Submitted in part fulfilment of the requirements for the degree of Master of Science in Business Analytics

Application of data mining and machine learning in retail marketing

By

Hoang Viet Ho

Faculty of Arts and Social Sciences

UNIVERSITY OF SURREY

November 2020

Word count: 16,000.

# Executive summary

The report provides an explanation of how data mining and machine learning can be applied in retail to enhance consumer understanding and predict their behaviour. By using different data mining techniques, retailer can have a detailed overview of their customer base and an understanding of how to conduct different marketing strategies. Moreover, tuning the hyperparameters of the binary classification models in the machine learning process helps a business anticipate consumer participation based on the personal details of a consumer such as their most frequently bought products or demographic data.

Literature review suggests various known insights of consumer behaviour and marketing in previous researches within the retail context. Mainly, by understanding main types of customers in their business, retailers can tailor their business to customer’s liking and maintain a competitive edge. Various methods such as clustering and K-mean technique have been used in different literatures on customer portfolios to develop different clusters in which consumers behave in a similar manner, while customer behaviour can also be identified using association and basket analysis. Moreover, machine learning can especially be used in the marketing aspect of retailing. Examples of the applications include using techniques such as Artificial Neural Network, Decision Trees or Support vector machines can examine churned customers which helps the retailer design specific marketing strategies to retain consumers, as well as using past information of individual consumers as insights to predict whether a consumer has an interest in a campaign.

For the analysis, the dataset “The complete journey” provided by Dunnhumby offers transaction data of 2,500 consumers along with other supplementary data. From household information to the types of campaign they are directly advertised, and which households redeem a coupon. As the data is highly complicated in terms of structure, the methodology section first aims to filter and construct the features and characteristics that is representative of each consumer which is then united as a final dataset that has the most illustrative variables for each consumer in the dataset, including the campaign participation status. Such features include demographic information, characteristics of most bought products and consumer purchase behaviours (such as most frequent transaction time of the day or spending per trip/per week).

Data mining techniques such as visualizations are used on the final dataset to reveal insights regarding consumers: such as most common age bracket, income group as well as spending per week of each consumer and the specific types of products most bought by households. Besides consumer specific information, insights regarding marketing effectiveness is also revealed. Specifically, the more a marketing campaign is advertised, the more effective it will be in increasing the spending over time of consumers during the campaign duration.

Classification algorithms are then performed on binary target classification problem presented in the final dataset, which aims to categorise whether a consumer will participate in a campaign or not based on the consumer-specific information. Random forest, boosted C5.0 decision trees and bagged decision trees are performed on the dataset, with a focus on predicting positives/consumer that participates (Sensitivity) as well as a balance on filtering negatives/consumer who aren’t interested (Specificity). Each model shows varying degree of effectiveness. Random Forest and Bagged decision trees have around 90% in sensitivity, suggesting that a majority of interested consumers can be correctly identified. Specificity also does not suffer too significantly with around an average of 60% of uninterested consumers identified correctly, with the exception of C5.0 boosted model with up to 87% specificity, in exchange of lower sensitivity with around 80%. Area Under the Curve, which indicates the ability of a classifier to distinguish a positive from a negative, are over 0.8 for all models which is close to a perfect classifier with an AUC of 1. This further suggests that the models executed are accurate in predicting interested consumers.

Models can be used in different scenarios depending on the marketing strategies used by the retailer. If the retailer focuses on mass and cheap marketing strategies in high quantity, Random Forest is the most advantageous as it is very efficient in predicting interested consumers in exchange for a lower accuracy in uninterested customer classification. This is done in order to maximise consumer awareness. On the other hand, if the retailer deploys cost intensive yet highly effective strategies (focuses on quality of the campaigns), boosted C5.0 decision trees offers a more balanced false positives and true positive rates, which allows retailers to still reach a decent number of interested consumers while waste of resources is reduced as false positive is remained at a minimum level.

The research also has some limitations. Due to limited processing and computing capabilities, full datasets cannot be used for classification. The complexity of a few categorical variables on a small subset of data chosen for analysis also proves a challenge as the models will be harder to generalize on new datasets unless similar transformation is conducted on unseen data to make sure that the mismatch of categorical values compared to the training set is eliminated. The second limitation suggests that the prediction of consumer participation in a more personalized level, such as predicting whether a customer will participate in a specific campaign among many other campaigns sent to them, is not yet examined. Hence, more opportunities for the study of direct marketing personalization can be further conducted.

***Section word count: 878***

# Declaration of Originality

“I hereby declare that this thesis has been composed by myself and has not been presented or accepted in any previous application for a degree. The work, of which this is a record, has been carried out by myself unless otherwise stated and where the work is mine, it reflects personal views and values. All quotations have been distinguished by quotation marks and all sources of information have been acknowledged by means of references including those of the Internet. I agree that the University has the right to submit my work to the plagiarism detection sources for originality checks.”

Author’s signature

A drawing of a face

Description automatically generated

Hoang Viet Ho

# Acknowledgements

First and foremost, I would like to express my deepest appreciation to my dissertation supervisor, Dr Vikas Grover. My dissertation would not be possible without his patience and thoughtful assistance as well as his detailed feedbacks for the direction of my work throughout each stage of writing, especially during these unusual times and circumstances that has been taking place when arrangements can be difficult. His calm and considerate guidance as well as expertise in machine learning has been my greatest source of support. I would also like to thank him for proofreading my work and provide many crucial constructive criticisms.

I would also like to extend my gratitude to my module leader Dr Masoud Fakhimi and my personal tutor Dr Tao Huang for their support and early advices of the crucial decisions towards the completion of my course modules. I also wish to thank the staffs of the Faculty of Art and Social Sciences, and specifically the teaching staff of Surrey Business School for providing me invaluable tools and knowledge necessary for my Dissertation.

Finally, I am forever grateful for my family with their unconditional love, comfort and reassurance. I’m extremely lucky to have them as my emotional support not only during my dissertation, but also in many crucial milestones.

# Table of contents

[Executive summary 1](#_Toc55080070)

[Declaration of Originality 4](#_Toc55080071)

[Acknowledgements 5](#_Toc55080072)

[Table of contents 6](#_Toc55080073)

[List of figures 9](#_Toc55080074)

[List of tables 11](#_Toc55080075)

[Chapter 1. Introduction 12](#_Toc55080076)

[1.1. Background 12](#_Toc55080077)

[1.2. Main aim and objectives. 13](#_Toc55080078)

[1.2.1. Aims 13](#_Toc55080079)

[1.2.2. Objectives 13](#_Toc55080080)

[1.3. Contributions 14](#_Toc55080081)

[1.4. Structure 14](#_Toc55080082)

[1.5. Summary 15](#_Toc55080083)

[Chapter 2: Literature review 16](#_Toc55080084)

[2.1. Grocery retail industry and customer profiles 16](#_Toc55080085)

[2.1.1. Grocery industry overview: Definition, importance and key statistics. 16](#_Toc55080086)

[2.1.2. Customer segmentation in the industry 17](#_Toc55080087)

[2.2. Behaviour and engagement choice of customers in grocery retailing 19](#_Toc55080088)

[2.2.1. Consumer behaviour and determination methods 19](#_Toc55080089)

[2.2.2. Factors affecting retailer engagement choice and purchase behaviour: 20](#_Toc55080090)

[2.3. Direct marketing’s effects on consumer engagement behaviour in grocery retailing 21](#_Toc55080091)

[2.4. Application of data science in targeted marketing enhancements 23](#_Toc55080092)

[2.4.1. Overview of Data Mining in the retail context 23](#_Toc55080093)

[2.4.2. Data mining’s application in CRM and Targeted marketing 24](#_Toc55080094)

[2.5. Summary 26](#_Toc55080095)

[Chapter 3: Methodology 27](#_Toc55080096)

[3.1. Research philosophy 27](#_Toc55080097)

[3.2. Research approach and analytical method 28](#_Toc55080098)

[3.2.1. Research approach 28](#_Toc55080099)

[3.2.2. Analytical method 28](#_Toc55080100)

[3.3. Research methodology 29](#_Toc55080101)

[3.3.1. Overview of research process 29](#_Toc55080102)

[3.3.2. Data collection and understanding 30](#_Toc55080103)

[3.3.2.1. Data collection 30](#_Toc55080104)

[3.3.2.2. Data description and understanding 31](#_Toc55080105)

[3.3.3. Data cleaning 32](#_Toc55080106)

[3.3.4. Data pre-processing 33](#_Toc55080107)

[3.3.4.1. First stage: Collect in concern and create variables for final dataset 33](#_Toc55080108)

[3.3.4.2. Second stage: Pre-process final dataset for classification modelling 35](#_Toc55080109)

[3.3.5. Classification model selection 39](#_Toc55080110)

[3.3.6. Modelling methodology 41](#_Toc55080111)

[3.4. Summary 43](#_Toc55080112)

[Chapter 4: Findings and classification results 44](#_Toc55080113)

[4.1. Finding and analysis of findings 44](#_Toc55080114)

[4.1.1. Customer characteristics 44](#_Toc55080115)

[4*.1.2. Customer purchase behaviour* 46](#_Toc55080116)

[4.2.2. Marketing information and effectiveness 53](#_Toc55080117)

[4.3. Classification results 58](#_Toc55080118)

[4.3.1. Random forest 58](#_Toc55080119)

[4.3.2. Boosted c5.0 decision tree 59](#_Toc55080120)

[4.3.3. Bagged decision tree 61](#_Toc55080121)

[4.4 Summary 64](#_Toc55080122)

[**Chapter 5. Discussions** 66](#_Toc55080123)

[5.1. Overall findings 66](#_Toc55080124)

[5.2. Classification performance 66](#_Toc55080125)

[5.2.1. Random forest 66](#_Toc55080126)

[5.2.2. Boosted C5.0 Decision trees 67](#_Toc55080127)

[5.2.3. Bagged decision trees 68](#_Toc55080128)

[5.2.4 Final remarks of classification models 68](#_Toc55080129)

[**5.3. Summary** 69](#_Toc55080130)

[Chapter 6: Conclusions 70](#_Toc55080131)

[6.1. Overall achievements 70](#_Toc55080132)

[6.1. Implications 70](#_Toc55080133)

[6.2. Limitations and opportunities 71](#_Toc55080134)

[References 73](#_Toc55080135)

[SUBMISSION DOCUMENTS 85](#_Toc55080136)

[ENTRY FORM 85](#_Toc55080137)

[ETHICAL FORM 86](#_Toc55080138)

# List of figures

[Figure 1. The CRISP-DM process for data mining 28](#_Toc55080560)

[Figure 2. The Complete Journey dataset table map (Source Files - dunnhumby, 2020). 30](#_Toc55080561)

[Figure 3. Final data used for classification model building and other data components 32](#_Toc55080562)

[Figure 4. Importance of chosen variables 37](#_Toc55080563)

[Figure 5. Performance of various models on the dataset. 39](#_Toc55080564)

[Figure 6. Visualization of household’s income composition 42](#_Toc55080565)

[Figure 7. visualization of customer age composition 43](#_Toc55080566)

[Figure 8. Visualization of household composition 44](#_Toc55080567)

[Figure 9.Visualization of most common transaction time. 45](#_Toc55080568)

[Figure 10.Visualization of spending per trip of households 46](#_Toc55080569)

[Figure 11. Visualization of spending per week of households 47](#_Toc55080570)

[Figure 12. Average spending per trip and per week of each age group. 48](#_Toc55080571)

[Figure 13.Average spending per trip and per week for each income group 48](#_Toc55080572)

[Figure 14.Visualization of most liked products 49](#_Toc55080573)

[Figure 15.Most liked grocery products 50](#_Toc55080574)

[Figure 16.Most liked drugs and general merchandise 50](#_Toc55080575)

[Figure 17.Most liked produce goods 51](#_Toc55080576)

[Figure 18.Tree map of campaigns sent, visualized by proportion. 52](#_Toc55080577)

[Figure 19.Weekly spending of customers with campaign type A and highlights of each campaign’s duration. 53](#_Toc55080578)

[Figure 20. Weekly spending of customers with campaign type B and highlights of each campaign’s duration. 54](#_Toc55080579)

[Figure 21.Weekly spending of customers with campaign type C and highlights of each campaign’s duration. 55](#_Toc55080580)

[Figure 22.ROC and AUC of two random forest models before (left) and after tuning(right) 57](#_Toc55080581)

[Figure 23.Best combination of hyperparameters for Boosted C5.0 decision trees. 58](#_Toc55080582)

[Figure 24.ROC and AUC of boosted decision tree before (left) and after tuning (right) 59](#_Toc55080583)

[Figure 25. Performance of the bagged decision trees model at different nbagg levels 60](#_Toc55080584)

[Figure 26. Visualization of the model performance at different nbagg levels 61](#_Toc55080585)

[Figure 27. ROC and AUC curve of bagged decision trees before (left) and after tuning (right). 62](#_Toc55080586)

# List of tables

[Table 1. Balance of the data before and after using SMOTE 34](#_Toc55080171)

[Table 2. High carinal variables before and after transformation 34](#_Toc55080172)

[Table 3. Chi-squared test results for categorical variables 35](#_Toc55080173)

[Table 4. Features with near-zero variance 36](#_Toc55080174)

[Table 5. Correlation between numeric variables 37](#_Toc55080175)

[Table 6. Marital status of households 43](#_Toc55080176)

[Table 7. Homeowning status of households 44](#_Toc55080177)

[Table 8. Number of weekly trips and percentage 45](#_Toc55080178)

[Table 9. Quantity and percentage of campaigns sent by each type 51](#_Toc55080179)

[Table 10. Confusion matrix of default random forest and random forest with tuned hyperparameters on test data 56](#_Toc55080180)

[Table 11. Summary statistics of two random forest models 56](#_Toc55080181)

[Table 12. Summary statistics of two random forest models 57](#_Toc55080182)

[Table 13. C5.0 boosted decision tree confusion matrix with default parameters and tuned parameters on test data 58](#_Toc55080183)

[Table 14. Summary statistics of two boosted c5.0 models. 58](#_Toc55080184)

[Table 15. Summary statistics of boosted C5.0 decision tree 59](#_Toc55080185)

[Table 16. Confusion matrix of bagged decision trees with default and tuned parameters on test data 61](#_Toc55080186)

[Table 17. Summary statistics of Bagged decision trees 61](#_Toc55080187)

[Table 18. Summary statistics of bagged decision trees 62](#_Toc55080188)

# Chapter 1. Introduction

1.1. Background

In the retail industry, consumers are always a central focus. The study of consumers and their behaviour is a major aspect in Marketing, and it is crucial to analyse, influence and predict the consumer’s psychology. This can help the business not only in inventory management but also in tailoring their products/services to the customer’s preferences. But most importantly, businesses can develop sufficient marketing strategies to attract, retain customers and increase overall sales.

