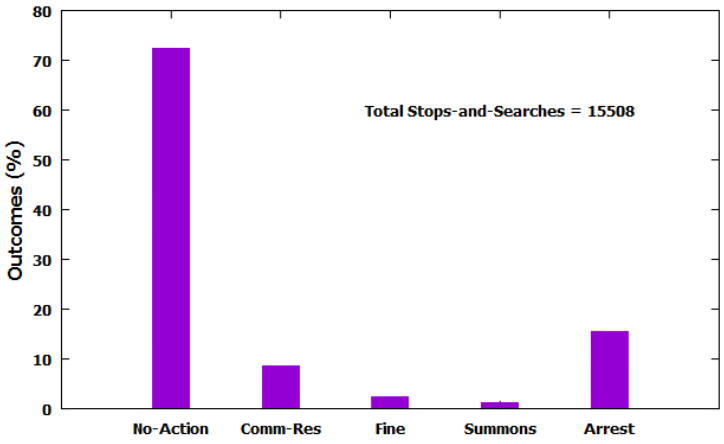


Figure 6.4: S&S Outcomes - December 2018, Met Police Data

In terms of racial demographics, the data presented in Figure 6.5 compared S&S data with the demographic data of greater London, which is in essence the area served by the Metropolitan Police [132]. From Figure 6.5, it can be seen that in general

a disproportionately large number of black people are stopped and searched, while white people, in general, are less-likely to be stopped and searched. The proportions for asians and other groups were broadly in line with demographics. These are not new findings and has been a point of contention for several years [26]. In this regard, community groups, academics and the media have noted that this disparity may be construed as racism, or at the very least unconscious bias [27][74]. Indeed, this problem is not confined to London, but has been observed in other major cities globally, including in New York and Los Angeles [163]. This is compounded by the fact that most S&S activity are "false positive" as shown in Figure 6.4. As a result, it is not surprising that S&S is perceived very negatively by black communities, who in general have very low levels of trust for the impartiality of police, and their crime prevention intentions [27][74].

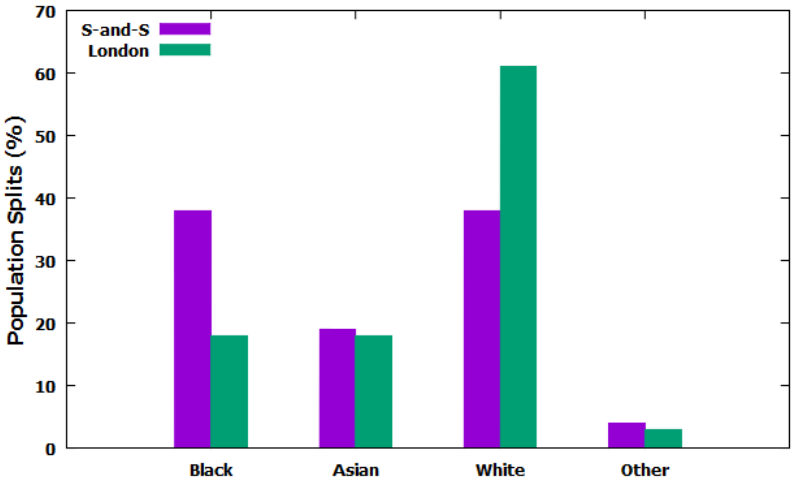


Figure 6.5: Racial Demographics - December 2018, Met Police Data

The majority of S&S is done with the police suspecting drug-related activity as shown in Figure 6.6. This is not surprising given the high rates of drug use in the UK today [16][132]. While this may be the case, critics of S&S suggest that a suspicion of a drug-related offence is usually easier for police to justify as it could be the police saying: "I saw smoke or smelled cannabis", and is sometimes used as an entry point to conduct a search for other offences that are sometimes more difficult to initial justify [15]. There

is some merit to this argument, as once the search begins, a person may be arrested for any offence, not necessarily related to drugs [15]. Indeed, the Metropolitan Police themselves have admitted to this in [132], stating: "The overall outcome rate from drug searches is currently 34 per cent and one third of all our weapons arrests from stop and search come from drug searches." The outcome rate of 34% is broadly in line with that in the December 2018 dataset, being 30%, however the arrest rate for drug-related searches is 13%, and is lower than the overall average for the dataset, which is at 16%.

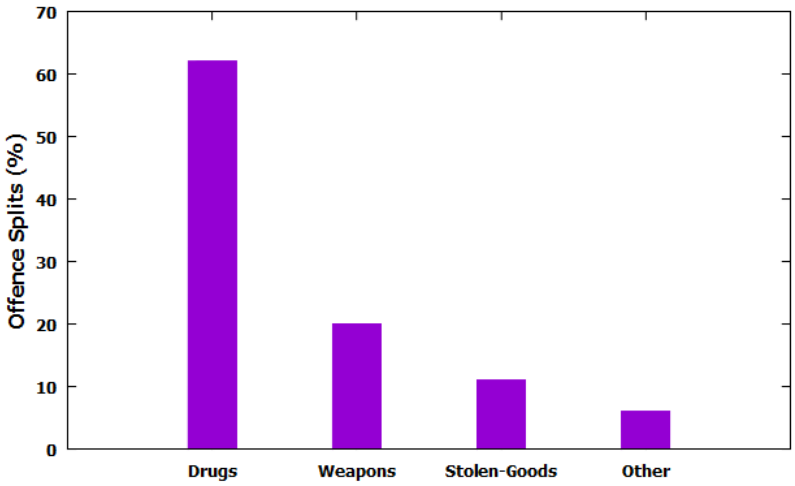


Figure 6.6: S&S by Offence - December 2018, Met Police Data

6.4.3.2 Stop and Search Efficiency

The top ten S&S categories by count, is shown in Table 6.11. Drugs-related searches make up eight of the top ten, with all categories involving males. As noted previously, this is not surprising given the requirements for suspicion of drugs-related offences may be seen as being less-stringent than other categories [15]. Indeed, nine of the ten categories detailed in Table 6.11 also have the highest arrests by count, as shown in Table 6.12. However, when comparing the data in Table 6.12 from an efficiency perspective, it is clear that all of the top ten S&S categories are below the average arrest efficiency of 20% that is published by the Metropolitan Police [132].

Thus it can be concluded, in line with several previous studies including in [27] and [111], that the top ten S&S categories are fairly inefficient, which not only leads to a waste of resources, but fosters an environment of higher than expected "false positives", which is damaging to both the police force and society at large.

Demographic	Offence	Support	Count
Black, Male, 18-24	Drugs	0.085	1326
Asian, Male, 18-24	Drugs	0.065	1007
White, Male, 18-24	Drugs	0.064	996
White, Male, 25-34	Drugs	0.050	774
Black, Male, 25-34	Drugs	0.050	772
White, Male, 34+	Drugs	0.035	548
Black, Male, 10-17	Weapon	0.034	531
Asian, Male, 25-34	Drugs	0.032	501
Black, Male, 34+	Drugs	0.031	479
Black, Male, 18-24	Weapon	0.029	457

Table 6.11: Top 10 S&S by count - December 2018

Demographic	Offence	Support	Count	Arrest Rate (%)
Black, Male, 18-24	Drugs	0.011	174	13%
White, Male, 18-24	Drugs	0.008	118	12%
White, Male, 25-34	Drugs	0.007	115	15%
Black, Male, 25-34	Drugs	0.007	114	15%
Asian, Male, 18-24	Drugs	0.007	103	10%
Black, Male, 34+	Drugs	0.005	83	17%
White, Male, 34+	Drugs	0.005	80	15%
Black, Male, 10-17	Weapon	0.005	76	14%
White, Male, 25-34	Stolen Goods	0.005	73	29%
Black, Male, 18-24	Weapon	0.004	69	15%

Table 6.12: Top 10 Arrests by count - December 2018

Arrest rate percent, as shown in Table 6.12, is an equivalent measure to the mt value and is widely used [27][132]. However, arrest rate percent, like the concept of percentages in general, is difficult to operationalise and be used as a tangible metric to both the public and front-line police [20]. In this regard, the mt value may be

considered to be a much more tangible metric and given that it is normalised, it readily allows for easy comparisons to be made across categories, and across police forces. For this application, the mt value is computed by letting $minsup =$ the number stopped and searched, i.e. $minsup = P(A)$ with $P(A, C)$ being the number arrested, and thus Equation (4.15) reduces to Equation (6.1).

$$\text{mt} = \frac{P(A)}{P(A, C)} - 1 \tag{6.1}$$

To illustrate the tangibility of the mt value, consider the data in Table 6.13 which shows the arrest percentage and mt value for the top ten arrest categories (by arrest rate percent) for the December 2018 dataset. The arrest rate percent and the mt value are inversely correlated, and point to same objective. However, the mt value may be considered easier to operationalise. For example: the data in Table 6.13 shows that there is a 37% arrest rate percent for black males over 34 years old being arrested for the possession of stolen goods. Operationalising this percentage at the "ground-level", may not mean anything tangible to the police officer, apart from the fact that this is the most efficient of all S&S categories. In contrast, the mt value, derived from Equation (6.1) represents the number of "false positives" per arrest. Using the above example, the police officer can then conclude that for every black male over 34 years old being arrested for the possession of stolen goods, a further 1.7 will be searched without an arrest. Put practically, there is a very high likelihood that every third black male over 34 years old being searched for the possession of stolen goods will be arrested, thus if the officer has searched two people in this category without an arrest, then by the laws of probability, the next person searched in this category will most likely be arrested.

