DEVELOPMENT OF CUSTOMER RESPONSE PREDICTION MODEL FOR PERSONAL LOAN USING LOGISTIC REGRESSION AND DECISION TREE

NUR HAFIZAH BINTI ZAKARIA

MASTER OF SCIENCE (DECISION SCIENCES)

UNIVERSITI UTARA MALAYSIA

2023

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

NUR HAFIZAH ZAKARIA

calon untuk Ijazah **MASTER OF SCIENCE (DECISION SCIENCE)**
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"DEVELOPMENT OF CUSTOMER RESPONSE PREDICTION MODEL FOR PERSONAL LOAN USING LOGISTIC REGRESSION AND DECISION TREE"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***26 Januari 2023.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
26 January 2023.

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Nerda Zura Zaibidi**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Assoc. Prof. Dr. Safwati Ibrahim**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Assoc. Prof. Ts. Dr. Mohd Faizal Omar**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof Ts. Dr. Izwan Nizal Mohd Shaharanee**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Ts. Dr. Jastini Mohd Jamil**	Tandatangan *(Signature)*

Tarikh:
(Date) ***26 January 2023***

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Secara umumnya, syarikat perkhidmatan kewangan menggunakan strategi pemasaran melalui media massa untuk menawarkan dan mempromosikan sesuatu produk atau servis kepada pelanggannya. Strategi ini dilihat kurang efektif kerana perbezaan minat pelanggan tidak diambil kira. Dengan pangkalan data pelanggan yang besar, syarikat perkhidmatan kewangan telah berubah dari strategi pemasaran yang tidak produktif melalui media massa kepada strategi pemasaran bersasar dengan mengenalpasti pelanggan yang berkemungkinan bertindak balas terhadap produk dan servis yang lebih spesifik. Namun, strategi ini kelihatan mencabar apabila dilaksanakan secara manual melalui pangkalan data pelanggan yang besar. Oleh itu, teknik perlombongan data digunakan untuk membangunkan model peramalan tindak balas pelanggan yang tepat. Matlamat kajian ini adalah untuk membangunkan model tindak balas pelanggan bagi meramal samada seseorang pelanggan bertindak balas terhadap produk pinjaman peribadi menggunakan teknik regresi logistik dan pepohon keputusan. Untuk tujuan pemodelan, sebanyak 10952 data pelanggan telah dikumpulkan dari salah satu bank komersial di Malaysia. Analisis Faktor dijalankan untuk mengenalpasti input terhadap model. Akhirnya, teknik regresi logistik dan pepohon keputusan digunakan bagi membangunkan model tindak balas pelanggan. Model yang berbeza daripada dua teknik ini dibandingkan. Model terbaik untuk meramal pelanggan yang berkemungkinan bertindakbalas terhadap produk pinjaman peribadi dipilih berdasarkan nilai ketepatan, 'precision' dan 'recall' paling tinggi. Kedua-dua model mempunyai ketepatan yang sama iaitu sekitar 98%. Namun, pepohon keputusan berprestasi sedikit lebih baik berbanding regresi terutama untuk pengesahan silang 10 kali ganda. Untuk pengesahan silang 5 kali ganda, regresi logistik berprestasi baik. Dalam kebanyakan masa, pepohon keputusan sepatutnya melaksanakan lebih baik dalam senario kes perniagaan yang mana regresi logistik berprestasi lebih baik apabila pencemaran data adalah kurang. Kajian ini boleh memberi faedah kepada bank di mana ia boleh memaksimumkan tindak balas pelanggan terhadap penawaran produk, meminimumkan keseluruhan kos pasaran dan menambahbaik pengurusan hubungan pelanggan.

Kata Kunci: Perlombongan data, Pemasaran bersasar, Model tindak balas pelanggan, Pepohon keputusan, Regrasi logistik