Almost all consumers in countries ranging from developed to developing countries have encountered and are influenced by advertising and promotional activities. Businesses in all industries, especially retail, have been finding different ways and methods of communicating with customers (Kazmi and Batra, 2008). With fragmented target audience, it is crucial for businesses to know their customer base and specific consumer segments, which can help them develop niche marketing mix that can influence the buying decisions and behaviour of consumers. A few studies suggest that effective mobile promotion strategies in retailing can motivate shoppers to move further inside the stores, which can increase unplanned spending. (Hui, Inman, Huang and Suher, 2013). Another positive side effect is that marketing strategies have an impact on customer royalty and retainment. Studies have shown that there is a connection between the two factors, in which loyalty reward programs and personalization has a positive impact on consumer behaviour. (Bojei, Julian, Wel and Ahmed, 2013)

The analysis of marketing activities and how it affects consumer behaviour can be assisted with the help of Big Data Analytics. In the context of retail where large number of transactional data is produced and available for examination, big data marketing analytics can aid businesses significantly in obtaining insights from data so that businesses can understand their customer base and enhance marketing performance which can maximise return on investments (Wedel and Kannan, 2016). For instance, firms can deploy clickstream analytics by using Google trends data to match consumer and non-consumer patterns and identify consumer segments for behavioural targeting (Du and Kamakura, 2012).

This work will focus on determining whether or not big data analytics and machine learning can be used in retail to determine the effectiveness of marketing campaigns and extract crucial insights of customers through different data mining techniques. Secondly, the insights gained are used as predictors to accurately predict customer behaviour and response to the direct marketing strategies via binary classification algorithms, including Random Forest, Boosted C5.0 decision trees and Bagged decision trees.

The dataset named “The Complete Journey” is provided by Dunhumby, a global customer data science company. This is a complex dataset that contains household level transactions of over 2,000 households who are frequent shoppers of a retailer. This data set spans over two years and contains the purchase data of all households as well as other information ranging from demographic, product information (of up to 90,000 products) to advertisement campaigns for each household. The dataset, if processed and analysed efficiently, will provide crucial insights about the relationship between promotional activities and consumer spending in the retail industry, specifically grocery retail.

1.2. Main aim and objectives.

1.2.1. Aims

There are two separate aims for this dissertation. The first aim is to find out if it is possible to use data mining techniques to explore consumer insights and the effects of marketing campaigns on consumer spending in retail.

The second aim involves classifying whether a customer will participate in a marketing campaign by using different customer’s demographic and transaction data as predictors to classify a binary target variable.

1.2.2. Objectives

Generally, there are three main objectives that need to be implemented.

1. Find out retail customer characteristics based on demographic information and transaction history. Identify retail spending pattern/habit of customers, in terms of overall transaction data and demographic data is the second sub-objective.
2. Identify relationships between marketing tactics and the change in retail spending/engagements by first looking at the main characteristics of the campaigns carried out by the retailer. This is used as supplementary insights for examining any changes in consumer spending during the occurrence of the marketing campaign.
3. Predict household’s behaviour to marketing campaigns by using predictive aspect of the machine learning process. Different predictors will be used in a few predictive models which aims to accurately classify whether a household will participate in a marketing campaign or not.

1.3. Contributions

Firstly, as many retail datasets can be complicated in terms of structure and size, the dissertation sets the direction of which consumer features in such data can be narrowed down to. This not only serves prediction but also knowledge mining purposes. Particularly, it is established that demographic variables (income, age, household size and composition), the characteristics of the product most frequently bought by each households (product department and sub-department) and their purchase behaviour (average spending per trip and trips per week) serves as great predictive variables as they are all dependent with the consumer participation indicator. Therefore, such features should be focused on when attempting to understand consumer features and their response behaviour.

Secondly, the dissertation contributes to the researches by combining between data mining and machine learning to produce effective classification models that can accurately predict the probability of a consumer participating in marketing campaigns. This allows the retailers to effectively increase consumer awareness by reaching the right consumer, while reducing waste of resources on unnecessary advertisements. As the primary focus is on participation prediction, the dissertation explores what algorithm is best chosen for the particular problem and how models are tuned for maximised Sensitivity (customers who are interested) while making sure that False Positives (uninterested consumers that are wrongly identified as having an interest) are reduced to a minimum.

1.4. Structure

Chapter 2 explains the literature review. It examines the known insights and studies conducted in previous researches relating to consumer behaviour, marketing strategies effectiveness and different applications of data science in the retail context. Specifically, this chapter determines the overview of the retail industry, customers profiles and how they are segmented. This is followed by a deeper look into how the consumer behaves, factors affecting their engagement choice and how these variables can be determined. Subsequently, marketing strategies’ effectiveness in changing the consumer psyche is also explained, followed by a detailed examination of how data science has been implemented in various settings to improve the effectiveness of the strategies.

Chapter 3 explains the methodology. This chapter clarifies the type of research philosophy and paradigm adapted in the research, followed by a clear explanation of how the research is conducted. Data collection method, data explanation, pre-processing and modelling techniques are also discussed in detail.

Chapter 4 explains the findings derived from data mining, and the results of the classification modelling. The chapter is divided into two parts. The first part involves data mining as visualizations through figures and graphs are presented. This is aimed to explain the characteristics of the consumers of the retailer, the information of how the marketing campaigns are carried out by the retailer and whether they are effective or not. The second part involves the optimization of different machine learning models which are used to predict a customer’s participation in a marketing campaign. The results of the process are presented

Chapter 5, built on the previous chapter, proceeds to discuss the results presented. For the findings of the data mining process, general observations will be discussed along with interpretation of the meaning. For modelling results, the section will explain in detail how each model performs, strength and weaknesses as well as some suggestions of how each model should be used depending on the marketing strategies deployed by the retailer.

Chapter 6 is the conclusion of the dissertation. Besides addressing key remarks, this chapter aims to provide the limitations of the current work and potential research that can be carried out with regards to the matter.

1.5. Summary

This chapter has introduced the background of the dissertation: To use data mining and machine learning for consumer and marketing insights discovery and prediction. This chapter also explains the structure of the dissertation as well as the aims and objectives that needs to be carried out to meet the aims. Next chapter explains the literature review, which focuses on exploring the known theories involving consumer segmentation, behaviour and prediction.

**Word count: 1,368**

Chapter 2: Literature review

* 1. Grocery retail industry and customer profiles
     1. Grocery industry overview: Definition, importance and key statistics.

When it comes to the general retailing industry, the activities have transformed into more sophisticated and complex processes, which results in an array of definitions. A traditional definition simply stated that retailing is the procedure of product procurement from other organisations in order to re-sell them to the customers (Zentes et al., 2017). However, it has evolved overtime from a buying and assorting process to a highly customer-oriented procedure in which effective exploitation of marketing and management results in deeper understanding of consumer behaviour and preferences. More precisely, retail is the process involve in marketing a product or service to consumers via availability administration in a large scope, followed by the supply to customers in a smaller extent (Amit and Kameshvari, 2012). Ultimately, a retailer provides goods or services via various channels to the end user, regardless of whether they are individual consumers, businesses or manufacturers.

The grocery industry can be defined into more specific categories. Specifically, it is classified as: Traditional and Non-Traditional (FMI | Supermarket Facts, 2018). The main difference between Traditional and Non-Traditional retailers is that Traditional retailers decide the products/services offered for the customers whereas Non-traditional retailers tailor their business to suit the demand of their consumer base. Traditional retailers include traditional supermarkets, small groceries and super-warehouses while Non-traditional retailers are represented by E-commerce, supercentres and mass stores (such as Walmart or Target). (FMI | Supermarket Facts, 2018). Among the different categories, major channels of grocery retailing include Supermarkets, Convenience Stores, Discounters, Online and Hypermarkets.

Grocery retailing industry is one of the key industries within the United Kingdom. It is enormous in size and influential to the development of UK economy and welfare, with the market value steadily increases over the years to £193.6 Billion in 2019 (IGD, 2019), while market size stays at £90 Billion (Mintel, 2019). From the channel perspective, large stores and supermarkets holds the highest Market value with £90 Billion followed by Convenience Stores and Discounters (IGD, 2019). The most noticeable feature of the Grocery market in the UK is that it is highly concentrated. Tesco, Sainsbury’s, Asda and Morrisons are the “Big 4” major players in the industry, accounts for 66% of the total market share (Grocery Market Share - Kantar, 2020). This not only shows that consumers are more likely to purchase in supermarkets rather than in other grocery channels, it also indicates the lack of negotiating power of grocery suppliers (FMI | Supermarket Facts, 2018) since there is a limited number of potential customers. Despite having a stronger market position, supermarkets are threatened by the development of online and discounters. Since 2012, the online grocery channel grew doubled in market size and it fulfils the big basket demands, which impact the performance of major supermarkets (Mintel,2019). The industry has high growth potential and it is expected to reach £204 Billion in value by 2024.

* + 1. Customer segmentation in the industry

The main characteristics of the grocery industry, such as constant development of online shopping and discounters, makes it highly competitive. This requires businesses in the industry to achieve sufficient competitive advantage by dealing with their customers in a more efficient manner. Products and services need to be tailored to the customer’s liking and personally customised to establish customer royalty and satisfaction. This can be achieved through a variety of methods, but the most important and crucial step is to always understands the customer base through customer segmentation, as they are divided into distinctive groups, so that they can be dealt with accordingly. Customer segmentation is very closely linked to identifying consumer behaviour. This is essential to understand the customer psyche, their value and how to retain high value customers (Webster, 1992). Since each customer segment responds differently to marketing strategies, segmentation helps businesses identify and target the best customer group in terms of profit and loyalty. This is the most crucial insight that in necessary for direct marketing since businesses can customised their marketing strategies to suit the desired customer group.

An array of methods has been used to study the customer profiles. Due to the simplicity and how easily it can be applied to real cases, one of the most common method of doing so is via the use of Recency, Frequency and Monetary (RFM) method, which is defined as a tool to evaluate value of customers by predicting the behaviour of customers through metrics such as time of last purchase, money spent and number of purchases within a period (Bult and Wansbeek, 1995). It has been widely exploited in different literatures (Khajvand and Tarokh, 2011; Fader et al., 2005, Hu and Yeh, 2014) and proven to be effective in identifying common types of consumers within the context of food retailing. The RFM model is then further modified and extended by adding extra variables such as customer relation length, periodicity of customers visits, time since first transaction and the probability of churn (Reinartz and Kumar, 2000; Yeh et al., 2009; Peker et al., 2017), which increases the accuracy of the original RFM model.

Another method of consumer classification includes using different data mining techniques. For example, classification techniques such as clustering can also be combined with RFM so that typical consumers can be identified based on Customer Lifetime Value (CLV), which can be defined as the future flow of cash attributed to customers during the entire relationship between them and the company (Definition of 'Customer Lifetime Value', 2020). Clustering analytical methodology such as K-mean technique can be used on customer portfolios to develop different clusters in which consumers behave in a similar manner, and is different from other clusters, which leads to the identification of consumer segments (Ngoc Do, 2011; You et al., 2015; Dogan et al., 2018).

Within the grocery retail industry, the consumer segments vary across multiple contexts, such as business types, locations and consumer-specific conditions. In general, the segmentation of customers is often divided into two groups: By demographic factors (such as gender, age, income, lifestyle and circumstances, etc.) and by purchasing behaviours (such as frequency/quantity of purchase and line of products usually consumed) (Wedel and Kamakura, 2000). This information can be deducted via the use of complete purchase history (Aeron et al., 2012, Miguéis et al., 2012). The behaviour of a consumer is analysed by looking at what the customer purchased, how much they purchased based on one and multiple visits, and frequency of visits which allows the division of customers into unique categories.

In order to achieve the analysis of behaviour and segments of customers, results derived from using clustering techniques can still be used as efficient estimators. However, the customer segmentation is unique and different depending on the context of the analysis. K-mean analysis done by Ngoc Do shows that three consumer groups are identified as unique (Ngoc Do, 2011) from the five categories based on the foundation of Portfolio analysis: Staples, Variety, Enhancers, Niches and Fill Ins (Dhar et al., 2001). Each customer group have a different combination of metrics such as purchase frequency and the value of each purchases, which makes each group homogenous. Customer Value Matrix, which uses frequency of purchases and the basket quantity, divides the customers into four different categories such as Spender, Frequent, Uncertain and Best (who shops frequent and buy a large number of items) (Marcus, 1998). Meanwhile, Clustering analysis done by Kinsey identified six different segments of shoppers: One Stop Socialist, Discriminating leisure shoppers, back to nature shoppers, middle of the road, no-nonsense and time-pressed meat eaters (Kinsey et al., 2001).

The clustering techniques can also be combined with RFM in which separate values of Recency, Frequency and Monetary can create subgroups with different indicator combinations. Based on the techniques, it also reveals effective classification methods yet unique to the context of the business such as classification by the loyalty of customers. The “Loyal” customer group will have better RFM score compared to other customer groups which also have their own RFM value (Dogan et al., 2018).

* 1. Behaviour and engagement choice of customers in grocery retailing
     1. Consumer behaviour and determination methods

In the retail industry, customer is the core of any business activities and it is important to understand consumer behaviour. Consumer behaviour can be defined as the psychological and physical activities that the consumers engage in during the process of selecting, purchasing and using the products/services that fulfil their needs (Keng et al., 2013). It also refers to the characteristics that affect the consumer as well as their beliefs and trust in different parties such as in the retailers (Orji, 2017).