Based on this, it can be seen that the mt value is a quick and easy measure for efficiency, and police officers "on the ground" can easily grasp the concept and target

those categories that have the lowest mt value, which in turn has the highest efficiency. The data extract in this analysis was a blend of race, gender, and age, however the data could be very easily be cut by location, which is another variable that is logged by police officers duing S&S. In the case of location, it will become easy to identify crime-related features including crime "hotspots" and places where criminals are generally concentrated, which can then be used to inform police officers as to where searches should be targeted.

Demographic	Offence	Arrest Rate (%)	mt value
Black, Male, 34+	Stolen Goods	37%	1.7
White, Male, 34+	Weapon	36%	1.8
Black, Male, 25-34	Stolen Goods	33%	2
Black, Male, 18-24	Stolen Goods	30%	2.3
White, Male, 25-34	Stolen Goods	29%	2.4
Black, Male, 34+	Weapon	29%	2.4
White, Male, 25-34	Weapon	29%	2.4
Black, Male, 10-17	Stolen Goods	26%	2.9
White, Male, 10-17	Stolen Goods	24%	3.1
White, Male, 18-24	Stolen Goods	24%	3.1

Table 6.13: Comparison of mt value and Arrest Rate % - December 2018

For completeness, it should be noted that the insight from the data in Table 6.13 is in stark contrast to the S&S data presented in Table 6.11. In Table 6.11, it was noted that eight of the ten S&S were drug related, however from Table 6.13, it can be seen that the most efficient operations are not in drug-related activity but rather in stolen goods, which account for seven of the top ten most efficient S&S categories. Further, it is also evident from Table 6.13, that older people, i.e. people 25 years and older, have a higher arrest rate than younger people, i.e. people younger than 25 years.

6.4.4 Summary Remarks on the use of the mt model in Crime Prevention

The mt model was used to analyse the effectiveness of the highly charged police power of stop and search (S&S). The notion of minimising "false positives" and "false negatives" also apply to this field [26]. While "false negatives" can indeed be quite dangerous in some instances as lives could be lost, all stakeholders (including the police, community and academic researchers) agree that "false positives" are far more damaging in the long term as it resource intensive, creates a sense of disharmony between police and communities, and ultimately contributes to the ongoing social decay as a result of victimisation and demoralisation of those being falsely searched [26].

The December 2018 S&S data from the Metropolitan Police was used as a case study. From an overview perspective, it was found that the majority of searches (62%) were conducted for drug-related offences on males. Several theories have been put forward as to why this is the case, including the fact that a drug-related search is usually far easier to justify than any other category [15]. From a demographic perspective, black people are more likely to be disproportionately searched than any other race group, while the split between young people (under 25 years) and older people (25 years and older) is approximately equal [74].

In terms of efficiency of S&S, measured as the arrest rate percent, it was found that the top ten search categories are indeed all below the stated average arrest of 20%. However, the top ten categories with the highest arrest rate include seven for stolen goods and three for weapons, with none of these featuring in the top ten S&S categories. Clearly, there is a case for the police force to redirect efforts to areas of greatest efficiency. Whilst the arrest rate percent metric is available to all within the police force, it is difficult to operationalise as it lacks tangibility, as is typically the case with

percentage metrics across other sectors. In this regard, the mt value may be far more effective from an operational perspective. Not only does it point to efficiency, like the percentage metric, but it also enhances tangibility for "on the ground" police officers as it quantifies the number of "false positives" per arrest. Given this, the conclusions in [26] best summarise the use of the mt model in S&S: "the mt model shows good promise in providing a simple, yet novel, efficiency metric that can be used by police forces with respect to S&S. This will potentially help them reduce "false positives" and redirect their resources to maximize the value of S&S.".

6.5 Summary

The main aim of this chapter was to demonstrate the applicability of the mt model in addressing decision making challenges across a wide and varied field of applications. In this regard, three everyday applications were considered namely: education, healthcare, and safety and security. In education, a novel approach, using the mt model, to tackling the serious problem of pupil absenteeism was developed, implemented and evaluated. Whilst in healthcare, the mt model was used to analyse public healthcare data, and showed potential usefulness in the generation of research hypotheses. Finally, the mt model was used in analysing S&S data, and showed good promise as being a simple, tangible efficiency metric that can be used by "on the ground" police to enhance their effectiveness in S&S. A summary of the key points for each application is highlighted below:

- **Education: Improving School Attendance**
  - An action research-based experiment was conducted on a "live" UK primary school, with the aim of using the mt model to enhance pupil attendance. The mt model correctly identified the most problematic, i.e. most impactful to overall absenteeism, of the ten school sessions, and in this case

it was Monday AM.

- School-wide initiatives were used to encourage attendance, both across all sessions, and focussing on Mondays. Results show that the school improved its overall attendance to meet the required 96% target, and that attendance in Monday AM sessions is no longer a problem.
- A significant positive impact of this study was also the improvement in attendance of persistent absentees. In this regard, the school saw the number of pupils classed as persistently absent decrease by more than half.
- This approach to tackling school absenteeism is novel, and may be seen as a unique contribution of this study.

- **Healthcare: Analysing the impact of the metabolic factors on CHD**
  - Publicly available, large scale, CHD data was analysed using the mt model, with a focus on the impact of the four metabolic conditions on CHD mortality.
  - Results show that some metabolic conditions, in some age groups are more important than others. This includes diabetes and high blood pressure across all age groups, but especially in young people, i.e. younger than 45 years old.
  - Results also showed that in general having a high BMI is not linked to CHD death, and in fact, having a high BMI could be beneficial. This is consistent with the well-known obesity paradox.
  - Overall it was concluded that the mt model could prove useful in healthcare research, as it is a quick and easy way of uncovering associations, which may be used to formulate research hypotheses.

- **Safety and Security: Using the mt model as an efficiency metric in**

S&S

- The mt model was applied on the Metropolitan Police S&S data for December 2018.
- Results show that the mt model quickly uncovered well-known inefficiencies and police officer biases in the data, and was able to highlight the most efficient S&S categories
- Whilst the police forces use arrest percentages as an efficiency metric, it was felt that this metric was difficult to operationalise as it lacked tangibility for "on the ground" police. In this regard, the mt value was a more tangible metric as it quantified the number of "false positives" for every arrest with a category.
- Overall it was concluded that the mt value shows promise as being a tangible metric that can be used by police forces.

Chapter 7

Conclusion

The grocery retail sector remains central to most citizens across the developed world, including the UK, as it is their main source of food and household goods [86]. However this sector has been undergoing significant changes over the past decade not only from the increased use of technology and online retailing, but also as a result of changes in societal behaviours around macro-factors including finance, the environment, diet and choice [154]. These changes are in part responsible for an increased appetite for individualisation and the resurgence of the "customer is king" mindset. Consequently, grocery retails now place significant emphasis on understanding their customers and tailoring their offers to each and every customer. As a result, data mining and data analytics is commonplace at most major grocery retailers [43][60].

Association rule mining, ARM, has been a cornerstone of data analytics research in grocery retail for several decades now, as the underlying principles in grocery retail sales have remained broadly unchanged [154]. In essence, customers still want high quality products, at their convenience, and at the lowest price. On the other hand, grocery retailers still want to maximise sales, and profitability, whilst retaining all of its existing customers, and attracting a proportion of new ones on an ongoing basis

[43][109]. In this regard, this study provides models based on ARM that straddle these objectives by offering grocery retailers tools to enhance sales of products, by nature of their association with other products, while at the same time, enhancing the offer to customers through promotions, or other incentives that result from retailers' attempts to reinforce these associations.