Abstract

In general, financial service companies use mass media marketing strategies to offer and promote a product or service to their customers. This strategy is seen as less effective because differences in customer interests are not taken into account. With a large customer database, financial service companies have shifted from unproductive mass media marketing strategies to targeted marketing strategies by identifying customers who are likely to respond to more specific products and services. However, this strategy appears challenging when implemented manually through a large customer database. Therefore, data mining techniques are used to develop accurate customer response forecasting models. The goal of this study is to develop a customer response model to predict whether a customer will respond to a personal loan product using logistic regression and decision tree techniques. For modelling purposes, 10,952 customer data were collected from one commercial bank in Malaysia. Factor analysis was conducted to identify inputs for the model. Finally, logistic regression and decision tree techniques were used to develop the customer response model. Different models from these two techniques were compared. The best model for predicting customers who are likely to respond to personal loan products was selected based on the highest accuracy, precision, and recall values. Both models had the same accuracy of around 98%. However, decision tree performed slightly better than regression, especially for 10-fold cross-validation. For 5-fold cross-validation, logistic regression performed well. In most cases, decision trees should perform better in business scenarios where logistic regression performs better when data contamination is low. This study can benefit banks by maximizing customer response to product offers, minimizing overall market costs, and improving customer relationship management.

Keywords: Data mining, Target marketing, Customers response model, Decision tree, Logistic regression

Acknowledgement

I would like to take this opportunity to express my sincere appreciation to my supervisor, Associate Prof. Ts. Dr. Izwan Nizal Bin Mohd Shaharanee and Ts. Dr. Jastini Binti Mohd Jamil. Their wise guidance, intellect and unwavering encouragement have provided me the essential motivation to study in this area. They were always supportive, committed and enthusiastic throughout the whole research and thesis writing. I am thankful to them for introducing me to different aspects of the academic life.

I gratefully acknowledge the scholarship granted by the Minister of Higher Education Malaysia (MyBRAIN15) and also thanks to the staff at School of Quantitative Sciences and Awang Had Salleh Graduate School, University Utara Malaysia for their help throughout my study.

My special acknowledgement goes to my beloved husband, Mohd Faizal Izwandi Bin Md Noor, for his love, tolerance and encouragement throughout my master study. Last, but not least, I am deeply indebted to my parents, Zakaria Bin Ariffin and Chik Binti Ahmad, for always praying for my success and happiness, and my family members for their love and support. Not forgotten also huge thanks to all of my friends for their encouragement and support during my study.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Businesses now have access to massive amounts of data and information on their consumer and client. This data can be empowered to develop and establish strong customer relationships, allowing them to be targeted for specific products and services. For various reasons, these vast client records and market data are retained. The targeted customers are chosen from a database based on multiple factors, including demographic information customer attributes such as occupation, age, sex, and purchase history. In advertising and promoting a new product or service, the selected clients are usually reached directly using one or more of the following channels are often used physical interaction, mailing, e-mail, telephones, and short messaging. Target marketing is the term for this form of marketing. Target marketing is becoming more popular among banks and insurance companies as a key means of connecting with customers. It has become a significant application for data mining, partly due to the increased interest in target marketing practices. Data mining is the most common approach used to discover potential customers for a certain campaign, product, or offer. Data mining techniques may be used to build a predictive response model based on past purchase data to forecast how a customer will react to a campaign or deal.

On the whole, Malaysian banks and other financial institutions use mass marketing to promote their services and products to their customers. This marketing effort wastes a lot of money and resources because only a small fraction of clients responds to these offers. This type of campaign is less effective due to client differences are not taken into account. As a result, mass marketing has grown less efficient in today's world of overwhelming products and services, as well as a highly competitive market.

In an effort to increase productivity, banks, financial institutions, and other businesses are shifting away from mass marketing and instead focusing on specific segments of their client base to provide tailored product and service offerings. This trend is reflected in the research of Pan and Tang (2014). Companies must target the most likely customers for their products to make more money. Identifying clients who are likely to respond to offers, on the other hand, will be challenging to do manually through a vast customer database. In ensuring reliable customer reaction predictions, a data mining technique was used to recognize patterns in customer datasets.

For target marketing, data mining can be very beneficial (Ling & Li, 1998). By determining which persons to contact, data mining can help enhance targeting. Target marketing has grown important as one of the data mining applications since the first half of the 1990s. One of the most prevalent data mining uses is to identify clients for marketing purposes. As a result, vast databases of consumer and market data are kept. Consumers or clients to be targeted in a specific campaign are chosen based on various factors, including demographic data and information about the customers' personal qualities such as occupation, age, and purchasing history. The primary goal is to find potential clients by recognizing profiles of customers who have previously expressed interest in the product. Consumer data from previous promotional campaigns may target customer preferences for a certain product.