It is important to grasp multiple aspects of consumer purchase such as problem formulation stage, purchase engagement (and how those actions satisfy their needs and desires) to post-purchase activities such as the tendency of re-purchasing or re-purchase frequency (Orji, 2017). Understanding the preferences, habits and how consumer interacts with different variables is crucial not only essential in developing, customising goods and services but also in designing effective customer retaining strategies. Knowing what the customers buy and how they make those decisions, organisations can explore and choose the best tactic that can influence those decisions and attract consumers. It is determined as the key aspect in the process of designing one-to-one marketing tactics and personalising customer services (Kim et al., 2003; Ha, 2007)

Consumer behaviour can be determined in a variety of ways. Data analytical techniques such as data mining is common, and it can be used on not only market segmentation but also to predict consumer behaviour. Specifically, retail businesses generate a lot of transactional and customer related data. This can be used as a basis for data mining to discover, analyse patterns of behaviour as well as to develop predictions of what the customer will buy in their next shopping trip (Giudici and Passerone, 2002). Specifically, a certain behaviour can be identified using association rules to discover the correlation between the product transaction (derived from the transaction database) and demographic variables (Song et al., 2001). It is an unsupervised mining method as it analyses quantity and types of items in each transaction of a consumer. Relationships and patterns are extracted as rules (Kotu and Deshpande, 2019). Association rules can also be used to discover the changes over time in consumer behaviour by analysing it in different time periods (Chen et al., 2005).

This data-driven technique is closely related and used alongside Basket Analysis, which is a powerful tool in the retail marketing context that can be used to specifically learn more about the consumer engagements, habits and preferences. Affinities among the items that are frequently purchased together (such as items in the shopping cart/basket of a shopper) are identified and they can reveal patterns and behaviours which can assist the business in designing a compatible marketing mix, creating offers to retain customers, positioning products close together in stores or cross-selling them (Loshin and Reifer, 2013).

* + 1. Factors affecting retailer engagement choice and purchase behaviour:

The factors that impact the engagement probability and purchasing behaviour of shopper is as plenty as the methods of determining it. One factor that impact the likelihood of consumer engagement is the store attributes. Specifically, retailers can impact the behaviour of consumers by having a positive image of different features and characteristics such as store atmosphere, layout convenience and design, social environment and even the name of the brand (Grewal et al., 2017), which will lead to positive reputation to customers, increasing loyalty and word of mouth. Service convenience can be considered a factor that impacts consumer behaviour. Retailers having a convenient service will create incentive for customer to spend their effort and scarce time and resources to engage with the retailer (Roy, et al., 2020). A retailer that provides a wide variety of complimentary services, as well as availability in a wide range of location and in flexible time frames will be more likely to worth the time and effort of the customers, which is known to be rare.

Not only store image, but price image is also an important influential factor upon the behaviour of the shopper. In developed European countries, discounters’ operations and constant expansion poses a threat to traditional retail stores as it affects the store choice (and retail brand choice) of shoppers due to the effectiveness of price-aggressive discounts programs (Colla, 2003). A shopper having a good impression of a retailer’s price level means a value is considered to be obtained in exchange of the money spent. This also means a reasonable compromise between perceived quality and sacrifice (Zielke, 2006). Hence, it is perceived as “Well-priced” and products are appropriate to the customer’s purchasing power (Jinfeng and Zhilong, 2009). This leads to the retailer not only having a good price image (which refers to how the customer perceive the store in terms of price range), but also a good value-for-money reputation. This positively impact a customer’s engagement intention, incentives and behaviour of buying in the store (Zielke, 2010). Sometimes, combination of different offers and discount gives the customers a lot of value in return for the money that they pay for, hence having a good reputation of price-to-performance ratio will attract customers to buy more, and also more likely to choose the retailer because of good value impression.

## Direct marketing’s effects on consumer engagement behaviour in grocery retailing

The importance of having a strong and active information power is gradually identified by grocery retailers. Enhanced brand image through maintaining the reputation of shopping experience and customer service can greatly improve customer loyalty and satisfaction. It is common among retailers to develop their brand image by creating centralised and uniform retail offers alongside different marketing strategies. Marketing strategies can be defined as marketing variables (which includes but not limited to promotions, price and product range) that can be utilised to satisfy a target consumer segment which leads to profit (McCarthy, 2011). This can significantly increase the customer awareness and perception of the retailer’s values (Burt, 2000) and this also correlates with the chance for the customers to become more loyal with the brand. The retailer can determine a suitable marketing mix, which contains multiple instruments that serve the purpose of establishing and maintaining retail consumer’s patronage (Hogreve et al., 2017).

Along with appealing product price, convenient services and corporate brand, marketing strategies such as communication with the customers via personal selling is one of the factors that has the strongest impact on customers satisfaction (Blut et al., 2018), while price deals and spending advertisements are proven in literature to be effective in the creation of customer loyalty for department stores and hypermarkets (Li et al., 2012). Marketing strategies are recommended to have loyalty programs and especially price deals, which makes customers visit the stores more often and increase spending. Another common method of improving the relationship with customers is to use relationship marketing tactics such as direct mailing, royalty offers/programs and the usage of coupons (Verhoef, 2003).

Generally, Direct marketing is defined as an advertising tactic in which a specific individual or business is the main target. This allows the retailer to build a personal relationship with the shopper/client without costly traditional methods include newspaper or TV advertisement (Business Queensland, 2020). The characteristic of a direct marketing campaign requires customers to act via a variety of methods: receive additional information, make extra purchases or visit the website. It is the goal of this tactic to create a first-step action or response for customers to engage with the retailers.

Direct marketing’s major benefit is to trigger growth in revenue from the frequent and current consumers via quantity and brand appeals. Within retail, direct marketing usually involve in coupons and promotions used together, accompanying an advertising message (Lund and Marinova, 2014). Direct marketing is favoured by retailers to be used on specific consumer segments, or individual shoppers who don’t like mass marketing (Steel, 2008). Direct marketing, when used on consumers and especially those who aren’t frequently exposed to this tactic, has a significant positive effect on overall sales (Venkatesan and Farris, 2012). Direct marketing, when done correctly, will not only build relationship with the consumer but also reflect the most suitable strategy for a given segment, test the attractiveness of a product/service and trigger sales growth (Subramanian R., 2017). This helps the business become more flexible and responsive to the market needs.

By using loyalty programs, food retailers have a high chance of maximizing their profits (Pauler and Dick, 2006) as it boosts cumulative buying or repeat behaviours. This is valuable in the retail industry that has low purchase frequency and supplier distinction (Bhattacharya and Sen, 2003). Another purpose of royalty programs, such as royalty cards, is to collect demographic and transactional data of customers in order to be utilized in the design of more profitable and personalized deals (Beenstock, 1999).

* 1. Application of data science in targeted marketing enhancements
     1. Overview of Data Mining in the retail context

Throughout different aspects of retail operation, Data Mining along with machine learning and its application plays an essential part that aids businesses in achieving its objective. Data mining can be defined as a procedure within the Knowledge Discovery in Databases process (KDD), which focuses on extracting information and unknown insights (such as unusual records, dependencies or data records) from a vast amount of information which is then made into a comprehensible format that serves future uses (Kamber and Han, 2011). Specifically, data mining can build models as well as analyse data and extracts knowledge depending on the business problems and objectives of the business and aid the them in understanding themselves and making decisions. Another application of data mining is to help a business predict future statistics, events, behaviours and trends that can be used to support business planning and even managing risks.

Data mining can be utilized effectively in retailing and its application is various, especially when combined with the vast quantity of data collected which would be wasteful if not exploited effectively and fully. It can give the retailers a competitive advantage especially when it comes to information and aid them in managing shopper retention, customer segmentation/ranking as well as identify cross-selling or up-selling opportunities due to the natures of the technique involve with it, including descriptive/predictive modelling and forecasting (Madan L., 2006).

Data mining can also be used to optimize retail stores. Classification rules such as Decision tree can be used as a simple and visually appealing method to find out the in-store attributes which are relevant to focus on in order to increase a store’s service quality (Min, 2006). A retail outlet’s features can also be examined practically and scientifically by reducing the quantity of store attributes into a smaller yet more meaningful and critical set of dimensions. Customer profiling done via data analytical technique such as cluster analysis and Chi-squared analysis allows retailers to identify which range of products are best placed and arranged in store that is appealing to a target segment and maximize the chance of consumer impulsive buying.

A retailer can also use demographic information derived from data mining to manage its logistics by identifying which product is the “hottest” to a consumer segment in a particular area which allows them to identify best-selling location, arrange merchandise stocking and decrease inventory movement (Ahmed, 2004).

* + 1. Data mining’s application in CRM and Targeted marketing

The most basic characteristic of Data Mining in CRM is in its capability of analysing the huge amount of data collected from various channels and extract most relevant pieces of information and insights relating to the customers which helps firms the demand of market segments and manage customer lifecycle stages. Data mining can play an important part in the customer relationship cycle, which includes: Customer obtainment, increasing consumer values and retaining consumers (Mittal, 2001).

With high competition, the retailer can use data mining to its advantage especially in acquiring and retaining consumers. Shopper’s past purchasing behaviour can be examined based on the transaction data collected within the database of the retailer (Madan L., 2006), while a consumer’s future behaviour and patronage can be predicted using data mining techniques such as decision trees which uses “IF-THEN” statements to establish prediction. This gives the retailer an idea of what marketing instrument is most appealing to the consumers on a personal level and leads to easier consumer obtention.

Using techniques such as Artificial Neural Network (Kim, 2006), logistic regression (Hwang et al., 2004) or Support vector machines (Archaux et al., 2004) data mining can also make customer retainment easy by modelling the behaviour of churned customers in the database and use it to predict consumers who has a high probability of switching to an alternative retailer in the near future (Chopra et al., 2011). This allows the organization to design appropriate marketing strategies and campaign to maintain consumers and save the cost of finding new ones and exploring new territories.

An increase in effort to obtain new customers also means an increase in the usage of data mining tools to filter valuable customers and specially to improve direct marketing campaigns. One of the reasons why data mining has becoming more and more common in the retail industry is due to business communities having the need to better understand customers to enhance the direct marketing process (Bose and Mahapatra, 2001), and data mining is an excellent tool to gain a competitive edge.

Direct marketing can be performed based on the usage of data mining software and techniques which utilizes the raw transactional data. For example, customer service representatives are equipped with customer profiles resulted from data mining that allows them to know what products/services are potentially appealing to the targeted consumers, and at the same time reducing the marketing cost and increase efficiency since the target of direct marketing will usually be a small subsets of customers that are found to behave similarly by the data mining algorithm (such as clustering and associations), rather than a large number of consumers (Chopra et al., 2011).

In direct marketing, besides knowing when and how to use it, it is also important to know which customer is the most suitable for a target marketing campaign (Elsner, Krafft and Huchzermeier, 2004). Data mining and machine learning can be used to determine this particular aspect. Data mining allows predictive modelling which searches for shoppers who have a high probability of responding to a promotional programmes and offers by using statistical models such as least-squared regression, logistic regression or neural network (Hand, 1981) to attach a score that characterize them than proceed to target those with the highest scores (Bhattacharyya, 2000).

As mentioned previously, customer behaviour and direct marketing have a close relationship, since knowing the customer preferences will lead to easier consumer identification for directed marketing. This is reflected in basket analysis as insights drawn can be used by marketing managers to design targeted marketing actions that is included in the loyalty programs of a retailer. Specifically, retailers can encourage the customers with loyalty-card at each transaction by using discounts and self-check offers (Passingham, 1998). The barcode on the royalty cards, combined with the Point-Of-Sale technology, allows the retailers to obtain a huge amount of transactional data relating to what each individual customer put in their shopping cart and even multi-categories purchase behaviours (Boztuğ and Reutterer, 2008). Past consumption behaviours of consumers can be drawn from the basket analysis, and retailers can offer the personalized offers for specific categories of items that they are most likely to buy based on the association rules , leads to an increase in spending (Nisbet et al., 2018).

* 1. Summary

The second chapter set the foundation for the dissertation, as insights from previous researches in the retail marketing context offers great insights on various methods of consumer analysis.

Overall, as the retail industry is highly competitive due to constant development of online shopping and discounters, retailers must achieve competitive advantage by tailoring their business to customer preferences. Understanding their consumer via segmentation is a crucial first step in achieving this advantage. Besides the frequently used RFM method, different data mining techniques such as clustering with K-mean analysis can also be an efficient method of determining consumer who behaves in the same way. Consumer behaviour is also the second valuable information which can be achieve via using association rules and basket analysis to determine purchase choice.

A lot of factor contributes to the decision a shopper makes in their purchases, such as store attributes, price image and convenience. However, one factor that affects their behaviour is in the marketing campaigns they are exposed to. This not only attracts consumers but also improve overall sells significantly. One exemplary tactic is direct marketing, which a personal relationship is built with a specific targeted type of consumer. Data science helps immensely in this particular aspect by analysing huge amount of data from various channel for future consumer behaviour. Artificial Neural Network, logistic regression and support vector machines can be used to detect consumer churn while decision trees give the idea of a particular consumer’s buying pattern which offers the retailer more ways of developing their campaigns.

**Word count: 4,187**

# Chapter 3: Methodology

## 3.1. Research philosophy

To identify the correct philosophy as a basis for the research, it is important to examine paradigm and ways of viewing them. A paradigm is a set of common agreements as well as beliefs which are commonly shared between researchers and scientist about how to understand and address the problems (Kuhn,2015). There are multiple paradigms that can be used in research, but the two most common paradigm in research are Constructivist and Positivist. In order to determine which paradigm is the most suitable, it is important to look at the perspective of each. Specifically, the main components of research paradigms include Ontology and Epistemology. Every paradigm is based upon an ontological and epistemological assumption; hence the two perspectives can then be used as a basis for paradigm selection.