The market target (mt) model, which is a central element of this study, will be useful to all retailers who are continuously having to decide on which items to promote, in what combinations, and to which customers; as they strive to maximise their revenue and profitability. However, this decision making process is not endemic to grocery retail, and indeed, other fields have similar decision making processes. In light of this, the mt model has been applied successfully to other fields, including in education, medicine and crime prevention.

7.1 Mathematical Modelling and Algorithm

The mt model was developed, based on the research methodology detailed in Chapter 3, using a first-principles approach from the well-known support and confidence parameters that are synonymous with ARM. The premise of the mt model is grounded in ARM theory which states that, in general, items or itemsets that are sold frequently in grocery retailers are generally the main drivers for customer attraction, retention and revenue expansion [42][43][109]. Thus, driving items/itemsets towards frequency should be a key objective. However, given that there are several hundreds, if not thousands of itemsets combinations at grocery retailers, the decision of which itemset to promote is not easy [43][109]. In this regard, the mt model simplifies this decision-making process, by selecting the itemset that will reach the target frequency with the least effort. Thus, the mt model may be defined as an "effort-based" decision

making model, where the notion of "effort" can be any quantity including: money, energy, time, or even lives.

Having an effective customer segregation mechanism is essential for grocery retailers that want to offer tailored marketing [120]. Whilst there is an ongoing desire by customers to be treated as individuals, with unique wants and needs, as noted in [19] and [120], having individualised marketing will be a monumental, and indeed impractical task for most grocery retailers, given that they have millions of customers on their records [131]. One approach of addressing this issue and which presents an effect of tailoring is to cluster customers based on their propensity to purchase an item that is being promoted. This approach has been used in this study, and in other studies as well including in [140]. However, unlike in [140], where there is a real possibility of "false positives" and consequently the loss of tailoring, this study has placed constraints on customer targeting with only those customers who have a history of purchasing an item being targeted. As a result, "false positives" are eliminated, and customers get a sense of tailored promotions as they are only offered promotions on items that they have bought in the past, and may likely want to buy again. This approach thus enhances customer trust in the grocery retailer, and consequently encourages customer loyalty [42][109].

The computer-based algorithm proposed in this study is the amalgamation of both the mt model, and the customer clustering model which leverages FCM. The algorithm first provides a decision making step around which item to target, and then creates nine clusters for all prospective customers, all of whom have a history of purchasing the target item. The nine clusters are then grouped into four treatment plans, with each plan potentially offering a different incentive to the other. The underlying principle of these offers is to be sufficiently attractive to the target customer audience, but

not to undermine the overall revenue of the grocery retailer. Thus, from a financial perspective, and as is now common practice, loyal customers are generally not given the most lucrative incentives, as their custom is already well-established, with the best rewards given to the non-loyal customers, also known as "switchers" in this study [42].

7.2 Conclusions from Grocery Retail Experiments

The models and algorithm proposed in Chapter 4 were tested using two sets of consumer scanner panel data obtained from Kantar [90]. For completeness, and as outlined in Chapter 3, the models and algorithm were also tested using synthetic data generated using the algorithm proposed in [83]. The effectiveness of the proposed approach was compared against (1) the approach proposed in [140], (2) against "reality", by considering the holistic shopping patterns of the customers in the scanner panel, and (3) by targeting customers who buy "top sellers". The results showed that:

- The mt model was effective in choice making between two alternatives ($A \rightarrow C$) and ($B \rightarrow D$), in that it always selected the alternative that minimised the "effort" required to reach frequency.

- The clustering approach successfully grouped customers who had similar shopping behaviours, so that targeted marketing schemes offered by the grocery retailers can be effective, and give the effect of tailoring, whilst ensuring that the erosion of revenue streams from its loyal customers were minimised.

- Simulation testing showed that it was possible for the targeted itemset to reach frequency with both the "conservative" and "aggressive" marketing campaigns. Indeed, the aggressive campaigns enabled the target itemset to reach frequency sooner.

- Comparative testing of the proposed algorithm with the algorithm detailed by

Reutterer et al. in [140], showed that the proposed algorithm in this study outperformed the approach in [140] by both reducing "false positives" and "false negatives". Further, it was able to sufficiently segregate customers to offer differentiated treatment to maximise the impact of the grocery retailer's marketing spend, whilst minimising the grocery retailer's revenue erosion. This concept was not considered in the approach outlined in [140].

- Comparative testing of the proposed algorithm with "reality" showed that in general, the proposed algorithm was effective in clustering customers into the four loyalty groups, with the most loyal group consistently buying more products at the selected grocery retailer compared to other grocery retailers, and vice versa for "switchers".

- Comparative testing of the proposed model with targeting customers who buy "top sellers" showed that in general, a significantly large customer base has to be targeted, which will drive up costs. Indeed, limiting the customer base to only those that have a history of purchasing both the antecedent and consequent target item reverts back to the population proposed in this study, which may be considered as the optimal target population.

- Testing the algorithm's robustness, in terms of its processing speed, was done using synthetic datasets. Tests showed that whilst processing time increased for larger, denser datasets, it was still quite manageable, even with laboratory-scale computers.

- Overall, adopting a "functionalist" research paradigm, and the amalgamation of well-known research methodologies has led to this study following an approach that resulted in unique contributions that advanced the body of knowledge in data mining and grocery retailing.

7.3 Conclusions from Experiments in Other Applications

The mt model was applied to other fields that faced a similar decision making scenario, i.e. choosing between two independent alternatives, $(A \rightarrow C)$ and $(B \rightarrow D)$. Three applications, education, healthcare, and safety and security, were considered.

7.3.1 Education: Improving School Attendance

A detailed action-based research project was conducted at a primary school, where the mt model was used to find the school session that had the largest impact on overall school absence. Using this result, the school was able to offer incentives to pupils to encourage them to improve their attendance. The results showed:

- Applying the mt model to historic attendance data showed that the Monday AM session had the largest impact on overall school absence over the past three academic years. Thus, it may be possible to improve overall attendance by targeting Monday AM sessions with incentives.
- Attendance incentive schemes conducted in the Spring and Summer terms of 2018/19 showed that:
  - Overall school attendance improved from 95% in 2017/18 to 96.2% in 2018/19, the highest it has been over the last four years, and above the national requirement of 96% for the first time during the four year period.
  - Monday AM attendance during the Summer term of 2018/19 was significantly higher than the three previous years, and also above 96% for the first time during the four year period.
  - As a result of the interventions, Monday AM attendance is no longer the most impactful to overall school absence, with Friday AM now the most

problematic.

- Persistent absenteeism, the most damaging form of school absenteeism, also more than halved to 5.8% during the 2018/19 academic year compared with the 2017/18 academic year. Over the last three years, the school consistently had a persistent absenteeism issue that was above the national average, and is now significantly below the latest reported national average figure of 8.7%.

- As a result of the success, the school is continuing with the Monday incentives, and is now also creating new initiatives to address Fridays.

7.3.2 Healthcare: Applying the mt model to CHD data

The mt model was applied to large scale, public health data pertaining to Coronary Heart Disease (CHD), with the aims of reducing complexity and presenting the data in an unambiguous way, and to establishing whether the mt model can aid in setting public healthcare priorities. The results, based on the mt model calculations showed that:

- Having a high BMI is not necessarily a major risk factor for CHD across all age groups. Whilst this is contrary to popular belief, it is indeed well-documented that a high BMI can actually be protective during heart surgery [102].

- Diabetes is rapidly becoming the most important risk factor for CHD, in particular amongst younger people, i.e. less than 55 years old.

- Similar to diabetes, high blood pressure is also a significant risk factor amongst younger people.

- Cholesterol, although controlled through the use of statins, is less of a risk than high blood pressure and diabetes.

The results of the mt model were tested with the University of Leicester Cardiology Research Team, who noted that while some of the findings require further investigation, the use of mt model "could quickly uncover some associations and generate hypotheses for further studies" [124].