The response model is a widespread method among target marketing analysts, and it's a valuable tool for fine-tuning direct marketing techniques (Potharst et al., 2002). A response model forecasts the likelihood of responding to a campaign or offer. The majority of response models are created using past purchase data. The model may identify potential consumers who are more likely to respond than others. The response modelling procedure consists of numerous processes, including data gathering, data pre-processing, variables selection, class balancing, classification, and evaluation from a modelling standpoint.

1.2 Problem Statement

The current marketing strategy employed by financial institutions and banks in Malaysia relies on mass marketing techniques, overlooking the unique characteristics of each customer. This results in significant waste since only a small percentage of customers are likely to purchase the product (Ling & Li, 1998). With the rise of competition, mass marketing has become less effective in achieving the desired response rate. Therefore, targeted marketing strategies are needed to identify potential customers who are most likely to respond positively to the product or service.

Personal loans are among the most popular products offered by financial institutions, and customer preferences and behaviours in selecting the right personal loan facility present a significant challenge for providers. Factors such as interest rates, loan amounts, repayment periods, and the provider's reputation are critical considerations for customers when choosing a personal loan. Additionally, customer preferences may vary depending on factors such as age, income level, and financial goals. Therefore, understanding customer behaviour and preferences in selecting a suitable personal loan product is essential for providers to develop effective marketing strategies and improve their offerings.

However, there is a lack of research on effective targeted marketing strategies for personal loans in Malaysian context. To improve the effectiveness and efficiency of targeted marketing strategies, there is a need for personalized marketing strategies for personal loans that could potentially increase the response rate and lead to more effective customer engagement (Setnes & Kaymak 2001).

One significant research gap in the field of targeted marketing in the Malaysian context is the lack of studies exploring the potential of data mining techniques, such as logistic regression and decision trees, in developing customer response prediction models. Additionally, there is a need for research on variable selection techniques, such as factor analysis, to reduce the complexity of models in predicting customer responses. Addressing these gaps is crucial for improving the efficiency and effectiveness of targeted marketing strategies in the Malaysian financial sector.

Furthermore, research should also focus on understanding the factors that influence customer responses to targeted marketing campaigns. Identifying relevant factors such as demographics, income levels, and previous purchase history will provide useful insights for developing effective marketing strategies tailored to the needs and preferences of individual customers. Overall, addressing these gaps will not only contribute to the academic knowledge in the field of targeted marketing but also provide practical benefits to the Malaysian financial institutions and banks.

1.3 Research Questions

The following research questions are to be answered at the end of this study:

1. What are the suitable data for customer response prediction of personal loan product?
2. How to identify a set of variables that influence customers response to personal loan product?
3. How to develop customer response prediction model for personal loan product?
4. How to evaluate the performances of different classifiers in predicting customer response to personal loan product?

1.4 Research Objectives

The main objective of this study is to develop customer response prediction model for personal loan product by applying logistic regression and decision tree techniques. In order to fulfil the main objective, the following sub-objectives must be achieved:

1. To investigate suitable data for customer response prediction of personal loan product
2. To identify a set of variables that influence customers response to personal loan product using feature selection techniques
3. To develop the customer response prediction model for personal loan product using logistics regression and decision tree
4. To evaluate the performances of different classifiers in predicting customer response to personal loan product

1.5 Significance of Study

Companies (marketers) can use a predictive response model to find a group of consumers who are more likely to reply than others and build a direct relationship with them. Companies may not only save their overall marketing costs by utilizing this technique, but they can also avoid irritating their customers. Carefully identifying and promoting a suitable product to correct customers is essential in marketing. Certain products may only be of interest to a subset of the total customer base. Sending too many uninteresting products offers and unwanted mail or phone calls can irritate potential and future customers. Additionally, sending product offers to all customers causes the cost of such a large-scale campaign to exceed the expected returns quickly. The relevance of response modeling for businesses has grown due to growing mailing costs and greater competition. By enhancing the targeting of offers (using a response model), businesses and organizations may combat the issues of rising competition and expensive marketing costs by reducing non-response. Furthermore, from the standpoint of the product offering's recipient (customers), businesses do not want to overburden them with options. A valuable technique for fine-tuning direct marketing campaigns is a response model. Even tiny increases attributable to modeling may result in significant advantages in terms of market value and product penetration among potential customers.