Ontology considers whether the central question/social entities should be viewed as objective or subjective. Positivist paradigm has an objective point of view which suggests that the existence of objects and entities are independent from social actors (Cohen et al., 2007), unknown to researchers (Pring, 2000) and not affected by external factors and senses (Scotland ,2012). In the context of the research, the primary concern with big data analytics and machine learning’s effectiveness in retail behaviour prediction and information extraction is highly objective. It is independent from any social factors and external influences, and the researcher does not know whether the effectiveness exists or not. It is dependent on the usage of data analytical techniques and optimisation of the prediction models for the conclusion of the existence to be made. Hence, the research adapts a positivist ontological position.

When it comes to the epistemology assumption, the positivist paradigm is also dormant in the research. With the entity’s meaning resides in itself rather than in the conscience of the researcher, it is up to the researcher to use reliable and valid tools to obtain them via research. Specifically, the aim of the research is to explain how data mining can be used in retail and predict customer behaviour based on direct marketing. Datasets will be examined in a detailed manner using different data analytical techniques which will result in crucial insights regards to consumer demographic and behaviour that can be inspected with the hypotheses of literatures taken into consideration. Moreover, the possibility of using machine learning to predict customer’s response to direct marketing efforts can also be measured using different evaluation criteria. By using observational and measurable facts along with scientific methods, causal explanation and predictions, the research’s epistemological stance is also Positivism. Therefore, the research’s paradigm is Positivism.

## 3.2. Research approach and analytical method

### 3.2.1. Research approach

Generally, there are two types of approaches in research: deductive or inductive. The deductive approach’s goal is to test and develop hypotheses based on existing theories in literature (Wilson, 2019). It begins with known patterns and hypotheses in which their validity is tested against observations within given circumstances, and as a result a rejection or acceptance of the hypotheses can be concluded (Snieder and Larner, 2009). The literature has many in common with the deductive approach. The literature starts with known patterns and hypotheses in literature involving with data mining’s application in consumer information extraction, behaviour examination and marketing response prediction. By extracting crucial and relevant information within the collected data in the research, the conclusions derived from the analysis and evaluation will either confirm or reject the effectiveness. Therefore, the research follows a deductive research approach.

### 3.2.2. Analytical method

The research adapts a quantitative method. Specifically, it follows a positivist and objective research philosophy with a deductive approach. The researcher believes in the existence of only one reality that is objective as it is independent from social actors (Cohen et al., 2007) and the researcher (Pring, 2000). The quantitative method examines determined theories in literature by testing theories that can be measured numerically, analysed with statistical techniques and researcher can make generalized predictions of the theory that holds true (Creswell, 1994). A quantitative research can end up with large quantity of output in which the researcher has to give meaning to (Choy, 2014). The researcher must then select data that most relevant with the question of interest, as well as selecting appropriate statistical methods to describe and examine trends/relationship in the data (Saunders,2009).

The quantitative method is consistent with the nature of the literature in which collected data involving customer transactions, demographic and product information as well as detailed marketing campaigns strategies deployed can be used to explain how data analytics can be exploited within the retail context and most importantly predict whether a customer will respond to the marketing campaigns directed to them by using the insights discovered in the aforementioned data mining procedure. Customer data in the research are measured numerically as well as analysed using different statistical techniques which produces new outputs that can be used in classification models to make prediction of customer behaviours that is expected via literature. The literature also follows a deductive approach as it reinforces known theories in other literature that involves with data science in retail, which is also consistent with the quantitative research method. Therefore, the literature will adapt a quantitative research method.

## 3.3. Research methodology

### 3.3.1. Overview of research process

In order to efficiently extract information from data, the data mining process will closely follow the Cross Industry Standard Process for Data Mining approach, or further known as CRISP-DM. Specifically, it is a framework model in data mining that consists of multiple phases and tasks. The main phases of data mining include: Business understanding, data understanding, data preparation, modelling, evaluation and deployment, which is seen in figure 1 (Wirth and Hipp, 2000).

Table

Description automatically generated

Figure 1. The CRISP-DM process for data mining

Firstly, it is important to understand the business in question, the overarching problem of the business that needs solving and the nature of the dataset. The dataset belongs to a real-life retailer, although the name of the business is not clarified for privacy preservation. Regardless, the nature of the data and the variables available in the dataset are very well structured and provide a strong basis for analysing the insights and relationship between consumers and marketing strategies within the retail context. The dataset is also collected by Dunnhumby for the same purpose. Hence, the overall goal is to extract knowledge from the dataset with regards to consumer insights and advertising efficiency, followed by predicting consumer participation in directed marketing.

The second phase involves collecting and understanding data. This starts with specifying how the data is initially collected and overview of the dataset and detect important subsets for the classification problem at hand. The section 3.3.2 will further clarify type of data, where it was collected and an overview of the entire dataset.

The third phase, covered in section 3.3.3 and 3.3.4, will further explain the subsequent phase which is data preparation. This covers multiple activities executed to construct the final dataset for classification in order to provide best results. This involves selecting relevant variables, create new variables through different statistical methodology and re-coding values.

Once when the data is prepared, modelling will be the next phase. The relevant machine learning models will be selected in section 3.3.5, and parameters of the selected models are tuned to provide optimal accuracy in predicting consumer participation. This is explored in detail in Chapter 4, section 4.3.

Chapter 5, section 5.2 includes an evaluation of how the models perform besides training data, and its effectiveness in applying for real world data. The section covers the last of the two phases: Evaluation and Deployment. Limitations and recommendations are also suggested in Chapter 6.

### 3.3.2. Data collection and understanding

#### 3.3.2.1. Data collection

There are two main categories of data that can be used in research: Primary and Secondary, in which the data belongs in this dissertation belongs to the latter. The data is from another sources rather than collected by the researcher (appraisal institute,2015) and was primarily collected in the past by other researchers and then made available for re-use by other individuals (Hox and Boeije, 2005). Therefore the data used for the research is secondary type as it focuses on an already available dataset, which includes a large amount of data and variables that serves the purpose of reinforcing effects of direct marketing to customers in retail.

The secondary data used in the literature is collected by Dunnhumby. Dunnhumby is an international data science company which specializes in utilizing customer data to help businesses in many aspects (About us, 2020). The company also has resources include Source Files, which is a data-sharing platform of real-world source-files that contains a multitude of datasets that varies in nature and sizes. Among the offered source files include The Complete Journey, which includes many smaller data with different dimensions. The dataset is chosen due to the details, complexity and size.

#### 3.3.2.2. Data description and understanding

The Complete Journey dataset belongs to a real-world retailer with all transaction records of 2,500 households that spans over two years, as well as other complementary data such as demographic information, product mailing/ display data and especially direct marketing contact history for a selected number of households. The dataset’s implication not only spans in a classroom setting, but also ideal for academic researches relating on the effect of direct marketing to shoppers. Figure 2 indicates the structure of The Complete Journey dataset.

Graphical user interface, diagram

Description automatically generated

Figure 2. The Complete Journey dataset table map (Source Files - dunnhumby, 2020).

There are two main categories of data: Data tables and Look-up tables. Data tables are the primary data which contains crucial information relating to customer transactions as well as demographic info and the list of sent marketing campaigns with recipient households. The Look-Up tables offers complementary data that serves to give extra information to the main tables. This includes campaigns duration, coupons redeemed along with eligible products as well as the information of all products sold by the retailer and how they were displayed in stores and sent in mailings.

### 3.3.3. Data cleaning

In order for classification to be executed, the dataset must first be cleaned. Relevant variables must be selected and created in order to form a final dataset that is then used for the classification model in the machine learning procedure. The whole data preparation and classification modelling process will be done using R coding language along with relevant packages in R Studio software.

The dataset is first imported and cleaned. For all datasets, sanity checks are performed. For the literature in particular, sanity check is performed to make sure that numeric variables remain logical, and that all products and campaigns are present in the retailer’s database. In transactional data, numerical variables such as dates, sales value and quantity are checked to make sure no values are negative. For marketing related data, such as campaign information and coupon redemption, the campaign and redemption start date must be positive and smaller than the end date, while the end date must not exceed 730 which is two years- the observation duration of the data.

All data are also checked for missing values as well as duplicates. Missing value check is performed on all data and most returns no missing values, with the exception of Product information data. 15 products have no information hence they are removed from the product information dataset.

Some values also need to be recoded in order to better indicate information. This is mostly done in Demographic data, in which some values either not having appropriate levels, names or data type. Specifically, values in “income” variable has been recoded to have appropriate levels while “Unknown” values in the column that specifies number of kids are recoded to 0 and “household size” variable is converted to numeric with special symbols in certain fields removed. The transformation is also observed in the transaction dataset since a few value in the “retail discount” variable is at positive, which is not logical since discounts deduct a transaction’s value rather than adding to it. Hence, all positive amounts in this variable are converted to negative.

Finally, extra demographic data needs to be generated. The demographic information is not entirely available for all households in the dataset, in order to preserve privacy. Therefore, extra demographic data is synthesized so that all households will have demographic information, which is crucial for classification modelling. The synthetisation is done in R Studio software using the “synthpop” package, which aims to fit the data to the original distribution and obtain estimates of its parameters (Nowok, Raab and Dibben, 2016). The synthesized data’s distribution is finally checked and compared with the original dataset to make sure it maintains unchanged characteristics.

### 3.3.4. Data pre-processing

#### 3.3.4.1. First stage: Collect in concern and create variables for final dataset

Once when the data tables are cleaned, the next procedure is to create the final data for classification models. The goal of the literature is to predict whether a customer will respond participate in any given marketing campaign, by looking at all households in the observation and the number of households that actually participates in a marketing campaign.

The complete journey dataset includes 8 different data tables. In order for the classification to work, predictors relating to a household’s personal demographic, transaction and behaviour information must be created. Therefore, it is crucial to extract relevant data tables from the dataset, which is then combined together and allows for the creation of additional fields from the variables. Figure 3 illustrates the final dataset’s variables and its components.

A picture containing screenshot

Description automatically generated

Figure 3. Final data used for classification model building and other data components

Figure 3 shows the final data consists of three components that acts as predictors. The components include each household’s demographic data, information of most purchased product and additional spending/ shopping data. The final component, participation, is the output variable and the main focus of the models.

The first seven variables are demographic data. It consists demographic information of all households. This is constructed by finding ID of household in the transaction data, which is then matched with their corresponding demographic information in the “household demographic” data.

The next six variables contain the information of each household’s most frequently purchased product. Transaction data is first combined with product data, so each purchase in the transaction data has detailed information of product characteristics. This is followed by finding the most frequent type of product purchased for each household in 2 years of observation. A data of most frequently purchased product for each household is created as “purchase behaviour” and it is linked with demographic data matching by household key.

The following six variables indicate additional purchase data of customers, which uses only transaction data as basis. Mean basket value is calculated by extracting the means of all basket values for each household, and weekly spending can be calculated by finding the spending total of each week for each household and then averaged. Trip summary is the total number of trips each household made in 2 years. Households with over 100 trips are indicated as “YES” in the “frequent” variable, and vice versa for households with less than 100 trips made. Finally, transaction time is calculated by converting time of transaction to time of days (such as morning or evening), which allows the discovery of most frequent time occurrence. Consequently, additional transaction information are grouped together as “additional purchase data” and it is linked with the aforementioned demographic and purchase behaviour data by household ID.

The final predictor variable “participation” indicates whether a household participates in a campaign (by redeeming coupons). Value of 1 indicates the household participate in any marketing campaign, and “0” indicates vice versa. This is found by extracting ID of households that redeem coupons of any given campaigns from the “redemption” data, and a separate “participation” column is created which contains only value of “1”. This household ID is then matched with the household ID of the final dataset, and household with no matching household ID have the participation value of “0”.

#### 3.3.4.2. Second stage: Pre-process final dataset for classification modelling

Once when a final data is gathered for classification, the final step is to do the necessary transformations and pre-processing to have an accurate classification results.

With only 434 households who participate out of 2,500, the final dataset must first be balanced in order to give equal priority to each class that is predicted. If not balanced, the classifiers will become more biased towards the majority class and performs poorly on predicting the minority class. Synthetic Minority Oversampling Technique (SMOTE) is the balancing method of choice as that it synthesizes brand new observations of the minor class instead of simply create replacements (Chawla et al., 2002) or reducing samples for a dataset that is already small in size. Using smote, the class achieves a more balance ratio between minority class and majority class, indicated in table 1 below.

Table 1. Balance of the data before and after using SMOTE



For factor variables, four distinct variables: Department, Brand, Commodity and Sub-commodity description are highly cardinal as each variable all have a significant number of unique levels with too little observations. In a small sized data of only 3,582 observations, the cardinal data can be an issue as it not only increases processing time significantly but the data will also have extremely rare or infrequent categorical levels, reduced model interpretability as well has high dimensionality since typically R will transform factor levels into dummy variables, thus adding one dimension per unique factor level (Mount and Zumel, 2018). Having levels with little observations means a reduction in predictive value hence it needs to be dealt with accordingly. By keeping all categories with at least certain percentage of representation in data (more than at least 1% of observations) and rename the categories that have less than 1% of presence in data to “others”, the factor variables will have a significantly more manageable number of categories. Table 2 shows the high cardinal variables before and after transformation.

Table 2. High carinal variables before and after transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | Description | Number of levels before grouping | Number of levels after grouping |
| SUB\_COMMODITY\_DESC | 16/266 levels take up 70% of total observations. The remaining levels have less than 1% observations | 266 | 16 |
| MANUFACTURER | 16/444 levels take up 70% of total observations. The remaining levels have less than 1% observations | 444 | 20 |
| DEPARTMENT | 4/13 levels take up 90% of the total observations. Remaining levels have less than 1% observations. | 13 | 6 |
| COMMODITY\_DESC | 12/126 levels take up 70% of the total observations. Remaining levels have less than 1% observations. | 126 | 17 |

Once the high cardinal categorical predictors are transformed, the next step is to conduct feature selection to remove irrelevant predictors. Chi-squared test of independence can be used to examine whether the categorical predictors have any relationship towards the target variable. A P value less than the chosen alpha value of 0.05 means that the null hypothesis is rejected, whereas a P value higher than the alpha of 0.05 means that the null hypothesis is accepted. Specifically, the hypotheses of the Chi-squared test of independence is as follows:

1. Null hypothesis (H0): The chosen categorical predictor has no relationship with customer participation in a marketing campaign.
2. Alternative hypothesis (H1): There is a relationship between the categorical predictor and whether a customer participate in a campaign.