7.3.3 Safety and Security: Applying the mt model to S&S data

The mt model was applied to publicly available, Stop and Search (S&S) data for the Metropolitan Police, for December 2018. S&S is highly charged with differing points of view, however all stakeholders are in agreement that preventing "false positives" in S&S is a key priority. In this regard, the mt model was used to understand efficiency of S&S operations, and offer a tangible metric that could be used by "on the ground" police officers to reduce "false positives", and enhance the effectiveness of S&S. Applying the mt model to the S&S dataset showed:

- An overview analysis confirmed that most people are stopped for drug-related searches, which in some cases, can be used as a gateway to conduct more harder-to-justify searches.
- Black people have a disproportionately larger volume of searches than their white counterparts
- The top ten search categories have an arrest rate that is significantly lower than the stated average of 20%, implying that the top ten categories are possibly not all driven by "reasonable suspicion".
- The top ten arrest categories are predominantly for stolen goods, with none of these appearing in the top ten search categories, clearly pointing to inefficiencies
- Arrest rate percentage is the typical metric that is used for S&S effectiveness,

but difficult to operationalise as it lacks tangibility. The mt mt model, is considerably easier to operationalise, as it provides the number of "false positives" per arrest. Clearly the smaller the mt value, the more efficient the search category.

Overall, the mt value can be an effective, simple and reliable metric that could be operationalised for S&S. In a discussion with a leading S&S academic, it was concluded that "the mt model shows good promise in providing a simple, yet novel, efficiency metric that can be used by police forces with respect to S&S. This will potentially help them reduce "false positives", and redirect their resources to maximize the value of S&S. Whilst minimizing "false positives" is important from a policing operational efficiency perspective, it is even more important from a societal perspective, as it reduces the tension between police and communities, who sometimes live with the long lasting impacts of being falsely accused" [26].

7.4 Unique Contributions of this Study

The several unique contributions of this study to both the study of Market Basket Analysis, and Association Rule Mining, and the various applications are as follows:

7.4.1 Market Basket Analysis and Association Rule Mining

- The application of the uninorm to market basket analysis in $(A \rightarrow C)$ and $(A \rightarrow D)$ decision making was the first of its kind, and showed that it was superior to other well-recognised metrics including the cosine and Jaccard coefficient measures.

- The mt model, developed from first principles using the well-known support and confidence metrics, is an effective, simple, and novel method for helping decision makers choose between $(A \rightarrow C)$ and $(B \rightarrow D)$ alternatives.

7.4.2 Targeted promotion in grocery retail

- The application of the mt model as a decision making framework for targeted promotions has not been done before, and will help decision makers rapidly select the right combinations for promotions, thus minimising effort whilst improving outcomes.

- The use of the clustering approach is indeed novel, and has shown to be very effective, based on the testing approach in Chapter 5. The clustering approach effectively segregates customers for a differentiated treatment approach, and is based on their purchasing history.

- The computer-based targeted promotions algorithm, together with its pruning approach is new to the field of targeted promotions. It is highly effective, in that it combines the mt model, and the clustering approach to target those customers with products that they are most likely to buy. Indeed the proposed targeted promotions algorithm has been shown to be superior to the novel targeted promotions algorithm developed recently by Reutterer et al. in [140].

- The application of the simple, Markov-based simulator is also new to the grocery retail sector. Whilst Markov-based simulators have been used in a variety of applications [136], there is no evidence that it has been used in targeted promotion simulations, as is done in this study.

7.4.3 Improving School Attendance

- Using data mining to improve school attendance is novel, and there is no clear evidence that this has been previously done in a school setting. Indeed there have been studies conducted in the past that used data mining, including a study to understand the factors influencing educational performance, however attendance was one parameter amongst several others [39].

- The application of the mt model and the related algorithmic approach to improve school attendance is new, and has been shown to be highly effective. The action-research based project that was conducted at WPS, see Section 6.2, showed that session-targeting can have a significant impact on overall school attendance, an equally-importantly on the attendance of those classified as persistently absent.

7.4.4 Using the mt model in Healthcare and Safety and Security

The mt model is very flexible, and thus lends itself to be used in a number of novel ways. Apart from being used for targeting, as shown in both the grocery retail and school attendance applications, it has been used as a tool for root cause/hypothesis generation, and efficiency metric. These are detailed as follows:

- The use of the mt model on large-scale, publicly available coronary heart disease (CHD) data not only easily confirmed well-known cardiology tenants, including the obesity paradox, where people with high BMIs have better CHD survival outcomes, it also pointed out to emerging trends [124]. In terms of emerging trends, it showed the role of diabetes and high blood pressure as key risk factors for CHD, particularly amongst young people. Given this, it was concluded that there is role for the mt model in CHD research, as it could be used to quickly uncover hypotheses for future research [124].

- Experiments conducted on Stop and Search (S&S) data showed that the mt value, could be used as an operational metric for "on the ground" police officers, as it is simple and tangible. Further, from an effectiveness perspective, the mt model can be easily used by all stakeholders in crime prevention to quickly measure the effectiveness of S&S operations, identify any biases, and redirect

efforts so as to minimize “false positives” [26].

7.5 Limitations

Whilst the mt model itself is based on robust mathematical concepts, the use of the mt model, and the target promotions algorithm in applications does have application-specific limitations. These limitations are borne out of assumptions that are taken to simplify the analysis. Having said this, the mt model and algorithm does produce effective results, even with these assumptions taken, as evidenced in the various case studies that formed part of this study. However, for completeness, the following list of assumptions and associated limitations are discussed.

7.5.1 Limitations in grocery retail analysis

7.5.1.1 Items

There has been three primary assumptions taken around items namely:

- Quality of the item is homogenous across grocery retailers, i.e. similar products from each grocery retailer are equivalent substitutes for each other.
- Variability in quantities have a negligible influence on shopping patterns, i.e. customers buy a quantity that is required for their individual circumstances.
- Customers do not stockpile items.

Whilst all of the above assumptions, can be countered with specific cases, in general, it is well-known that customers are now more willing to swap items, and purchase quantities that are appropriate for their families to both minimise waste, and maintain cashflow [73][87][138]. As a result the assumptions on quality, quantity and stockpiling are reasonable for a grocery retailer-wide customer base study like this.

7.5.1.2 Customer loyalty

It is assumed that customer loyalty is purely out of customer choice, and not due to the lack of competition. This assumption is valid for the UK grocery retail sector where competition is ubiquitous across all postcodes, and customers have access to multiple online retailers, most of them delivering throughout the UK. For example, both Tesco, and the Coop have stores in every postcode, whilst ASDA, Sainsbury's and Tesco deliver to over 97% of all UK postcodes [89].

7.5.2 Limitations in improving school attendance

The underlying assumption at WPS was that children are generally homogenous as a collective from year to year, with no significant peaks or troughs in pupil behaviour or health-related issues that may impact overall attendance. Further, whilst weather does impact attendance, the last four years have been fairly stable with the occasional "snow day" in the Spring term being excluded from the attendance data. Hence, weather had a negligible impact on this study. Both these assumptions were discussed with WPS leadership, and it was confirmed that they are valid for the period under consideration [176].

7.5.3 Limitations in Other Applications

The application of the mt model to medical data was limited by the influence of confounding factors, e.g. people living with other diseases simultaneously with CHD, or lifestyle habits such as smoking, which is generally an issue for medical studies. In this regard, it was concluded that despite these influences on the large-scale data used for this study, the mt model can be used to rapidly identify areas to probe for future research [124]. As part of future work, it will be good to investigate the use of the mt model to datasets that have been controlled for confounding factors.

7.6 Wider Impact

The wider impact of the mt model, the clustering approach, and the proposed algorithm, which have all been developed against the backdrop of grocery retail has also been shown to have good applicability in other fields. For grocery retail, which has typically been a highly confidential and proprietary field, this study provides a foundation from which other researchers in the field can build on, based on their own datasets and research objectives. For example, researchers working on projects to tackle the impact of food consumption on climate change can use the mt model to find foods that are either highly positive or negatively associated with climate change, then segment the customers based on their consumption of these products, and finally target these customers with appropriate substitute products.

The approach taken for improving attendance at WPS can be used to support other schools, thus enhancing outcomes for a greater number of pupils across the country, and indeed across the world. Further, the approach can be adjusted to understand, among other concepts, the notion of "take-up". For example, universities could use the approach to understand why science, technology, engineering and mathematics (STEM) related subjects are currently less-attractive to female students, as compared to their male colleagues. In this regard, a series of "STEM attractiveness factors" can be developed and quantified for the student population to identify strong associations with overall take-up of STEM courses. Following this, these factors can then be promoted or addressed to encourage take-up.