1.6 Scope of Study

This study focuses on the development of customer response prediction model for personal loan in one of the commercial banks in Malaysia. Even though there are many products offered by a single bank, such as credit cards, fixed deposit, mortgage loan, car loan, housing loan and many more, this thesis solely focuses on customer response to for personal loan only. The original dataset consisted of 10,952 customers and 110 variables. In addition, it can be observed that while there are 110 variables, not all the variables are useful for the development of the prediction model. Initial data cleaning was done to remove unwanted variable thus only 35 variables (plus one dependent variable) were included in the thesis. These variables were discarded as they consist of 50% missing values. Even though manually removing irrelevant variables takes a lot of time, tedious and laborious, it is still vital since it reduces the complexity of the models.

1.7 Organization of Thesis

There are five chapters in this thesis. The background of the chosen study subject is described in the first chapter, followed by a description of the issue area, research question, research objectives, the significance and the scope of the study. The literature review in chapter two will provide the reader with an overview of ideas pertinent to the study field and a sneak peek at studies on the subject. Additionally, previous works on data mining, target marketing, and response modelling were discussed and argued in this chapter. Chapter 3 focuses on the research methodology employed in this work, namely the knowledge discovery in database (KDD) process. This chapter provided a detailed explanation for each phase in the KDD process related to this research work. The results of the analysis are presented in Chapter 4. Chapter 5 is dedicated to the conclusion and the achievement of the study. In addition, the limitations of the study are revealed with some suggestions for future works.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

"Target marketing," which is the practice of marketing goods and services to a specific group of customers, was formerly referred to as "mail order," "direct marketing," or "the catalogue business" before the late 1960s. Even though some people associate "target marketing" with "direct mail," the term was first introduced in the late 1960s to combine the modern concepts of targeting and long-term value. Target marketing has been recognized by marketers for a long time as an effective and efficient way of engaging with customers. This modern business activity enables businesses to send messages or offer goods and services to a specific, targeted group of customers through various direct means, such as mail, telephone, email, and others, to optimize profits (Wang et al., 2005). These days, target marketing activities are also referred to as relationship marketing, interactive marketing, database marketing, and integrated marketing (Scovotti & Spiller, 2005).

Jonker et al. (2002) define target marketing as a marketing approach that seeks to create and maintain direct relationships between particular suppliers and buyers in one or more product/market combinations. Shin and Cho (2006) describe target marketing as the process of identifying potential buyers for specific products or services and promoting them through various channels. The primary aim of target marketing is to locate potential customers for a new product by examining customer data from past marketing campaigns and generating customer profiles. This entails creating response models that can identify the customers who are most likely to be interested in a particular product offering, maximizing the number of targeted customers who respond to the offer (Kaymak, 2001). The goal of response models is to identify customers who will be most interested in a specific product offer, so that as many of the targeted customers as possible respond to the product offer. Hence, Target marketing is carried out by creating a response model and evaluating past purchase data from customers using data mining techniques.

2.2 Customer Response Modelling in Target Marketing

Target marketing analysts employ response modelling, which is a well-known technique (Desarbo & Ramaswamy, 1994). It's proven to be a useful tool for fine-tuning target marketing efforts, as even minor improvements attributable to modelling can result in significant revenue gains (Elsner et al., 2004). The importance of response modelling comes from the fact that even a one-percentage-point increase in response can result in significant profit increases (Viaene et al., 2002). The necessity of response modelling has grown as a result of growing postal costs and more competition. Improving the targeting of offers could help address these issues by reducing non-response. The model can identify a group of consumers who are more likely to respond than others. In a more accurate response model, the subset will have more respondents and fewer non-respondents. As a result, the entire marketing cost can be greatly reduced without sacrificing opportunities.

The response model was created using a variety of ways. Among the methods used to address response modelling are judgmentally based Recency, Frequency, Monetary Value or Frequency, Recency, Amount and Type of goods (RFM/FRAT) methods, clustering methods, and supervised classification methods. RFM/FRAT or "manual" segmentation based on judgement is still commonly used to divide a client list into "homogenous" segments for targeting purposes (Ullah, Mohmad, Hussain, Johar, Khan, Ahmad, Mahmoud, & Huda, 2023).

RFM and FRAT analysis are two popular methods used in target marketing. Both these methods help to identify the most valuable customers and create targeted marketing campaigns for them. RFM analysis is a technique that helps to identify the customers who are most likely to respond to a marketing campaign. It involves analyzing the recency, frequency, and monetary value of each customer's purchases. Recency pertains to the recentness of a customer's purchase, frequency pertains to the frequency at which a customer makes purchases, and monetary value pertains to the amount of money a customer has expended on purchases.