The Chi-squared test of independence is conducted for all categorical variables, which returns the following values:

Table 3. Chi-squared test results for categorical variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | X-squared | Degrees of Freedom (df) | P value |
| AGE\_DESC | 27.127 | 5 | <5.389e-05 |
| MARITAL\_STATUS\_CODE | 76.477 | 2 | <2.2e-16 |
| INCOME\_DESC | 127.19 | 11 | <2.2e-16 |
| HOMEOWNER\_DESC | 199.83 | 4 | <2.2e-16 |
| ADULT\_NUM | 0.81617 | 1 | 0.3663 |
| KID\_CATEGORY\_DESC | 16.164 | 3 | 0.00105 |
| MANUFACTURER | 133.22 | 19 | <2.2e-16 |
| DEPARTMENT | 19.605 | 5 | 001482 |
| BRAND | 2.3139 | 1 | 0.1282 |
| COMMODITY\_DESC | 162.01 | 16 | <2.2e-16 |
| SUB\_COMMODITY\_DESC | 186.48 | 15 | <2.2e-16 |
| FREQUENT\_TIME | 153.43 | 3 | <2.2e-16 |

The Brand and Adult number variable have a P value of 0.3 and 0.1, which is higher than the chosen alpha level of 0.05 and the null hypothesis is accepted. This means that knowing the brand that the consumer usually purchases from and number of adults per household has no effects on whether they will participate in a marketing campaign or not. Therefore, Brand and Adult number should be removed from the dataset to improve prediction.

With numerical variables, features with near zero variance will be removed as they often have less predictive power and including them sometimes causes models to crash. Near zero-variance variables have a few unique values with low frequencies as the fraction of these values over the sample size is low, and the ratio of the frequency of the most occurring value and the second most occurring value is large (Kuhn and Johnson,2013). Upon further examination, the variable Quantity is categorized as near-zero variance (shown in table 4) and should be removed.

Table 4. Features with near-zero variance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | freqRatio | percentUnique | zeroVar | Near zero |
| QUANTITY | 23.0625 | 2.909482759 | FALSE | TRUE |
| MEAN\_BASKET\_VALUE | 1 | 90.55316092 | FALSE | FALSE |
| WEEK\_SPEND | 1.333333 | 94.93534483 | FALSE | FALSE |
| TRIP\_SUM | 1 | 51.43678161 | FALSE | FALSE |
| TRIPS\_PER\_WEEK | 1.451138 | 23.16810345 | FALSE | FALSE |

Removing highly correlated variables also proven effective in improving a dataset as it helps avoid multicollinearity. Table 5 shows trip\_sums variable is highly correlated with trips\_per\_week with a value of 0.79. Trip\_per\_week will be removed due to high correlation with trip\_sum and it is less important, which is shown by using random forest to determine feature importance. Figure 4 shows random forest algorithm’s ranking of the predictors based on their predictive value. And it further reinforces the removal of variables Brand, Adult\_num, Quantity as these variables are the least important, and Trips\_per\_week is not as important as trip\_sum. Therefore, Brand, Adult\_num, Quantity and Trips per week will be removed from the dataset.

Table 5. Correlation between numeric variables



*Table

Description automatically generated*

Figure 4. Importance of chosen variables

With the transformation and removal of irrelevant variables, the dataset is finalized for classification models. The dataset is split into two separate training and testing dataset with a ratio of 70:30. All modelling process is done on the training set, and the models are used to predict the outcome in the test set.

### 3.3.5. Classification model selection

To select models, it is important to do spot checking to determine with models will have the best performance by implementing various algorithms with the default parameters on the dataset. Models with the best performance will be chosen for further tuning in classification. In the literature, 3 models with the best Receiver Operating Characteristic (ROC), Sensitivity and specificity will be chosen. The models tested has the performance shown in figure 5 below.

A picture containing table

Description automatically generated

A picture containing person, person, people, large

Description automatically generated

A close up of a person

Description automatically generated

Figure 5. Performance of various models on the dataset.

Looking at the results of the spot-checking process, it is seen that among all algorithms selected and fitted on the dataset, ensemble algorithms are the most accurate. Ensemble learning refers to the training of multiple baseline learners trained over the same data (Lopez-Gracia et al.,2019). The chosen algorithms in this category include Random Forest, Boosted C5.0 Decision Tree and Bagged Decision Tree that, when implemented on the dataset, provides sufficient mean ROC, sensitivity and specificity. This can be seen in figure 5.

Boosting and Bagging are two exemplary algorithms within ensemble learning. In boosting, early learners fit simple models to the data and it is followed sequentially with other learners that aims to analyse for error (Garg, 2018). The latter trees add extra weight to inputs that are incorrectly specified which results in weighted vote taken for prediction (Liaw and Weiner, 2002) so that it will correctly identify the input. Therefore, this can be effective as a succession of multiple weak classifiers turns into a strong one, which is highly accurate compared to a data’s best weak classifier (Quinlan, 1996). Bagging, on the other hand, trains multiple base learners on bootstrap samples which consists of the dataset’s subsamples. Finally, most voted class is determined via majority voting of the base learners (Zhou, 2009).

Random Forest, introduced in 2001 (Breiman 2001) features a combination of bagging on samples (Breiman 2001) and making decisions based on random subsets of variables (Ali et al., 2012). In essence, it draws a number of bootstrap samples from the dataset, and in each sample an unpruned tree is grew with the best split among the predictors. Finally, the prediction is conducted by aggregating the prediction of all trees pruned via majority votes (Liaw and Weiner, 2002). The algorithm has many advantages which includes its ability in bypass overfitting issues as well as being easy to set up and simultaneously provides feature importance which allows any researcher to identify crucial variables (Horning, 2013). Especially, it can also work with data that is small sized with high dimensional feature spaces (Biau and Scornet, 2016) hence make it useful for this particular dataset.

Therefore, Random Forest, Boosted C5.0 Decision Tree and Bagged Decision Tree will be chosen for classification.

### 3.3.6. Modelling methodology

The chosen models will be implemented on the training dataset and tuning of hyperparameters is required to observe for a change in efficiency. Each model’s performance is examined on two different stages. In the first stage, model will first be executed on the training dataset using the default parameters and then used to predict the test dataset. The model’s performance metrics on the test set, which also uses optimized classification threshold, are recorded and used as a comparison for the models after tuning. The second stage involves finding the best hyperparameters which are then implemented on the models. The models produced in the second stage is evaluated using ROC/Area Under the Curve (AUC) and confusion matrix. The classification threshold is also determined so that it maximizes the sum of specificity and sensitivity. The output of the tuned model is compared to the default model for any improvements in quality.

To evaluate the classification models, three different metrics are the main priority, assuming that “positive” refers to customers who are likely to participate in a campaign and “negative” indicates customers who aren’t interested and won’t likely to participate.

1. Sensitivity, or “True Positive Rate” is the rate of which the positive cases are correctly classified. In the context of the literature, it is crucial to correctly determine which customer will participate (positive). The model with high Sensitivity is considered a well-performing model and is the most prioritized metric.
2. Specificity, or “True Negative Rate” and “False positive rate” must also be taken into consideration. True negative rate is the rate of negative cases correctly identified. As sensitivity increases, the false positive rate will also increase and more customer who are not interested will be wrongly classified as participating in a campaign. The best model is the one with a balanced sensitivity and specificity, so that it can correctly identifies interested customers without letting the retailer waste resources on customers who are not attracted to marketing efforts.
3. ROC curve and AUC: ROC indicates the performance of a classification model in all classification thresholds by mapping the ratio of True Positive Rate to True Negative rate, in which a classification threshold that provides the best balance of the two will be determined and will provide the best results. AUC measures the probability of the model successfully distinguish between a positive and a negative class. The model with a high AUC is an efficient model.

All modelling and tuning are done with Caret package in R studios.

## Summary

The first half of the chapter explains the main research philosophy used in the literature. The research adapts a positivist philosophy which aims to prove the existence of the “effectiveness” of machine learning in consumer insights discovery. The research is quantitative with a deductive reasoning method.

The latter half of chapter 3 approaches and different methodologies used for the literature are also discussed. Specifically, the process of data collection, cleaning and pre-processing for machine learning classification is explained in detail. The models chosen, as well as optimization procedure and evaluation metrics are also discussed.

**Word count: 4,400**

# Chapter 4: Findings and classification results

## 4.1. Finding and analysis of findings

### 4.1.1. Customer characteristics

According to Figure 6, the majority of the households are between low and medium income. Most households earn between $35,000 to $74,000, in which more households are in the medium bracket ($50,000 to $74,000) compared to the lower medium bracket ($35,000 to $49,000). This is followed by the low to medium brackets, in which 26% of households earn between $15,000 to $34,000. The medium to high earning households between $75,000 to $99,000 takes up 12% of total. High earning households between $100,000 and $250 are the minority of the total customer base, which takes up 16.07% of total customers.

Chart, bar chart, histogram

Description automatically generated

Figure 6. Visualization of household’s income composition

According to Figure 7, the majority of households ages between 45 to 54 years. The households age between 35-44 are the second most common with 24.22% of total, followed by the younger age bracket of 25-34 years old. Households with older consumers are more common than younger customers which only takes up of approximately 6% of total consumer base, compared to the total of 16% that ages over 55.

Chart, bar chart

Description automatically generated

Figure 7. visualization of customer age composition

Table 6 indicates that there are significantly more household that consists married individuals rather than singles. However, marital status is not only unavailable to all customers, but also mostly unknown even for customers with demographic information. Within 801 households with demographic details, up to 43% of customers don’t have any marital status known. Hence, the actual marital status distribution of the retail store cannot be accurately represented in the data with such little information. Nevertheless, it can give insights to the retailer that there are more married than single household.

Table 6. Marital status of households

|  |  |  |
| --- | --- | --- |
| Marital status | Number of households | Percentage |
| Married | 340 | 42.45 |
| Single | 117 | 14.61 |
| Unknown | 344 | 42.95 |

The lack of data can also be seen for the homeowning status, although not too severe with only 29% of households has unknown house owning data. For the households with known data, a significant amount of over 60% of households are homeowners, followed by renter with 5.24%. Households with a probability of being either a renter or houseowner only takes 1.37% each. Table 7 further shows this distribution.

Table 7. Homeowning status of households

|  |  |  |
| --- | --- | --- |
| Homeowning status | Number of households | Percentage |
| Homeowner | 504 | 69.92 |
| Unknown | 233 | 29.09 |
| Renter | 42 | 5.24 |
| Probable owner | 11 | 1.37 |
| Probable Renter | 11 | 1.37 |

Next, it is useful to look at household composition which is shown in figure 8. The majority of households consist of only 2 adults. For household with children, there are significantly more households with two adults than households with only one adult. For household with only one individual, the gender composition can also be determined as there are slightly more female customers than male. The lack of data can once again be observed in 9% of total households.

Chart, bar chart

Description automatically generated

Figure 8. Visualization of household composition

### 4.1.2. Customer purchase behaviour

Looking at transaction data, it is possible to examine the overall behaviour of all households when making purchases. By constructing new variables from already existing transactional data, distinctive purchase features for each household is extracted.

Knowing when the customer usually purchase can be informative. By looking at unique basket ID and transaction time for each households, it is determined that half of all transactions occur in the afternoon, between 12pm and 6pm, while proportion of purchases in the morning (between 6am and 12pm) and evening (6pm until midnight) is the same with 23% of total transactions. Transactions during the midnight/early morning (between 12am and 6am) are the rarest, account only 2% of total transactions. It is therefore determined that customer most frequently purchases in the afternoon. Figure 9 further demonstrates this pattern.

Chart

Description automatically generated

Figure 9.Visualization of most common transaction time.

As for number of trips per week for each household on average, an expected pattern can be observed as the majority of customers shop once or twice per week. As shown in table 8, only 10% of people shop three times per week, and only 6% of people shops more than three times per week.

Table 8. Number of weekly trips and percentage

|  |  |  |
| --- | --- | --- |
| Trips per week on average | Number of households | Percentage |
| 1 | 1,050 | 42 |
| 2 | 1,049 | 41.96 |
| 3 | 236 | 9.44 |
| 3+ | 165 | 6.6 |

Spending information can indicate a lot of information about the household’s behaviour. On average, households spend $31.48 per trip, with the majority spends between $18.31 to $40.37 and the maximum of up to $165 per trip on average. Hence a right skewed distribution is spotted. The same patterns can be seen in terms of spending per week on average. Most households spend between $31 to $72 per week and $56.42 on average, right skewed distribution is also present with the maximum spending per week of up to $455.25. This is most likely influenced by the minority of high spending households in the observations. Figure 10 and 11 further shed light on the pattern.

Chart, line chart

Description automatically generated

Figure 10.Visualization of spending per trip of households

A picture containing graphical user interface

Description automatically generated

Figure 11. Visualization of spending per week of households

In terms of income and age, it is also very insightful to find out which type of household has the highest spending per shopping trip and spending per week.

As for age, although the majority of consumers are between 45 to 54 years old (according to figure 5), the average spending per trip and per week is the highest for people aged between 35 to 44 years old. While spending of the 45-54 age bracket is almost the same compared to the 25-34 age group. Figure 12 indicates spending and shopping behaviour of each age group.

Graphical user interface, chart, scatter chart

Description automatically generated

Figure 12. Average spending per trip and per week of each age group.

According to figure 13, although most of customers have low to medium income, the households in the higher income bracket spends more on their purchases and per week.

Chart, scatter chart

Description automatically generated

Figure 13.Average spending per trip and per week for each income group

Finally, determining what each household buys the most also allows the retailer to understand more about each consumer, which will help retailers design a suitable target marketing strategy. By combining transactional and product data, it is possible to extract the most frequently bought products for each households in the duration of 2 years. For example, figure 14 below indicates the information of the product departments that are most liked for all 2,500 households.

A picture containing application

Description automatically generated

Figure 14.Visualization of most liked products

Figure 14 further indicates that up to 75% of all customers purchases grocery most frequently when they visit the retail chain. This is followed by dugs/general merchandise and produce with 11% and 7% respectively. The rest of the product categories only takes up of 5.92% of total of most liked products. This could also be explained that grocery is a necessity, hence household most frequently purchases them.