Finding further applications for the use of the mt model will expand its wider impact. ARM models have been used across a variety of applications from weather prediction to medical diagnosis. However, as noted with the applications tested in this study, gaining access to both reliable data, and expertise in the relevant fields is not easy,

and in many ways are the key ingredients for success. Given this, it is anticipated that researchers using ARM in diverse fields will also test the applicability of the mt model as part of their work.

7.7 Future Work

There are several areas for extending the use of the mt model within the grocery retail sector. This study was based on product quantities, and this could be extended to include other variables such as revenue, profit, environmental friendliness, calories, etcetera. Another area of focus within the grocery retail sector can be the evaluation of changes to the product mixes on customers. For example, a store can use the mt model, and customer clustering approach to identify the best products to retire. Products that have a very large mt value, and has a high loyalty customer mix will be prioritised as these products are less-likely to become frequent, and less-likely to see customer defections as a result of it being retired.

With regards to improving school attendance, at WPS, the school will continue to focus on Mondays and also look to address Fridays, which has now become the new problematic school day in terms of absenteeism. The approach developed in this study will also be taken to other schools where attendance is a problem. In the medium term, this attendance improvement model will be developed into an easy-to-use software program with a graphical user interface, and will be made available to all schools that want to use it. This will enable schools to conduct their own attendance analysis without the need for expert data analytical, or computer coding skills, thus becoming empowered, and self-sufficient in improving educational outcomes for the communities that they serve.

On using the mt model for further analysis on the factors that influence CHD, the model has been made available to the University of Leicester Cardiology Research Team, so that they can use it on their own internal data sets where corrections have been made for confounding factors. It is hoped that the mt model will help them in uncovering new hypotheses, or rethinking some of their existing ones.

Finally on S&S, the mt model will be used to create a "heat map" of S&S effectiveness across the London boroughs. This will then be mapped to a crime "heat map" and comparisons drawn. For example, these comparisons will reveal where there is a high propensity for arresting people with stolen goods, or where there is likely community bias in which S&S activity disproportionately targets certain communities, with very poor arrest outcomes. It is envisaged that this activity will be in collaboration with University College London's Jill Dando Policing Institute were contacts have been established with S&S experts.

7.8 Concluding Remarks

The motivation for this study was to understand the purchasing behaviour of customers in grocery retail, with an attempt to enhance product targeting. This led to adopting a "functionalist" research paradigm, and an amalgamation of well-known research methodologies, to develop of the mt model, the clustering approach, and the associated algorithm, which has been shown to be successful in achieving this objective. However, a strong, consequential impact of this study was the application of the mt model to other applications. Indeed, all the other applications detailed in this study demonstrate that the mt model can be used as a force for public good.

At educational institutions, the mt model can be used to enhance educational out-

comes, particularly in people that come from disadvantaged backgrounds, by uncovering root causes in absenteeism or societal biases. In healthcare, the mt model can be used to challenge conventional thinking, or uncover new hypotheses that will likely lead to improved patient outcomes for some of the world's most dreaded illnesses. On crime prevention, the mt model can be used to bring more rigour to S&S activity, thus fostering a "win-win" environment where communities can benefit from its crime prevention potential, whilst not being unfairly discriminated against.

Overall, it is believed that this study has laid the foundation from which other researchers, in various fields, can build on, and most likely enhance their research outcomes.

Bibliography

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, 23(1):103–145, 2005.

[2] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[3] Tommaso Agasisti and Alex J Bowers. 9. data analytics and decision making in education: Towards the educational data scientist as a key actor in schools and higher education institutions. In *Handbook of contemporary education economics*, page 184. Edward Elgar Publishing, 2017.

[4] Chain Store Age. Five ways walmart uses big data. `https://www.chainstoreage.com/operations/five-ways-walmart-uses-big-data/`, 2017. Accessed: 2019-02-28.

[5] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[6] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[8] Emel Aktas and Yuwei Meng. An exploration of big data practices in retail sector. *Logistics*, 1(2):12, 2017.

[9] Shahriar Akter and Samuel Fosso Wamba. Big data analytics in e-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2):173–194, 2016.

[10] Waleed A Aljandal, William H Hsu, Vikas Bahirwani, Doina Caragea, and Tim Weninger. Validation-based normalization and selection of interestingness measures for association rules. In *Intelligent Engineering Systems through Artificial Neural Networks Volume 18*. ASME Press, 2008.

[11] Sara Allin, Elias Mossialos, Martin McKee, Walter Werner Holland, World Health Organization, et al. Making decisions on public health: a review of eight countries. Technical report, Copenhagen: WHO Regional Office for Europe, 2004.

[12] Azarnoush Ansari and Arash Riasi. Customer clustering using a combination of fuzzy c-means and genetic algorithms. *International Journal of Business and Management*, 11(7):59, 2016.

[13] Aviva. The uk grocery market: ripe for disruption. `https://www.avivainvestors.com/en-gb/views/aiq-investment-thinking/`

2018/12/the-uk-grocery-market-ripe-for-disruption/, 2018. Accessed: 2019-05-20.

[14] Robert Balfanz and Vaughan Byrnes. Chronic absenteeism: Summarizing what we know from nationally available data. *Baltimore: Johns Hopkins University Center for Social Organization of Schools*, 1(1):1–46, 2012.

[15] BBC. Stop-and-search: A bbc reporter's own experience. https://www.bbc.co.uk/news/uk-41849256, 2017. Accessed: 2019-08-10.

[16] BBC. Crime: What has boris johnson promised on law and order? https://www.bbc.co.uk/news/uk-49318400, 2019. Accessed: 2019-08-30.

[17] BBC. What do the people of the world die from? https://www.bbc.co.uk/news/health-47371078, 2019. Accessed: 2019-08-10.

[18] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.

[19] Anil Bilgihan, Jay Kandampully, and Tingting Zhang. Towards a unified customer experience in online shopping environments: Antecedents and outcomes. *International Journal of Quality and Service Sciences*, 8(1):102–119, 2016.

[20] Michael Blastland and Andrew W Dilnot. *The tiger that isn't: seeing through a world of numbers*. Profile books, 2008.

[21] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[22] Booz and Co. U.s. grocery shopper trends. http://www.icn-net.com/docs/12086_FMIN_Trends2012_v5.pdf, 2012. Accessed: 2017-01-30.

[23] Christian Borgelt. Simple algorithms for frequent item set mining. In *Advances in machine learning II*, pages 351–369. Springer, 2010.

[24] Christian Borgelt. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):437–456, 2012.

[25] Yasemin Boztuğ and Thomas Reutterer. A combined approach for segment-specific market basket analysis. *European Journal of Operational Research*, 187(1):294–312, 2008.

[26] Ben Bradford. Personal communication with prof. ben bradford, ucl. personal communication, 2019. Date: 2019-09-10.

[27] Ben Bradford and Matteo Tiratelli. Does stop and search reduce crime? 2019.

[28] John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[29] Wilson J Brown, Allison K Wilkerson, Stephen J Boyd, Daniel Dewey, Franklin Mesa, and Brian E Bunnell. A review of sleep disturbance in children and adolescents with anxiety. *Journal of sleep research*, 27(3):e12635, 2018.

[30] Robin Burke. Hybrid web recommender systems. In *The adaptive web*, pages 377–408. Springer, 2007.

[31] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *Principles of Data Mining and Knowledge Discovery*, pages 74–86. Springer, 2002.

[32] Robert L Cannon, Jitendra V Dave, and James C Bezdek. Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):248–255, 1986.

[33] Zeynel Cebeci and Figen Yildiz. Comparison of k-means and fuzzy c-means algorithms on different cluster structures. *Agrárinformatika/Journal of Agricultural Informatics*, 6(3):13–23, 2015.

[34] David Coghlan. *Doing action research in your own organization*. SAGE Publications Limited, 2019.

[35] Philip J Cook, Kenneth A Dodge, Elizabeth J Gifford, and Amy B Schulting. A new program to prevent primary school absenteeism: Results of a pilot study in five schools. *Children and Youth Services Review*, 82:262–270, 2017.

[36] MK Council. Guidance for improving school attendance. `https://www.milton-keynes.gov.uk/assets/attach/51484/`, 2019. Accessed: 2019-07-13.

[37] Ben Daniel. B ig d ata and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, 46(5):904–920, 2015.

[38] Data.Police. Met police stop and search data. `https://data.police.uk/`, 2019. Accessed: 2019-08-10.