Similar to RFM analysis, FRAT analysis also considers a customer's tenure, or the length of time they have been purchasing from a business. The four factors in a FRAT analysis are tenure, quantity, recency, and frequency. An organisation would first gather information on the purchase behaviours of its clients before using RFM or FRAT analysis for target marketing. To determine which clients were the most valuable, they would then divide their customer base depending on the RFM or FRAT characteristics. By segmenting its consumer base, a business may develop tailored marketing strategies for each segment that are more likely to appeal to their particular requirements and preferences.

For instance, if a business discovers a group of high-value clients who frequently make purchases but haven't done so in a while, it may develop a focused campaign to entice those clients to do so. Alternately, if a business recognises a group of clients who buy little but spend a lot each time, it may develop a focused campaign to persuade those customers to buy more frequently (Lenin & Zahavi, 2022). In conclusion, RFM and FRAT analysis are useful methods for target marketing because they help businesses better understand their consumers and develop more individualised and successful marketing programmes that are more likely to increase customer loyalty and generate sales.

2.2.1 Judgmentally based RFM/FRAT Methods

Instead of employing objective data-driven methodologies, judgmentally based RFM/FRAT methods relate to the process of developing consumer groups based on the subjective assessments of marketers or analysts. In other words, a business could use its own discretion to establish segments based on characteristics like demographics, geographic location, or psychographics rather of depending on data analysis to segment clients based on their purchasing behaviour (Lenin & Zahavi, 2022).

RFM variables are used to operationalize strategy for segmentation (Shepard, 1995). Recency, frequency, and monetary are the three factors that are used to categorize customers based on their past behaviour such as their purchase history. Recency measures the length of time since a customer's most recent purchase, while frequency looks at the number of purchases made over a given period or the proportion of mailings to which the customer responded. Finally, monetary measures the total amount of money spent on all purchases or the average amount spent per purchase, either within a product category or across all purchases. The basic rule is that the more recently a customer placed an order, the more goods s/he purchased from the firm in the past, and the more money s/he spent on the company's products, the more likely s/he is to purchase the next offering and the better target.

Additionally, more sophisticated manual approaches take into account the proximity of products/attributes when segmenting a dataset. The greater the probability of a purchase, the more similar the previously purchased products are to the current product offering, or the more relevant the product features (e.g., themes) become. For instance, when advertising a sports product, an individual who has previously purchased a sports product is more likely to respond to a new sports product offer, followed by those who enjoy camping, outdoor activities, and so on. In circumstances where males and females respond differently to a product offering, gender can be utilized to divide customers into groups. Generally, the list is divided into three categories such product or attribute type, RFM, and gender, which is known as FRAT (Frequency, Recency, Amount, and Type) (of product).

However, RFM and FRAT methods can be influenced by subjective and judgmental factors. Additionally, the primary assumption of the RFM method may not always hold true. In durable products like cars or refrigerators, for example, recency may work in the opposite direction, where the longer the time since the last purchase, the higher the likelihood of a new purchase. Finally, the RFM/FRAT may need to be performed iteratively to meet segment size limits, combining small segments and splitting large ones until a satisfactory solution is reached. This may result in a significant increase in computation time.

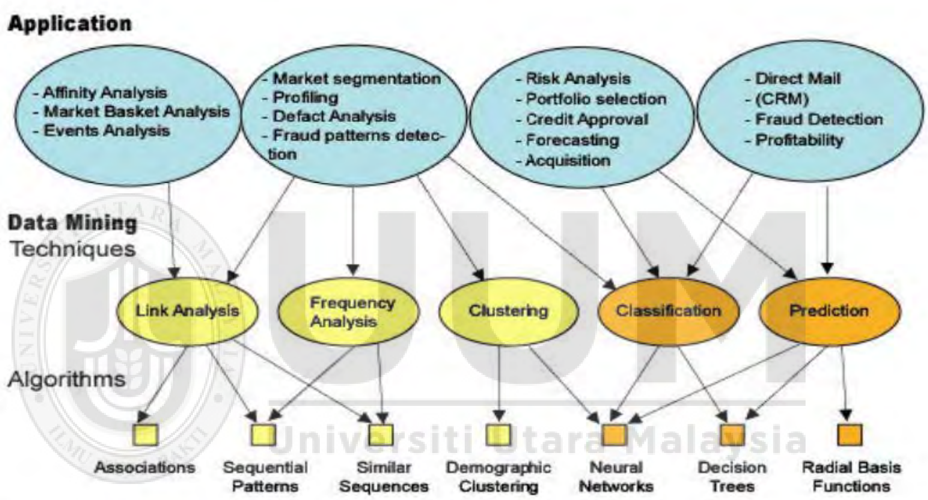
RFM/FRAT approaches that rely on judgement can be helpful in some circumstances when data is few or difficult to get, but they are often less efficient than segmentation methods that are driven by data. RFM and FRAT analysis are two examples of data-driven segmentation techniques that are based on objective standards and can offer more precise insights into consumer behaviour and preferences.