A further examination into the sub-categories of the most liked products gives further information about what specific product type is a frequent purchase for consumers. As for grocery, products like milk, soft drinks and frozen meat/meat dinner are the most liked products among all households, further indicated in figure 15. Similarly, the customer preferences in drug/general merchandise and produce department can also be found. For drugs and general merchandise, baby food is the most commonly bought repeatedly by households followed closely by candy (purchased at the check-out lane), as well as cigarettes and magazine/newspaper. For produce, tropical fruit is most widely liked. A point needs taken consideration that unlike grocery, the number of customers who frequently purchase drugs/general merchandise and produce is significantly smaller, hence the observation can’t be generalized to the real population unless more data is available. Nevertheless, categories of well-liked drugs/general merchandise and produce goods can be somewhat estimated, which is shown in figure 16 and 17 respectively.

Chart, bar chart

Description automatically generated

Figure 15.Most liked grocery products

Chart, funnel chart

Description automatically generated

Figure 16.Most liked drugs and general merchandise

A picture containing text

Description automatically generated

Figure 17.Most liked produce goods

### 4.2.2. Marketing information and effectiveness

A first step in examining how effective marketing campaigns are is to understand what strategies are being deployed.

In general, there are 30 unique campaigns sent to 1,584 households throughout the duration of the observation. The campaigns are divided into three different categories: Campaign type A, B and C. As aforementioned, it is not specified the details and characteristics of the campaigns or the methods of how they are deployed and sent. But it is possible to determine the most sent ones as well as other background information that can potentially reveal a connection to marketing effectiveness. According to table 9, more than half of the campaigns sent belongs to type A, followed by type B and C with 36% and 7% respectively.

Table 9. Quantity and percentage of campaigns sent by each type

|  |  |  |
| --- | --- | --- |
| Campaign type | Number of campaigns sent | Percentage |
| A | 3,979 | 55.2 |
| B | 2,655 | 36.83 |
| B | 574 | 7.96 |

Next is to determine which campaigns are sent the most. Among 30 campaigns, there are three campaigns that are sent more than others. Campaign 18 is the most sent with 15.72 percent, followed closely by campaign 13 and 8 with approximately 14% in each. All three campaigns are in type A, which shows that campaign type A is deployed significantly more compared to others. Figure 18 summarises the insights.

Chart

Description automatically generated

Figure 18.Tree map of campaigns sent, visualized by proportion.

It is important to determine whether a marketing campaign is effective or not based on observing whether there are any changes in spending per week for customers that are advertised that campaign. Therefore, a weekly spending pattern is established for all households that are sent a type of campaign, and the duration of each campaign will be highlighted to examine for any change in spending.

For campaign selection, all campaigns in type A and C will be chosen since each type only has 5 and 6 campaigns sent out. For campaign type B, 5 most sent campaigns out of 19 campaigns will be selected as representatives for examination. The households selected for each campaign type are only advertised that type, so that influence of others can be separated. For instance, households participate in campaign type C are only advertised type C and no other types, so it is clarified that any change in spending of the selected customers cannot be influenced by other campaigns in type A and B. The graphs in figure 19,20 and 21 indicates spending per week overtime for all households that receives campaign type A, B and C respectively. The coloured rectangles highlight the duration of corresponding campaigns.

A picture containing box and whisker chart

Description automatically generated

Figure 19.Weekly spending of customers with campaign type A and highlights of each campaign’s duration.

Figure 19 indicates that spending per week increases during four out of five type A campaigns. Campaign 18 (highlighted in green) is not only the most sent campaigns overall, but also seems to be the most effective among all campaigns in the same type. Average spending per week during campaign 18 increases up to $105 and $120 during the campaign, whereas spending during other campaigns only varies around $95 to $110 weekly. Campaign 26 (highlighted in blue) is also very effective as there is a surge in spending from around $100 to $110 on average towards the end of the campaign. Campaign 30 (highlighted in orange) is semi-effective as the average spending is lower than the weeks before, although a slight increasing trend is still spotted during the campaign. Campaign 8 (highlighted in purple) managed to increase spending to over $100 per week. Finally, campaign 13 (highlighted in red) seems to be the least effective among all type A campaigns, as the spending fluctuates between $95 and $105, although a few spikes is still observed.

*Chart, histogram

Description automatically generated*

Figure 20. Weekly spending of customers with campaign type B and highlights of each campaign’s duration.

Figure 20 shows that among 5 campaigns in type B, 4 campaigns are also valuable in increasing spending of customers. Campaign 22 (highlighted in orange), which is the most advertised type B campaign, is also the most effective as spending during the campaign increases up to over $125 per week on average. Campaign 16 and 17 (highlighted in green and blue) overlaps each other, and the average range of spending during the campaigns is also higher compared to the weeks before, excluding one week ending campaign 16 with a drop in spending down to $75. Campaign 11 (highlighted in red) sees a downward trend in spending but increases gradually towards the end of the campaign. Campaign 7 (highlighted purple) is the least effective, with average spending significantly lower than the weeks in advance.

Chart, scatter chart

Description automatically generated

Figure 21.Weekly spending of customers with campaign type C and highlights of each campaign’s duration.

Finally, according to figure 21, 4 out of 6 campaigns in type C are efficient in spending motivation. Campaign 27 (highlighted in orange) provides the most visible boost as spending increases in the middle of the campaign and skyrocketed towards the end of the campaign, reaching up to an all-time high of $175. As campaign 27 is the 4th most advertised campaign overall and the most advertised type C campaign, this could partially explain the success. Two overlapping campaign 3 and 6, highlighted in purple and cyan respectively, are both semi successful. Among 10 observable weeks in the duration of both campaigns, spending of 6 weeks average between $75 to $75 which is higher than before. Indicated in red, green and blue, campaign 14, 15 and 20 also overlaps each other. However, only campaign 14 is effective as spending increases until it ends. After campaign 14 ends, the weekly spending reduces significantly overtime during campaign 15 and 20, varying between $25 and $50 per week.

## 4.3. Classification results

The second goal is to use the information gained as inputs for predicting whether a customer will participate in a marketing campaign and this can be done using predictive analysis. This section will specify the process of modelling chosen algorithms, and their results.

The test set has 84% customers who do not participate (denoted as 0) and 16% customers that participate in a campaign (denoted as 1), in which the model with the best sensitivity and the balanced between sensitivity and specificity is classified as well-performing.

### 4.3.1. Random forest

The random forest model is first implemented on the training data with the default parameters. The model trained with 500 trees and 4 variables are split at each tree node. The trained default model is implemented on the test dataset with a determined optimal threshold of 0.275. Next, the model’s hyperparameters are tuned on the trained dataset and implemented on the test dataset. Using grid search and manual search, the optimal parameters are found with 2,500 trees and 12 variables split at each node. The performance of the random forest classification model is then evaluated using probability classification output with the optimized classification threshold of 0.205. Table 10,11 and 12 shows the detailed statistics of each model and figure 22 shows the ROC curve and AUC of each.

Table 10. Confusion matrix of default random forest and random forest with tuned hyperparameters on test data

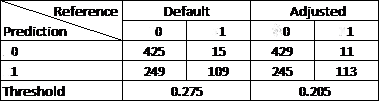


Table 11. Summary statistics of two random forest models



Table 12. Summary statistics of two random forest models



Chart, line chart

Description automatically generated

Figure 22.ROC and AUC of two random forest models before (left) and after tuning(right)

### 4.3.2. Boosted c5.0 decision tree

For Boosted C5.0 decision tree, the model will first be implemented on the training dataset with the default parameters chosen by the Caret model of 20 boosting iterations, and the threshold that provides the best balance between Sensitivity and Specificity is chosen with 0.45. Grid search is than conducted by finding the best combination of boosting iterations and modelling methods (decision trees or rules). With a grid of 100 boosting iterations, the best combination of 98 boosts and a tree-based method is used on the model, as indicated in figure 23.

A picture containing graphical user interface

Description automatically generated

Figure 23.Best combination of hyperparameters for Boosted C5.0 decision trees.

With the selected hyperparameters, the model is trained and fitted on the testing dataset. The best threshold for the balance of sensitivity and specificity is also determined at 0.299. Table 13 to 15 examine the statistics of the two models, and figure 24 shows the ROC and AUC of each model.

Table 13. C5.0 boosted decision tree confusion matrix with default parameters and tuned parameters on test data

Table

Description automatically generated

Table 14. Summary statistics of two boosted c5.0 models.



Table 15. Summary statistics of boosted C5.0 decision tree



Chart, line chart

Description automatically generated

Figure 24.ROC and AUC of boosted decision tree before (left) and after tuning (right)

### 4.3.3. Bagged decision tree

For bagged decision trees, the default number of decision trees voting is determined at 25. With the most optimized classification threshold of 0.225, the default model is implemented on the test data for prediction. For the second stage, the model’s number of trees is manually searched on various levels ranging from 1 to 200. Figure 25 and 26 shows the quality of the model at different levels of nbagg.

Table, calendar

Description automatically generated

Figure 25. Performance of the bagged decision trees model at different nbagg levels

A picture containing chart

Description automatically generated

Figure 26. Visualization of the model performance at different nbagg levels

At 200 trees, it will have better ROC but at 100 trees, the Sensitivity and Specificity are the highest. Since the main focus is the positive class, the nbagg value of 100 is selected for the model. Upon predicting the test dataset using an optimized threshold of 0.28, the results are shown in Table 16 to 18 and figure 27 indicates the ROC and AUC of the models before and after tuning.

Table 16. Confusion matrix of bagged decision trees with default and tuned parameters on test data

Table

Description automatically generated

Table 17. Summary statistics of Bagged decision trees



Table 18. Summary statistics of bagged decision trees



Chart, line chart

Description automatically generated

Figure 27. ROC and AUC curve of bagged decision trees before (left) and after tuning (right).

## 4.4 Summary

This chapter specifies the results of the mining and machine learning process on the finalized dataset. The first half of the chapter explains the main features of the households that shops with the retailer such as their spending tendency, favourite products as well as their main demographical features. This is followed by a detailed examination of the marketing campaigns deployed by the retailers. Knowing the campaigns that are the most sent, a relationship can also be established with their successfulness.

The second half of the chapter provides the results of the chosen classification models in predicting consumer participation. Using the aforementioned insights of each consumer, they act as variables for classification models to predict whether a consumer will participate in a marketing campaign or not.

**Section word count: 2,980**

# **Chapter 5. Discussions**

## 5.1. Overall findings

Through the application of data mining techniques, insights relating to the consumers have been discovered. In the context of the specific retailer in question, the majority of the households are houseowners consists of 2 individuals aged between 45-54 with medium income between $50,000 and $74,000. Despite this observed pattern, people in the high-income bracket (Over $125,000) spends more per week compared to the medium-income earners, and that people in the younger age bracket (35 to 44 years old) actually spends more per shopping trip and per week despite taking less percentage of the overall consumer demographic. The retailer therefore needs to design appealing campaigns to attract their majority consumer segment as they aren’t spending as much compared to the other household segments with less presence in total demographic.

As for marketing successfulness, increased advertising can potentially improve marketing efficiency. The findings indicate that the most advertised campaigns of each campaign type all show varying degree of effectiveness, with weekly spending increase during the campaign compared to the weeks before the campaigns start. Another interesting pattern can be seen that customers most likely participate in the middle or towards the end of campaigns, rather than in the beginning as spikes in spending can be observed in the middle or at the end of each campaign duration.

## 5.2. Classification performance

### 5.2.1. Random forest

Both random forest models before and after tuning do a well job in identifying potential customers who are interested in a marketing campaign. The default model with the parameters chosen does a good job with the AUC of 0.804 and the sensitivity of 87%, indicating that the model is accurate in predicting interested consumers. The Specificity of the original model is also not suffered too much from the high sensitivity with 63% of customers who are not interested correctly classified. However, the adjusted model with optimized parameters does even a better job as it correctly identifies 91% of customers who are interested while maintaining even a slightly better specificity with 4 less consumers wrongly classified as interested. AUC improves from model tuning with 0.81. Accuracy and balanced accuracy are also slightly improved with 67% and 77% respectively.

Therefore, the random forest model with adjusted hyperparameters does a good job in predicting consumer behaviour, with a strong specialization in predicting interested consumer rather than filter uninterested consumers. Retailers who wants to focus deploying a large quantity of marketing campaigns with cheap cost will have an advantage is using Random Forest model since it has very high accuracy in targeting interested consumers for maximum consumer awareness. This is done in exchange for more uninterested consumers wrongly advertised but if the cost the mass marketing campaign is cheap, the penalty for falsely classify unenthusiastic consumers will be minimized.

### 5.2.2. Boosted C5.0 Decision trees

After tuning, despite a 4% improvement in sensitivity, the specificity suffers the most from this boost with a decrease from 87% to 64%. This also results in overall accuracy dropped from 86% to 68%. AUC also reduced with only 0.81 compared to 0.92 as of before tuning. Hence, the default models with parameters chosen by Caret is more optimized and hence should be used.

For boosting C5.0 decision trees, it performs better than Random Forest models in terms of accuracy with 86 % compared to 67%. AUC is also higher with 92%. However, the boosted model has an underwhelming performance when it comes to predicting consumer participation as sensitivity is only 83% compared to 91% of Random Forest and performs significantly better in filtering uninterested consumers with a high true negative rate of 87% as opposed to 63% for Random Forest.

Therefore, this model has a relatively balanced True Positive to True negative rate (83% to 87%), with an emphasize on cost aspect of direct marketing by reducing waste from targeting uninterested consumers, in exchange for less accuracy in predicting interested consumers. As opposed to Random Forest, the C5.0 boosted model can be especially useful if the retailer wishes to deploy expensive yet effective strategies, in which the unnecessary cost of wrong identification is reduced as False Positives are kept to a minimal while a decent True Positive rate still ensures an efficient number of customers gets correctly advertised. On the other hand, if the retailer wishes to deploy cheap mass marketing strategies, Random Forest is a better classification algorithm of choice since a higher sensitivity is more preferable in order to increase consumer reach/awareness, despite of higher false positives.