[39] Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D Lytras, Farhat Abbas, and Jalal S Alowibdi. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion*, pages 415–421. International World Wide Web Conferences Steering Committee, 2017.

[40] Renan de Padua, Exupério Lédo Silva Junior, Laís Pessine do Carmo, Veronica Oliveira de Carvalho, and Solange Oliveira Rezende. Preprocessing data sets for association rules using community detection and clustering: a comparative study.

[41] DMU. Guidelines for good research practice. https://www.dmu.ac.uk/research/ethics-and-governance/research-integrity-and-ethics.aspx, 2019. Accessed: 2019-08-28.

[42] Matilda Dorotic. Keeping loyalty programs fit for the digital age. *NIM Marketing Intelligence Review*, 11(1):24–29, 2019.

[43] Dunhumby. Heartbeat market basket analytics. https://www.dunnhumby.com, 2015. Accessed: 2019-02-28.

[44] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[45] Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian. Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, 5(2):112, 2015.

[46] Retail Economics. Ten trends that are changing the landscape of the uk retail sectors. http://www.retaileconomics.co.uk/insights/ten-trends-that-are-changing-the-landscape-of-the-UK-retail-sector.asp, 2015. Accessed: 2017-01-30.

[47] Fiona Ellis-Chadwick, Neil F Doherty, and Leonidas Anastasakis. E-strategy in the uk retail grocery sector: a resource-based analysis. *Managing Service Quality: An International Journal*, 17(6):702–727, 2007.

[48] Public Health England. Making sense of data: A challenge and a responsibility. https://publichealthmatters.blog.gov.uk/2015/11/13/making-sense-of-data-a-challenge-and-a-responsibility/, 2019. Accessed: 2019-07-30.

[49] Money Saving Expert. Can you take kids on term-time holidays without being fined? https://www.moneysavingexpert.com/family/school-holiday-fines/, 2019. Accessed: 2019-03-25.

[50] Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. Rfm and clv: Using iso-value curves for customer base analysis. *Journal of marketing research*, 42(4):415–430, 2005.

[51] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[52] Ling Feng, Tharam Dillon, and James Liu. Inter-transactional association rules for multi-dimensional contexts for prediction and their application to studying meteorological data. *Data & Knowledge Engineering*, 37(1):85–115, 2001.

[53] Ling Feng, Jeffrey Xu Yu, Hongjun Lu, and Jiawei Han. A template model for multidimensional inter-transactional association rules. *The VLDB Journal*, 11(2):153–175, 2002.

[54] George S Fishman. *Discrete-event simulation: modeling, programming, and analysis*. Springer Science & Business Media, 2013.

[55] Nathan M Fong. How targeting affects customer search: A field experiment. *Management Science*, 63(7):2353–2364, 2016.

[56] Nathan M Fong, Yuchi Zhang, Xueming Luo, and Xiaoyi Wang. Targeted promotions and cross-category spillover effects. 2016.

[57] Dept for Education. School attendance and schools: Statutory guidance. https://www.gov.uk/government/organisations/department-for-education/, 2019. Accessed: 2019-07-13.

[58] The Institute for Health Metrics and Evaluation. Global burden of disease study 2017. http://ghdx.healthdata.org/gbd-2017, 2019. Accessed: 2019-06-10.

[59] Forbes. Big data at tesco: Real time analytics at the uk grocery retail giant. https://www.forbes.com/sites/bernardmarr/2016/11/17/big-data-at-tesco-real-time-analytics-at-the-uk-grocery-retail-giant/#3496de7961cf, 2016. Accessed: 2019-05-20.

[60] Forbes. Big data at tesco: Real time analytics at the uk grocery retail giant. http://www.forbes.com/sites/bernardmarr/2016/11/17/big-data-at-tesco-real-time-analytics-at-the-uk-grocery-retail-giant/#43bd506a519a, 2016. Accessed: 2017-01-30.

[61] British Heart Foundation. Eggs and cholesterol. https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2015/may/eggs-and-cholesterol, 2015. Accessed: 2019-06-10.

[62] British Heart Foundation. Eggs linked to heart disease and death, study suggests. https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2019/march/eggs-and-cardiovascular-risk, 2019. Accessed: 2019-06-10.

[63] British Heart Foundation. Heart statistics. https://www.bhf.org.uk/what-we-do/our-research/heart-statistics, 2019. Accessed: 2019-06-10.

[64] Minos N Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Spirit: Sequential pattern mining with regular expression constraints. In *VLDB*, volume 99, pages 7–10, 1999.

[65] Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

[66] Els Gijsbrechts, Katia Campo, and Mark Vroegrijk. Save or (over-) spend? the impact of hard-discounter shopping on consumers' grocery outlay. *International Journal of Research in Marketing*, 2018.

[67] C Giraud-Carrier and O Povel. Characterising data mining software. *Intelligent Data Analysis*, 7(3):181–192, 2003.

[68] Bart Goethals. Frequent set mining. *Data mining and knowledge discovery handbook*, pages 321–338, 2010.

[69] L Gordon. Leading practices in market basket analysis: How top retailers are using market basket analysis to win margin and market share. *Factpoint Group*, 2008.

[70] Anjana Gosain and Sonika Dahiya. Performance analysis of various fuzzy clustering algorithms: a review. *Procedia Computer Science*, 79:100–111, 2016.

[71] Karam Gouda and Mohammed Zaki. Efficiently mining maximal frequent itemsets. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 163–170. IEEE, 2001.

[72] Guardian. How supermarkets get your data. `https://www.theguardian.com/money/2013/jun/08/supermarkets-get-your-data`, 2013. Accessed: 2019-02-28.

[73] Guardian. Uk grocery sales in decline for first time in 20 years. `https://www.theguardian.com/business/2014/nov/18/uk-grocery-sales-decline-price-war-asda-sainsburys-morrisons-tesco`, 2014. Accessed: 2018-11-30.

[74] Guardian. Met police 'disproportionately' use stop and search powers on black people. `https://www.theguardian.com/law/2019/jan/26/`

met-police-disproportionately-use-stop-and-search-powers-on-black-people, 2019. Accessed: 2019-08-10.

[75] Eui-Hong Han, George Karypis, Vipin Kumar, and Bamshad Mobasher. Hypergraph based clustering in high-dimensional data sets: A summary of results. *IEEE Data Eng. Bull.*, 21(1):15–22, 1998.

[76] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[77] Jiawei Han, Micheline Kamber, and Jian Pei. Mining frequent patterns, associations, and correlations: Basic concepts and methods. 2012.

[78] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[79] Gregory L Hanna, Daniel J Fischer, and Thomas E Fluent. Separation anxiety disorder and school refusal in children and adolescents. *Pediatrics in Review*, 27(2):56, 2006.

[80] Hossein Hassani, Xu Huang, Emmanuel S Silva, and Mansi Ghodsi. A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(3):139–154, 2016.

[81] Trude Havik, Edvin Bru, and Sigrun K Ertesvåg. School factors associated with school refusal-and truancy-related reasons for school non-attendance. *Social Psychology of Education*, 18(2):221–240, 2015.

[82] Wu He, Jiancheng Shen, Xin Tian, Yaohang Li, Vasudeva Akula, Gongjun Yan, and Ran Tao. Gaining competitive intelligence from social media data:

evidence from two largest retail chains in the world. *Industrial Management & Data Systems*, 115(9):1622–1636, 2015.

[83] Jeff Heaton. Comparing dataset characteristics that favor the apriori, eclat or fp-growth frequent itemset mining algorithms. In *SoutheastCon 2016*, pages 1–7. IEEE, 2016.

[84] David Heyne, Malin Gren-Landell, Glenn Melvin, and Carolyn Gentle-Genitty. Differentiation between school attendance problems: Why and how? *Cognitive and Behavioral Practice*, 26(1):8–34, 2019.

[85] Nada Hussein, Abdallah Alashqur, and Bilal Sowan. Using the interestingness measure lift to generate association rules. *Journal of Advanced Computer Science & Technology*, 4(1):156, 2015.

[86] IGD. Uk food and grocery evolution 2014 - 2019. `http://www.igd.com/Research/Retail/UK-food-and-grocery-evolution-2014---2019/`, 2014. Accessed: 2017-01-30.

[87] IGD. Uk food and grocery market to grow 14.8 `https://www.igd.com/articles/article-viewer/t/uk-food-and-grocery-market-to-grow-148-by-282bn-by-2023/i/19052`, 2018. Accessed: 2019-05-20.