In conclusion, whereas data-driven segmentation techniques depend on objective data analysis to build segments based on consumer behaviour and preferences, judgmentally based RFM/FRAT approaches are a subjective approach to customer segmentation that relies on the judgement of marketers or analysts.

2.2.2 Application of Data Mining in Target Marketing

Data mining applications can be useful in a wide range of fields, including marketing, customer profiling, and retention for identifying potential customers, market segmentation, fraud detection, text and web mining, e-commerce, as well as production control, manufacturing, and scientific discovery. The use of data mining technology in marketing is a fairly common occurrence. Because it encompasses a wide range of scientific theories, such applications are referred to as Boundary Science. To begin, there are two fundamental disciplines: information technology and marketing. Statistics is another crucial foundation. It also has anything to do with psychology and sociology. This area's allure stems from the diverse range of fields available for study.

Data mining-based marketing may usually provide a customer a sales promotion based on his previous buy history. It is important to note that data mining is applicationoriented. Banking, insurance, traffic systems, retail, and other commercial fields have multiple common uses. Data mining tools can be applied to various market analysis techniques, including database marketing, customer segmentation and classification, profile analysis, and cross-selling. Additionally, data mining is commonly used for detecting fraud, credit rating, and churn analysis, as stated by Dunham in 2002. The interconnection between data mining applications and strategies can be visualized through Figure 2.1.



Source: Dunham (2002)

Figure 2.1 Application of Data Mining and Selected Strategies

Marketing, mainly database marketing, is home to many of the most productive data mining applications. The goal is to use database analysis to produce focused and thus effective marketing and promotional efforts. Currently, databases have the capability to store an immense amount of data pertaining to customers. By determining which persons to contact, data mining can help enhance targeting. According to Cabena et al. (1999), reducing the number of advertising messages benefits both consumers and marketers. Consumers appreciate fewer messages, while marketers can save money by lowering distribution costs and increasing the effectiveness of their campaigns.

The objective of this study is to support commercial banks in creating a target marketing approach that utilizes data mining techniques such as decision tree and regression analysis. The goal is to identify potential customers who are highly likely to borrow money from a market segment consisting of individuals who have previously used another bank's revolving credit service, and are therefore deemed as potential personal loan customers. This proposed model can help banks capture potential personal loan customers, leading to increased profitability.

2.3 Knowledge Discovery Database (KDD) Process

The KDD process is a way to turn enormous volumes of data into valuable information and knowledge (Fayyad, Shapiro, & Smyth, 1996). Figure 2.2 depicts the step involved in the KDD process from the beginning of data collection to the end of knowledge extraction proposed by Fayyad, Shapiro and Smyth (1996).