### 5.2.3. Bagged decision trees

The default bagged decision tree model and the tuned model has little difference in sensitivity. However, the specificity of the tuned model is higher than the default model, with 69% to 63%. This helps raising the accuracy to 72% overall, but the change in AUC is minimal. Regardless, the tuned model does a better job classifying uninterested consumer with minimal reduction in interested consumer prediction. Hence, the tuned model should be used if bagging is the method of choice.

The bagging model is the “middle ground” between the Random Forest and C5.0 Boosted models. Specifically, sensitivity of the bagged model is higher than that of C5.0 (with 87% to 83%) but lower than Random Forest’s sensitivity (87% to 91%). Similarly, specificity of the bagged model is higher than Random Forests (69% to 63%) yet lower than C5.0 (with 69% to 87%) significantly. Therefore, the model does not have any particular strength compared to the other two models. Although the bagging model surpasses C5.0 in consumer participation prediction, it is not as specialized for the same purpose as Random Forest. Likewise, the model does a better job of minimizing type 1 error (which prevents waste of resources) compared to Random Forest but C5.0 surpasses the model in doing so.

Nevertheless, with the specificity of approximately 70%, the bagged decision tree also has a relatively balanced True Positives to True Negatives (87% to 70%). Therefore, it can still be used if the retailer does not have a particular marketing strategy in mind, but rather wanting to have a balanced approach between quantity and quality of their marketing campaign. In other words, they can successfully direct advertise 87% of consumers while only wasting resources in falsely advertise 30% of unimpressed consumers.

### 5.2.4 Final remarks of classification models

In general, Random Forest, Boosted C5.0 decision trees and Bagged decision trees all yield good results. For random forest, the balanced accuracy of prediction in general is 77%. With an area under the curve of 0.81, the algorithm can confidently predict which consumer will participate in a marketing campaign as its sensitivity is the highest of all models with 91%. The specificity of the algorithm, although not suffered too significantly, only remains at 63%. The algorithm is therefore more specified in focusing on reaching as many potentially interested consumer as possible. On the other hand, C5.0 boosted decision trees has a solid performance with an AUC of 0.92 and can accurately filter out uninterested consumers with 87% specificity. However, the model is not as accurate is predicting consumer participation as Random Forest with only 83% sensitivity. Therefore, the model can be used to maximum efficiency if expensive marketing campaigns were to be deployed as false positive rate is kept to a minimum. Bagged decision trees have the least remarkable performance out of all models with 88% and 70% sensitivity and specificity respectively. However, with a decent AUC of 0.81, the model can be used if the retailer has a neutral marketing strategy which allows them to reach a decent number of consumers without wasting resources on false positives.

## **5.3. Summary**

In general, this chapter further specify and discuss the results that were presented in the previous chapter.

Specifically, the general features and behaviours of the retail’s consumer base are presented. Consumer insights as such can be especially useful for retailer as this allows a more robust approach in planning to appeal for a particular market segment especially if it takes the majority of their consumer base. Knowing consumer’s spending per week and what they most usually spend on also offers a more general overview of how different strategies such as discounts and loyalty programs should be appropriately advertised towards each individual in a direct matter.

Using the aforementioned insights of each consumer, it is possible to predict which particular consumer will participate in marketing campaigns with high efficiency. The results also reflect how the models should be uses in various contexts that is dependent on the way the retailer wants to deploy their marketing campaigns.

**Word count: 1,416**

# Chapter 6: Conclusions

## 6.1. Overall achievements

The dissertation further reinforces the known effects and application of data mining and machine in helping retailer extracts important insights with regards to the consumers, as well as the characteristics of marketing campaign and what makes them effective.

Specifically, the first objective involves using data mining for consumer knowledge. The goal is reached as various tables and visualizations painted a picture of atypical consumer of the retail chain as well as their spending behaviour and what type of product category is most liked by households

The dissertation also successfully answers the second objective, which aims to explains the relationship between a marketing strategy and consumer spending. In terms of marketing, the retailer understands more about the characteristics of how their campaigns are deployed and a brief examination of consumer spending during the campaigns is examined. A general pattern is observed that the more heavily advertised a marketing campaign, the more likely it will motivate consumers to spend more during the duration of a campaign.

Finally, the second aim and third objective has been investigated, which involves predicting consumer response to marketing efforts. This has been determined by using a consumer’s personal information such as their household data and their spending behaviour as variables of predicting whether they will participate in marketing campaigns or not. The results derived from modelling shows some degree of effectiveness of each model in one particular aspect that can be optimized for various scenarios in the retail environment.

## 6.1. Implications

Many retail datasets can be complicated in terms of structure and size. The dissertation sets the direction of which consumer features in such data can be narrowed down to, which helps with not only prediction but also knowledge mining purposes. Particularly, it is established that consumer demographic information, description of each consumer’s most frequently bought products and their purchase behaviour all serves as great predictive variables as they are all related to the consumer participation probability. Therefore, such features should be focused on when attempting to understand consumer features and their response behaviour.

As the goals are met, the dissertation helps the retailer sets out a direction when it comes to advertising design. By identifying the main characteristics of the consumers and how they behave can offer the retailer a crucial overview of what a marketing campaign should include and how it should be advertised to appear attractive to the majority consumer segment.

However, the main insights when it comes to classification modelling lies in the retailer’s decisions in research and deploy their direct marketing strategies for maximum effectiveness. According to the “no free lunch” theorem, no particular model that works well on all given problems (Wolpert, 1996). Therefore, the choice of what algorithm to use is entirely dependent on the retailer and their goal. For instance, if the retailer wishes to amplify consumer reach and awareness, models that offers high sensitivity such as Random Forest allows the consumers to focus on the interested consumers. Whereas models with high specificity such as Boosted decision trees is shown to be useful in deploying cost-intensive campaigns by reducing type 1 error, without much sacrifice in sensitivity. A retailer who prefers a middle-ground approach might prefer model with a good balance between sensitivity and specificity to not lose any opportunities or waste resources. The dissertation has not only identified the effectiveness of the algorithms, but also what evaluation metrics that need focused on for various scenarios, which will assist retailers if an application of the algorithms in the real retail environment occurs.

## 6.2. Limitations and opportunities

Despite satisfactory results in mining and modelling, a few limitations are also identified. As the whole process requires computational power, especially modelling algorithms, the whole process takes a lot of time and errors are identified in the process. The second limitation is in the lack of data as opposed to the complexity of the data. Some categorical variables are highly cardinal with up to over 100 levels, while the data sampled for classification is not sufficient in terms of size. This also makes the algorithms harder to generalize on unseen data as the occurrence of mismatch in values in a certainty. Unless similar data pre-processing techniques is carried out on the new dataset, the models will not be functional when applied.

There are also a few opportunities that can be further looked in to in the future researches. For instance, despite accurately identifying which consumer will participate in marketing campaigns, a more personalized approach can also be taken. Specifically, it is also possible to look at which campaign a household will most likely engage in among all the campaigns sent to them based on the characteristics of each campaign. This will further help retailers optimize the features of a marketing campaign to further appeal to as many consumer as possible. Another possibility is to predict whenever a consumer will increase or decrease their spending based on their past transaction history or predict what product will most suit their liking.

**Word count: 832**

# References

Aeron, H., Kumar, A. and Moorthy, J., (2012). Data mining framework for customer lifetime value-based segmentation. Journal of Database Marketing & Customer Strategy Management, 19(1), pp.17-30.

Ahmed, S., (2004). Applications of data mining in retail business. In: International Conference on Information Technology: Coding and Computing. Las Vegas, NV, USA: ITCC, pp.455-459.

Ali, J., Khan, R., Ahmad, N. and Maqsood, I., 2012. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), p.272.

Amit R, P. and Kameshvari J, B., (2012). A STUDY ON CONSUMER BEHAVIOUR OF ORGANIZED AND UNORGANIZED RETAIL OUTLES IN VADODARA CITY. International journal of Engineering and Management sciences, 3(4), pp.466-474.

Appraisal institute, 2015. The Dictionary Of Real Estate Appraisal. 6th ed. Chicago: Appraisal Institute, p.219.

Archaux, C., Laanaya, H., Martin, A. and Khenchaf, A., (2004). An SVM based Churn Detector in Prepaid Mobile Telephony. In: International Conference on Information and Communication Technologies: From Theory to Application. Boulogne Billancourt, France, pp.pp.459–460.

Baran, R., Galka, R. and Strunk, D., (2008). Principles Of Customer Relationship Management. Mason, Ohio: Thomson/South-Western.

Beenstock, S., (1999). ANALYSIS: Supermarkets Entice The ’Ultra’ Customers - Tiered And Tailored Loyalty Cards Are Likely To Replace Retailers’ Basic Schemes As They Seek To Motivate And Retain High-Spenders, As Sue Beenstock Reports. [online] Campaignlive.co.uk. Available at: <https://www.campaignlive.co.uk/article/analysis-supermarkets-entice-ultra-customers-tiered-tailored-loyalty-cards-likely-replace-retailers-basic-schemes-seek-motivate-retain-high-spenders/58236> [Accessed 29 July 2020].

Bernard, H., 2017. Research Methods In Anthropology. 6th ed. Rowman & Littlefield Publishers.

Berry, L., (2002). Relationship Marketing of Services Perspectives from 1983 and 2000. Journal of Relationship Marketing, 1(1), pp.59-77.

Bhattacharya, C. and Sen, S., (2003). Consumer–Company Identification: A Framework for Understanding Consumers’ Relationships with Companies. Journal of Marketing, 67(2), pp.76-88.

Bhattacharyya, S., (2000). Evolutionary Algorithms in Data Mining:Multi-Objective Performance Modeling for Direct Marketing. In: sixth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, US: Association for Computing Machinery, pp.465-473.

Biau, G. and Scornet, E., 2016. A random forest guided tour. Test, 25(2), pp.197-227.

Bojei, J., Julian, C., Wel, C. and Ahmed, Z., 2013. The empirical link between relationship marketing tools and consumer retention in retail marketing. Journal of Consumer Behaviour, 12(3), pp.171-181.

Bolton, R., Kannan, P. and Bramlett, M., (2000). Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value. Journal of the Academy of Marketing Science, 28(1), pp.95-108.

Bose, I. and Mahapatra, R., (2001). Business data mining — a machine learning perspective. Information & Management, 39(3), pp.211-225.

Bose, R., (2002). Customer relationship management: key components for IT success. Industrial Management & Data Systems, 102(2), pp.89-97.

Boztuğ, Y. and Reutterer, T., (2008). A combined approach for segment-specific market basket analysis. European Journal of Operational Research, 187(1), pp.294-312.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Bryman, A. and Bell, E., 2012. Social Research Methods. 5th ed. Oxford University Press.

Bult, J. and Wansbeek, T., (1995). Optimal Selection for Direct Mail. Marketing Science, 14(4), pp.378-394.

Burt, S., (2000). The strategic role of retail brands in British grocery retailing. European Journal of Marketing, 34(8), pp.875-890.

Business.qld.gov.au. (2020). Using Direct Marketing | Business Queensland. [online] Available at: <https://www.business.qld.gov.au/running-business/marketing-sales/marketing-promotion/direct-marketing/using> [Accessed 28 July 2020].

Carlsson, F. and Löfgren, Å., (2006). Airline choice, switching costs and frequent flyer programmes. Applied Economics, 38(13), pp.1469-1475.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, pp.321-357

Chawla, N.V., Japkowicz, N. and Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), pp.1-6.

CHEN, M., CHIU, A. and CHANG, H., (2005). Mining changes in customer behavior in retail marketing. Expert Systems with Applications, 28(4), pp.773-781.

Chopra, B., Bhambri, V. and Krishan, B., (2011). Implementation of Data Mining Techniques for Strategic CRM Issues. International Journal of Computer Technology and Applications, 2(4), pp.879-883.

Choy, L., 2014. The Strengths and Weaknesses of Research Methodology: Comparison and Complimentary between Qualitative and Quantitative Approaches. IOSR Journal of Humanities and Social Science, 19(4), pp.99-104.

Cohen, L., Manion, L. and Morrison, K., 2007. Research Methods In Education. 6th ed. London: Routledge.

Colla, E., (2003). International expansion and strategies of discount grocery retailers: the winning models. International Journal of Retail & Distribution Management, 31(1), pp.55-66.

Crotty, M., 1998. The Foundations Of Social Research. 1st ed. SAGE.

Dhar, S., Hoch, S. and Kumar, N., (2001). Effective category management depends on the role of the category. Journal of Retailing, 77(2), pp.165-184.

Dogan, O., Ayçin, E. and Atıl Bulut, Z., (2018). CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. International Journal of Contemporary Economics and Administrative Sciences, 8(1), pp.1-19.

Du, Rex Yuxing and Wagner A. Kamakura (2012), “Quantitative Trendspotting,” Journal of Marketing Research, 49 (August), 514–36.

Dudwick, N., Kuehnast, K., Jones, V.N. and Woolcock, M., 2006. Analyzing social capital in context. A guide to using qualitative methods and data, pp.1-46.

dunnhumby. 2020. Source Files - Dunnhumby. [online] Available at: <https://www.dunnhumby.com/source-files/> [Accessed 10 October 2020].

dunnhumby.com. 2020. About Us. [online] Available at: <https://www.dunnhumby.com/retailers/customer-insights-engagement/> [Accessed 10 October 2020].

dunnhumby.com. 2020. Customer Insights & Engagement. [online] Available at: <https://www.dunnhumby.com/retailers/customer-insights-engagement/> [Accessed 10 October 2020].

Elsner, R., Krafft, M. and Huchzermeier, A., (2004). The 2003 ISMS Practice Prize Winner: Optimizing Rhenania's Direct Marketing Business Through Dynamic Multilevel Modeling (DMLM) in a Multicatalog-Brand Environment. Marketing Science, 23(2), pp.192-206.

Fader, P., Hardie, B. and Lee, K., (2005). RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. Journal of Marketing Research, 42(4), pp.415-430.