[88] London Imperial College. Heart disease deaths nearly halved in uk - but condition remains top killer. `https://www.imperial.ac.uk/news/191414/heart-disease-deaths-nearly-halved-uk/`, 2019. Accessed: 2019-09-10.

[89] This is Money. Which supermarket really delivers? `https://www.thisismoney.co.uk/money/bills/article-6955339/We-reveal-supermarkets-compare-online-deliveries`, 2019. Accessed: 2019-08-30.

[90] Kantar. Kantar's consumer panel shopping data - uk. Private Communication, 2017.

[91] Kantar. Supermarkets await easter sales boost. `https://uk.kantar.com/consumer/shoppers/2019/supermarkets-await-easter-sales-boost/`, 2019. Accessed: 2019-05-20.

[92] Jonathan Karnon. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. *Health economics*, 12(10):837–848, 2003.

[93] Christopher A Kearney and Patricia Graczyk. A response to intervention model to promote school attendance and decrease school absenteeism. In *Child & Youth Care Forum*, volume 43, pages 1–25. Springer, 2014.

[94] Orhan Kesemen, Özge Tezel, and Eda Özkul. Fuzzy c-means clustering algorithm for directional data (fcm4dd). *Expert Systems with Applications*, 58:76–82, 2016.

[95] Farnoosh Khodakarami and Yolande E Chan. Exploring the role of customer relationship management (crm) systems in customer knowledge creation. *Information & Management*, 51(1):27–42, 2014.

[96] Shah Khusro, Zafar Ali, and Irfan Ullah. Recommender systems: issues, challenges, and research opportunities. In *Information Science and Applications (ICISA) 2016*, pages 1179–1189. Springer, 2016.

[97] Paul Klemperer. Price wars caused by switching costs. *The Review of Economic Studies*, 56(3):405–420, 1989.

[98] Severin Klingler, Tanja Käser, Barbara Solenthaler, and Markus Gross. Temporally coherent clustering of student data. *International Educational Data Mining Society*, 2016.

[99] Yun Sing Koh and Russel Pears. Rare association rule mining via transaction clustering. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pages 87–94. Australian Computer Society, Inc., 2008.

[100] Yun Sing Koh and Russel Pears. Transaction clustering using a seeds based approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 916–922. Springer, 2008.

[101] V Kumar and Werner Reinartz. Loyalty programs: Design and effectiveness. In *Customer Relationship Management*, pages 179–205. Springer, 2018.

[102] Carl J Lavie, Alban De Schutter, Parham Parto, Eiman Jahangir, Peter Kokkinos, Francisco B Ortega, Ross Arena, and Richard V Milani. Obesity and prevalence of cardiovascular diseases and prognosis—the obesity paradox updated. *Progress in cardiovascular diseases*, 58(5):537–547, 2016.

[103] Richard D Lawrence, George S Almasi, Vladimir Kotlyar, Marisa Viveros, and Sastry S Duri. *Personalization of supermarket product recommendations*. Springer, 2001.

[104] Tien-Duy B Le and David Lo. Beyond support and confidence: Exploring interestingness measures for rule-based specification mining. In *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pages 331–340. IEEE, 2015.

[105] Andrew Leicester. The potential use of in-home scanner technology for budget surveys. Technical report, National Bureau of Economic Research, 2013.

[106] Friedrich Leisch and Bettina Grün. *Extending standard cluster algorithms to allow for group constraints*. na, 2006.

[107] Qing Li, Ling Feng, and Allan Wong. From intra-transaction to generalized inter-transaction: landscaping multidimensional contexts in association rule mining. *Information Sciences*, 172(3):361–395, 2005.

[108] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments: A survey. *Decision Support Systems*, 74:12–32, 2015.

[109] Manthan. Food and grocery analytics in 2018 – the new age realities. `https://www.manthan.com/blogs/food-and-grocery-analytics-in-2018-the-new-age-realities/`, 2018. Accessed: 2019-02-15.

[110] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719, 2017.

[111] Rhydian McCandless, Andy Feist, James Allan, and Nick Morgan. Do initiatives involving substantial increases in stop and search reduce crime? assessing the impact of operation blunt 2. *London: Home Office*, 2016.

[112] Margaret McCartney. Media's misrepresentation of science. `https://www.bmj.com/bmj/section-pdf/914631?path=/bmj/352/8044/`, 2019. Accessed: 2019-07-30.

[113] Areej Fatemah Meghji, Naeem Ahmed Mahoto, Mukhtiar Ali Unar, and Muhammad Akram Shaikh. Analysis of student performance using edm methods. In *2018 5th International Multi-Topic ICT Conference (IMTIC)*, pages 1–7. IEEE, 2018.

[114] Agathe Merceron, Kalina Yacef, C Romero, S Ventura, and M Pechenizkiy. Measuring correlation of strong symmetric association rules in educational data. *Handbook of educational data mining*, pages 245–256, 2010.

[115] Raymond Moodley, Francisco Chiclana, Fabio Caraffini, and Jenny Carter. Application of uninorms to market basket analysis. *International Journal of Intelligent Systems*, 34(1):39–49, 2019.

[116] María N Moreno, Saddys Segrera, Vivian F López, María Dolores Muñoz, and Ángel Luis Sánchez. Web mining based framework for solving usual problems in recommender systems. a case study for movies[U+05F3] recommendation. *Neurocomputing*, 176:72–80, 2016.

[117] Juho Muhonen and Hannu Toivonen. Closed non-derivable itemsets. In *Knowledge Discovery in Databases: PKDD 2006*, pages 601–608. Springer, 2006.

[118] Maryam Khanian Najafabadi, Mohd Naz'ri Mahrin, Suriayati Chuprat, and Haslina Md Sarkan. Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. *Computers in Human Behavior*, 67:113–128, 2017.

[119] Janmenjoy Nayak, Bighnaraj Naik, and HS Behera. Fuzzy c-means (fcm) clustering algorithm: a decade review from 2000 to 2014. In *Computational intelligence in data mining-volume 2*, pages 133–149. Springer, 2015.

[120] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[121] NHS. New flu vaccine available this winter for those aged 65 and over. https://www.gov.uk/government/news/

new-flu-vaccine-available-this-winter-for-those-aged-65-and-over/, 2018. Accessed: 2019-08-28.

[122] Nielson. The future of grocery. https://www.nielsen.com/content/dam/nielsenglobal/vn/docs/Reports/2015/Nielsen%20Global%20E-Commerce%20and%20The%20New%20Retail%20Report%20APRIL%202015%20(Digital).pdf, 2015. Accessed: 2017-01-30.

[123] Jay F Nunamaker Jr, Minder Chen, and Titus DM Purdin. Systems development in information systems research. *Journal of management information systems*, 7(3):89–106, 1990.

[124] University of Leicester Cardiology Research Team. Personal communication with leadership. personal communication, 2019. Date: 2019-06-24.

[125] Ofsted. Willen primary school inspection reports. https://reports.ofsted.gov.uk/provider/21/110388/, 2019. Accessed: 2019-07-10.

[126] Özalp Özer, Ozalp Ozer, and Robert Phillips. *The Oxford handbook of pricing management*. Oxford University Press, 2012.

[127] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[128] Ken Peffers, Tuure Tuunanen, and Björn Niehaves. Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research, 2018.

[129] Gregory Piatetsky-Shapiro and Christopher J Matheus. Knowledge discovery workbench for exploring business databases. *International Journal of Intelligent Systems*, 7(7):675–686, 1992.

[130] Marie Plasse, Ndeye Niang, Gilbert Saporta, Alexandre Villeminot, and Laurent Leblond. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis*, 52(1):596–613, 2007.

[131] Tesco PLC. Tesco - open heart surgery in public. `https://www.tescoplc.com/media/264601/`, 2019. Accessed: 2019-07-03.

[132] The Metropolitan Police. How we use stop and search. `https://www.met.police.uk/advice/advice-and-information/st-s/stop-and-search/how-we-use-stop-and-search/`, 2019. Accessed: 2019-09-10.

[133] Michael E Porter. How competitive forces shape strategy. In *Readings in strategic management*, pages 133–143. Springer, 1989.

[134] Michael E Porter. *Competitive strategy: Techniques for analyzing industries and competitors*. Simon and Schuster, 2008.

[135] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.

[136] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[137] Nicholas J Radcliffe and Patrick D Surry. Differential response analysis: Modeling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI. Edinburgh, Scotland*, 1999.

[138] The Register. Big data, empty bellies: How supermarkets tweak prices just for the sake of your love. `https://www.theregister.co.uk/2015/02/05/big_data_tech_weapons_in_supermarket_price_wars/?page=2`, 2015. Accessed: 2017-01-30.

[139] IBM Research. Synthetic data generator code for associations and sequential patterns. `http://www.research.ibm.com/labs/almaden/index.shtml#assocSynData`, 2017. Accessed: 2017-01-30.

[140] Thomas Reutterer, Kurt Hornik, Nicolas March, and Kathrin Gruber. A data mining framework for targeted category promotions. *Journal of Business Economics*, 87(3):337–358, 2017.

[141] Hongjai Rhee and David R Bell. The inter-store mobility of supermarket shoppers. *Journal of Retailing*, 78(4):225–237, 2002.

[142] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[143] Carly D Robinson, Monica G Lee, Eric Dearing, and Todd Rogers. Reducing student absenteeism in the early grades by targeting parental beliefs. *American Educational Research Journal*, 55(6):1163–1192, 2018.

[144] Thomas Rolfsnes, Leon Moonen, Stefano Di Alesio, Razieh Behjati, and Dave Binkley. Improving change recommendation using aggregated association rules. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pages 73–84. IEEE, 2016.

[145] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce*, pages 158–167. ACM, 2000.

[146] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[147] M Lewis Saunders and Philip Lewis. P. & thornhill, a.(2009). *Research methods for business students*, 4, 2009.

[148] Willen Primary School. School website. https://willenprimary.com/, 2019. Accessed: 2019-07-10.

[149] National Health Service. The healthy way to eat eggs. https://www.nhs.uk/live-well/eat-well/eggs-nutrition/, 2019. Accessed: 2019-06-10.

[150] Luke Sibieta, Chris Belfield, et al. 2018 annual report on education spending in england. 2018.

[151] Kit N Simpson, Alvin Strassburger, Walter J Jones, Birgitta Dietz, and Rukmini Rajagopalan. Comparison of markov model and discrete-event simulation techniques for hiv. *Pharmacoeconomics*, 27(2):159–165, 2009.

[152] Surbhi K Solanki and Jalpa T Patel. A survey on association rule mining. In *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, pages 212–216. IEEE, 2015.

[153] David Solnet, Yasemin Boztug, and Sara Dolnicar. An untapped gold mine? exploring the potential of market basket analysis to grow hotel revenue. *International Journal of Hospitality Management*, 56:119–125, 2016.

[154] Zeynab Soltani and Nima Jafari Navimipour. Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. *Computers in Human Behavior*, 61:667–688, 2016.

[155] Herb Sorensen, Svetlana Bogomolova, Katherine Anderson, Giang Trinh, Anne Sharp, Rachel Kennedy, Bill Page, and Malcolm Wright. Fundamental patterns of in-store shopper behavior. *Journal of Retailing and Consumer Services*, 37:182–194, 2017.

[156] Alexander Strehl and Joydeep Ghosh. A scalable approach to balanced, high-dimensional clustering of market-baskets. In *International Conference on High-Performance Computing*, pages 525–536. Springer, 2000.

[157] Charlotte Sturley, Andy Newing, and Alison Heppenstall. Evaluating the potential of agent-based modelling to capture consumer grocery retail store choice behaviours. *The International Review of Retail, Distribution and Consumer Research*, 28(1):27–46, 2018.

[158] Karan Sukhija, Manish Jindal, and Naveen Aggarwal. The recent state of educational data mining: A survey and future visions. In *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pages 354–359. IEEE, 2015.

[159] Joffre Swait and Rick L Andrews. Enriching scanner panel models with choice experiments. *Marketing Science*, 22(4):442–460, 2003.

[160] Pang-Ning Tan. *Introduction to data mining*. Pearson Education India, 2018.

[161] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[162] Pham Huy Thong et al. Picture fuzzy clustering: a new computational intelligence method. *Soft computing*, 20(9):3549–3562, 2016.

[163] Jose Torres. Race/ethnicity and stop-and-frisk: past, present, future. *Sociology Compass*, 9(11):931–939, 2015.

[164] Mauricio A Valle, Gonzalo A Ruz, and Rodrigo Morrás. Market basket analysis: Complementing association rules with minimum spanning trees. *Expert Systems with Applications*, 97:146–162, 2018.

[165] Harald J van Heerde and Scott A Neslin. Sales promotion models. In *Handbook of marketing decision models*, pages 13–77. Springer, 2017.

[166] Rajkumar Venkatesan and Paul W Farris. Measuring and managing returns from retailer-customized coupon campaigns. *Journal of marketing*, 76(1):76–94, 2012.

[167] Neha Verma and Jatinder Singh. A comprehensive review from sequential association computing to hadoop-mapreduce parallel computing in a retail scenario. *Journal of Management Analytics*, 4(4):359–392, 2017.

[168] Bay Vo and Bac Le. Interestingness measures for association rules: Combination between lattice and hash tables. *Expert Systems with Applications*, 38(9):11630–11640, 2011.

[169] Baoying Wang, Imad Rahal, and Aijuan Dong. Parallel hierarchical clustering using weighted confidence affinity. *International Journal of Data Mining, Modelling and Management*, 3(2):110–129, 2011.

[170] Ke Wang, Chu Xu, and Bing Liu. Clustering transactions using large items. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 483–490. ACM, 1999.

[171] Yu Wang and Fusheng Wang. Association rule learning and frequent sequence mining of cancer diagnoses in new york state. In *VLDB Workshop on Data Management and Analytics for Medicine and Healthcare*, pages 121–135. Springer, 2017.

[172] Marketing Week. Why sainsbury's isn't waiting for 'big silver bullets' to drive business transformation. `https://www.marketingweek.com/2018/12/13/sainsburys-transforming-business-data/`, 2018. Accessed: 2019-05-20.

[173] Elizabeth Wilkins, L Wilson, Kremlin Wickramasinghe, Prachi Bhatnagar, Jose Leal, Ramon Luengo-Fernandez, R Burns, Mike Rayner, and Nick Townsend. European cardiovascular disease statistics 2017. 2017.

[174] Roland Winkler, Frank Klawonn, and Rudolf Kruse. Problems of fuzzy c-means clustering and similar algorithms with high dimensional data sets. In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, pages 79–87. Springer, 2012.

[175] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[176] WPS. Personal communication with wps leadership. personal communication, 2019. Date: 2019-07-10.

[177] Jian Wu, Ruoyun Xiong, and Francisco Chiclana. Uninorm trust propagation and aggregation methods for group decision making in social network with four tuple information. *Knowledge-Based Systems*, 96:29–39, 2016.

[178] Wanli Xing, Rui Guo, Eva Petakovic, and Sean Goggins. Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, 47:168–181, 2015.

[179] Hui Xiong, P-N Tan, and Vipin Kumar. Mining strong affinity association patterns in data sets with skewed support distribution. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 387–394. IEEE, 2003.

[180] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[181] Jinmei Xu, Hui Xiong, Sam Yuan Sung, and Vipin Kumar. A new clustering algorithm for transaction data via caucus. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 551–562. Springer, 2003.

[182] Shuo Yang, Mohammed Korayem, Khalifeh AlJadda, Trey Grainger, and Sriraam Natarajan. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach. *Knowledge-Based Systems*, 136:37–45, 2017.

[183] Yinghui Yang, Hongyan Liu, and Yuanjue Cai. Discovery of online shopping patterns across websites. *INFORMS Journal on Computing*, 25(1):161–176, 2013.

[184] Ching-Huang Yun, Kun-Ta Chuang, and Ming-Syan Chen. An efficient clustering algorithm for market basket data based on small large ratios. In *Computer Software and Applications Conference, 2001. COMPSAC 2001. 25th Annual International*, pages 505–510. IEEE, 2001.

[185] Ching-Huang Yun, Kun-Ta Chuang, and Ming-Syan Chen. Using category-based adherence to cluster market-basket data. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 546–553. IEEE, 2002.

[186] Mohammed J Zaki. Efficient enumeration of frequent sequences. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 68–75. ACM, 1998.

[187] Mohammed J Zaki. Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3):372–390, 2000.

[188] Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2):31–60, 2001.

[189] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[190] Mohammed Javeed Zaki and Ching-Jiu Hsiao. Charm: An efficient algorithm for closed itemset mining. In *SDM*, volume 2, pages 457–473. SIAM, 2002.