Fmi.org. (2018). FMI | Supermarket Facts. [online] Available at: <https://www.fmi.org/our-research/supermarket-facts> [Accessed 21 July 2020].

Garg, R., 2018. A Primer To Ensemble Learning – Bagging And Boosting. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/> [Accessed 31 October 2020].

Garvey Orji, M., (2017). Impact of Personality Factors on Consumer Buying Behaviour Towards Textile Materials in South Eastern Nigeria. International Journal of Business and Economics Research, 6(1), p.7.

Gay, L.R., Mills, G.E. and Airasian, P.W., 1992. Educational research: Competencies for analysis and application.

Giudici, P. and Passerone, G., (2002). Data mining of association structures to model consumer behaviour. Computational Statistics & Data Analysis, 38(4), pp.533-541.

Goddard, W. and Melville, S., 2004. Research Methodology. 2nd ed. Lansdowne: Juta.

Grewal, D., Roggeveen, A. and Nordfält, J., (2017). The Future of Retailing. Journal of Retailing, 93(1), pp.1-6.

Ha, S., (2007). Applying knowledge engineering techniques to customer analysis in the service industry. Advanced Engineering Informatics, 21(3), pp.293-301.

Hand, D., (1981). Discrimination And Classification. John Wiley & sons.

Hatch, M.J., 2018. Organization theory: Modern, symbolic, and postmodern perspectives. Oxford university press.

Hitchcock, D.H., Hitchcock, G. and Hughes, D., 1995. Research and the teacher: A qualitative introduction to school-based research. Psychology Press.

Ho, R., Huang, L., Huang, S., Lee, T., Rosten, A. and Tang, C., (2009). An approach to develop effective customer loyalty programs. Managing Service Quality: An International Journal, 19(6), pp.702-720.

Hogreve, J., Iseke, A., Derfuss, K. and Eller, T., (2017). The Service–Profit Chain: A Meta-Analytic Test of a Comprehensive Theoretical Framework. Journal of Marketing, 81(3), pp.41-61.

Horning, N., 2013. Introduction to decision trees and random forests. Am. Mus. Nat. Hist, 2, pp.1-27.

Hox, J.J. and Boeije, H.R., 2005. Data collection, primary versus secondary.

https://www.statcan.gc.ca/eng/start. (2012). NAICS 2012 - 44511 - Supermarkets And Other Grocery (Except Convenience) Stores - Industry. [online] Available at: <https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=118464&CVD=118468&CPV=44511&CST=01012012&CLV=2&MLV=5> [Accessed 21 July 2020].

Hu, Y. and Yeh, T., (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. Knowledge-Based Systems, 61, pp.76-88.

Hudson, L. and Ozanne, J., 1988. Alternative Ways of Seeking Knowledge in Consumer Research. Journal of Consumer Research, 14(4), p.508.

Hui, S., Inman, J., Huang, Y. and Suher, J., 2013. The Effect of In-Store Travel Distance on Unplanned Spending: Applications to Mobile Promotion Strategies. Journal of Marketing, 77(2), pp.1-16.

Hwang, H., Jung, T. and Suh, E., (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications, 26(2), pp.181-188.

Igd.com. (2019). UK Food Sales To Grow By £24Bn By 2024. [online] Available at: <https://www.igd.com/articles/article-viewer/t/uk-food-sales-to-grow-by-24bn-by-2024/i/21868> [Accessed 21 July 2020].

Johnson, P. and Clark, M., 2006. Business And Management Research Methodologies. London: SAGE.

Kamber, M. and Han, J., (2011). Data Mining: Concepts And Techniques. 3rd ed. Waltham, USA: Morgan Kaufmann Publishers.

Kantarworldpanel.com. (2020). Grocery Market Share - Kantar. [online] Available at: <https://www.kantarworldpanel.com/en/grocery-market-share/great-britain/snapshot> [Accessed 22 July 2020].

Kazmi, S. and Batra, S., 2008. Advertising & Sales Promotion. New Delhi: Excel Books, p.7.

Keng, C., Tran, V. and Thi, T., (2013). Relationships among Brand Experience, Brand Personality, and Customer Experiential Value. Contemporary Management Research, 9(3), pp.247-262.

Khajvand, M. and Tarokh, M., (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, pp.1327-1332.

Kim, Y., (2006). Toward a successful CRM: variable selection, sampling, and ensemble. Decision Support Systems, 41(2), pp.542-553.

Kinsey, J., Wolfson, P.J., Katsaras N., (2001). Data mining: A segmentation analysis of U.S. grocery shoppers. The Retail Food Industry Center, University of Minnesota, St. Paul, USA

Kivetz, R. and Simonson, I., (2002). Earning the Right to Indulge: Effort as a Determinant of Customer Preferences toward Frequency Program Rewards. Journal of Marketing Research, 39(2), pp.155-170.

Kotu, V. and Deshpande, B., (2019). Data Science. 2nd ed. Cambridge, United States: Elsevier Inc., pp.199-220.

Kuhn, M. and Johnson, K., 2013. Applied predictive modeling (Vol. 26). New York: Springer.

Kuhn, T., 2015. The Structure Of Scientific Revolutions. Chicago, Ill.: The University of Chicago Press.

Li, M., D.Green, R., A. Farazmand, F. and Grodzki, E., (2012). CUSTOMER LOYALTY: INFLUENCES ON THREE TYPES OF RETAIL STORES’ SHOPPERS. Electronic copy available at: http://ssrn.com/abstract=1957207 INTERNATIONAL JOURNAL OF MANAGEMENT AND MARKETING RESEARCH, 5(1), pp.1-19.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

Loshin, D. and Reifer, A., (2013). Using Information To Develop A Culture Of Customer Centricity. Waltham: Elsevier Inc, pp.68-78.

Lund, D. and Marinova, D., (2014). Managing Revenue across Retail Channels: The Interplay of Service Performance and Direct Marketing. Journal of Marketing, 78(5), pp.99-118.

Madan L., B., (2006). Data Mining: A Competitive Tool in the Banking and Retail Industries. The Chartered Accountant, pp.588-594.

Marcus, C., (1998). A practical yet meaningful approach to customer segmentation. Journal of Consumer Marketing, 15(5), pp.494-504.

McCarthy, E., (2011). Basic Marketing: A Managerial Approach. 18th ed. McGraw-Hill Higher Education.

Meyer‐Waarden, L., (2008). The influence of loyalty programme membership on customer purchase behaviour. European Journal of Marketing, 42(1/2), pp.87-114.

Miguéis, V., Camanho, A. and Falcão e Cunha, J., (2012). Customer data mining for lifestyle segmentation. Expert Systems with Applications, 39(10), pp.9359-9366.

Miles, M. and Huberman, A., 1994. Qualitative Data Analysis. 2nd ed. Thousand Oaks, Calif.: Sage.

Min, H., (2006). Developing the profiles of supermarket customers through data mining. The Service Industries Journal, 26(7), pp.747-763.

Mittal, R., (2001). E-CRM IN INDIAN BANKS: AN OVERVIEW. Delhi Business Review, 2(1).

Mount, J. and Zumel, N., 2018. The vtreat R package: a statistically sound data processor for predictive modeling. Journal of Open Source Software, 3(23), p.584.

Ngoc Do, T., (2011).  Apply Data Mining To Segment Retail Market Based On Purchasing Portfolios. Master thesis. högskolan i borås.

Nisbet, R., Miner, G., Yale, K., Elder, J. and Peterson, A., (2018). Handbook Of Statistical Analysis And Data Mining Applications. 2nd ed. London, United Kingdom: Academic Press, Elsevier.

Nowok, B., Raab, G. and Dibben, C., 2016. synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11).

Oliver, R., (1999). Whence Consumer Loyalty?. Journal of Marketing, 63, p.33.

Passingham, J., (1998). Grocery Retailing and the Loyalty Card. Market Research Society. Journal., 40(1), pp.1-8.

Pauler, G. and Dick, A., (2006). Maximizing profit of a food retailing chain by targeting and promoting valuable customers using Loyalty Card and Scanner Data. European Journal of Operational Research, 174(2), pp.1260-1280.

Pauler, G. and Dick, A., (2006). Maximizing profit of a food retailing chain by targeting and promoting valuable customers using Loyalty Card and Scanner Data. European Journal of Operational Research, 174(2), pp.1260-1280.

Peker, S., Kocyigit, A. and Eren, P., (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. Marketing Intelligence & Planning, 35(4), pp.544-559.

Pradhan, S., (2012). Retailing Management: Text And Cases. 3rd ed. New Delhi: Tata McGraw-Hill, pp.3-20.

Pring, R., 2000. Philosophy Of Educational Research. London: Continuum.

Quinlan, J.R. (1996). Bagging, boosting, and C4.5. In: Thirteenth National Conference onArtificial Intelligence. Cambridge: MA: AAAI Press/MIT Press., pp.725–730.

Reinartz, W. and Kumar, V., (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. Journal of Marketing, 64(4), pp.17-35.

Reports.mintel.com. (2019). Supermarkets - UK - Nov 2019 - Market Research Report. [online] Available at: <https://reports.mintel.com/display/925746/?fromSearch=%3FcontentType%3DReport%26freetext%3Dgrocery%2520retail#> [Accessed 22 July 2020].

Rew, L., Koniak-Griffin, D., Lewis, M., Miles, M. and O'Sullivan, A., 2000. Secondary data analysis: New perspective for adolescent research. Nursing Outlook, 48(5), pp.223-229.

Roy, S., Shekhar, V., Quazi, A. and Quaddus, M., (2020). Consumer engagement behaviors: do service convenience and organizational characteristics matter?. Journal of Service Theory and Practice, 30(2), pp.195-232.

Rudrabasavaraj, M., (2010). Dynamic Global Retailing Management. 1st ed. Mumbai India: Himalaya Pub. House, pp.1-25.

Saunders, M., Lewis, P. and Thornhill, A., 2009. Research Methods For Business Students. 5th ed. Pearson Education, pp.124-125.

SCHERMERHORN, J., (2008). MANAGEMENT. 9th ed. [S.l.]: WILEY.SCHERMERHORN, J., 2008. MANAGEMENT. 9th ed. [S.l.]: WILEY.

Scotland, J., 2012. Exploring the Philosophical Underpinnings of Research: Relating Ontology and Epistemology to the Methodology and Methods of the Scientific, Interpretive, and Critical Research Paradigms. English Language Teaching, 5(9).

Snieder, R. and Larner, K., 2009. The Art Of Being A Scientist. Cambridge: Cambridge University Press.

Song, H., Kim, J. and Kim, S., (2001). Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications, 21(3), pp.157-168.

Steel, E., (2008). Marketers Reach Out To Loyal Customers. [online] WSJ. Available at: <https://www.wsj.com/articles/SB122766322705958805> [Accessed 28 July 2020].

Subramanian R., K., (2017). Building customer relation through direct marketing. International Journal of Combined Research & Development (IJCRD), 6(9), pp.799-810.

The Economic Times. 2020. Definition Of 'Customer Lifetime Value'. [online] Available at: <https://economictimes.indiatimes.com/definition/customer-lifetime-value> [Accessed 30 September 2020].

Venkatesan, R. and Farris, P., (2012). Measuring and Managing Returns from Retailer-Customized Coupon Campaigns. Journal of Marketing, 76(1), pp.76-94.

Verhoef, P., (2003). Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development. Journal of Marketing, 67(4), pp.30-45.

Walker, J.C. and Evers, C.W., 1988. The epistemological unity of educational research. Educational research methodology and measurement: An international handbook, pp.28-36.

Webster, F., (1992). The Changing Role of Marketing in the Corporation. Journal of Marketing, 56(4), p.1.

Wedel, M. and Kamakura, W., (2000). Market Segmentation: Conceptual And Methodological Foundations. 2nd ed. New York: Springer science + Business media LLC.

Wedel, M. and Kannan, P., 2016. Marketing Analytics for Data-Rich Environments. Journal of Marketing, 80(6), pp.97-121.

Wilson, J., 2019. Essentials Of Business Research. Milton: Routledge.

Windle, P., 2010. Secondary Data Analysis: Is It Useful and Valid?. Journal of PeriAnesthesia Nursing, 25(5), pp.322-324.

Wirth, R. and Hipp, J., 2000, April. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.

Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. Neural computation, 8(7), pp.1341-1390.

Yeh, I., Yang, K. and Ting, T., (2009). Knowledge discovery on RFM model using Bernoulli sequence. Expert Systems with Applications, 36(3), pp.5866-5871.

Yi, Y. and Jeon, H., (2003). Effects of Loyalty Programs on Value Perception, Program Loyalty, and Brand Loyalty. Journal of the Academy of Marketing Science, 31(3), pp.229-240.

You, Z., Si, Y., Zhang, D., Zeng, X., Leung, S. and Li, T., (2015). A decision-making framework for precision marketing. Expert Systems with Applications, 42(7), pp.3357-3367.

Zakaria, I., Rahman, B., Othman, A., Yunus, N., Dzulkipli, M. and Osman., M., (2014). The Relationship between Loyalty Program, Customer Satisfaction and Customer Loyalty in Retail Industry: A Case Study. Procedia - Social and Behavioral Sciences, 129, pp.23-30.

Zentes, J., Morschett, D. and Schramm-Klein, H., (2017). Strategic Grocery retailing Management. 3rd ed. Wiesbaden, Germany: Springer Gabler, pp.3-27.

Zhou, Z.H., 2009. Ensemble Learning. Encyclopedia of biometrics, 1, pp.270-273

Zielke, S., (2006). Measurement of retailers' price images with a multiple-item scale. The International Review of Retail, Distribution and Consumer Research, 16(3), pp.297-316.

Zielke, S., (2010). How price image dimensions influence shopping intentions for different store formats. European Journal of Marketing, 44(6), pp.748-770.

# SUBMISSION DOCUMENTS

## DISSERTATION ENTRY FORM

Graphical user interface

Description automatically generated

## ETHICAL FORM

Table

Description automatically generated

Table

Description automatically generatedTable

Description automatically generatedGraphical user interface, table

Description automatically generatedTable

Description automatically generatedGraphical user interface, application, table

Description automatically generatedGraphical user interface

Description automatically generatedA picture containing text

Description automatically generated