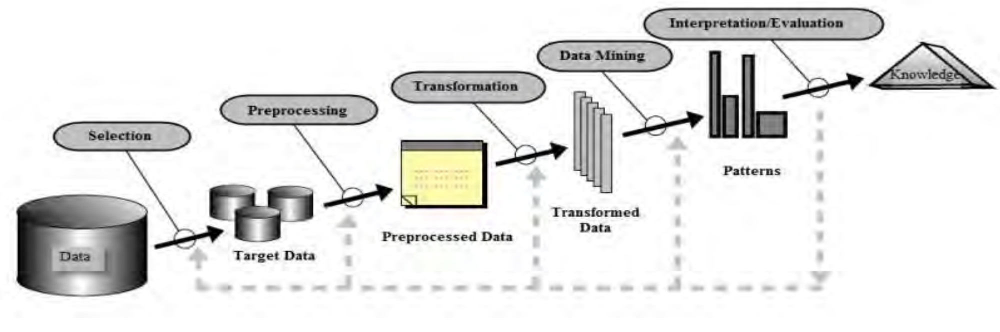


Figure 2.2. A Knowledge Discovery Database (KDD) process

Data can come from many sources, different databases, and formats; thus, the first step in the KDD process is data selection, which entails obtaining suitable data for a specific purpose at the discretion of subject experts. Pre-processing is the process of converting raw data into data that can be easily mined. Data pre-processing techniques include data cleaning, integration, transformation, and data reduction (Han, Kamber, & Pei 2012). Domain specialists generate data, and real-world data may not be in a format suited for mining activities, necessitating a pre-processing stage. Data cleaning is done to remove inconsistencies, noise, and incompleteness from the data. Problems with missing values can be solved by either discarding the record(s) or manually inserting the value, or by entering a global constant, a mean value of the attribute, or the most possible value. While binning, regression and outlier analysis can handle the problem of noise.

Data integration occurs when information is spread across multiple databases from different sources and needs to be merged into a unified data warehouse. During this process, it is crucial to address issues such as entity identification, inconsistencies in attribute naming, and the resulting correlation and duplication of data, as well as the usage of varying measurement scales or representations for attribute values. With techniques including smoothing, attribute building, aggregation, normalisation, discretization, and idea hierarchy development, data transformation aims to make mining more efficient. Data reduction, the third pre-processing step, also aims for a more efficient mining process by offering smaller data sets. Dimensionality reduction (feature subset selection), numerosity reduction, and data compression are all methods for reducing data.

Data mining is a process in KDD that involves automating pattern extraction. Even though the phrases KDD and data mining are sometimes used (Han, Kamber, & Pei

2012), there is a significant contrast between them. Data mining is an aspect of the process, whereas KDD refers to the whole process of knowledge discovery. One issue in KDD and data mining, according to Mattheus (1993), is the constantly changing and flawed character of real-world datasets. The successful application of KDD and data mining may be seen in a variety of industry fields, as well as in domains such as sport (Delen, Cogdell, & Kasap 2012), health (Li et al. 2004), marketing (Ngai, Xiu, & Chau 2009), finance (Zhou & Kapoor, 2011) and banking (Hormozi, 2004).

Data mining tasks generate patterns that reflect a subset of the data's model (Fayyad, Shapiro, & Smyth, 1996). Patterns can be turned into prediction rules or decision trees to guarantee that they are easily understood by users. The utility of these rules is determined by a post-processing procedure that assesses the rules' quality and interestingness. Although there is no universal agreement on what constitutes interestingness, a broad definition includes coverage, reliability, conciseness, diversity, peculiarity, novelty, utility, and surprisingness (Geng & Hamilton 2006). Once quality and interestingness metrics are employed to the extracted patterns, they transform into dynamic knowledge relevant to the field. Notwithstanding the patterns limited generalisation potential, which is limited to the data subset's similar context, KDD's overall capacity to extract novel and meaningful knowledge from vast datasets remains a good addition.

2.4 Data Pre-processing

The primary objective of data pre-processing is to ensure that the data used in the data mining process are of high quality and appropriate. Since data may come from multiple sources and in different formats, there is a possibility of data being anomalous, incorrect or missing. The purpose of the pre-processing phase is to ensure accurate and complete data by correcting or removing erroneous information, disregarding records with missing data, or predicting missing data through data mining tools. Moreover, data from various sources are often transformed into a standardized format for processing, and some data may be converted or encoded to enhance usability. Techniques such as data reduction, cleaning, integration, transformation, discretization, and others are used to manage pre-processing. Since the original data is typically unrefined, it is important to have an early and thorough understanding of the data, and a subject matter expert may be necessary to translate the information into a format that can be effectively modeled by the data miner to address the problem at hand (Han, Kamber, & Pei, 2012).

2.4.1 Factor Analysis for Variable Selection

Factor Analysis known as statistical approach that identifying the factors in a set of related variables and making it a useful tool for researchers (Shrestha, 2021). The technique allows researchers wisely choose which variables are relatively independent of one another and useful to create more accurate model. There are two common approaches to factor analysis, which are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is used to collect information among a set of variables, and CFA is widely used to test specific hypotheses or theories regarding the structure of variables set.

In the data pre-processing phase, factor analysis is essential in reducing the complexity of the data by identifying the underlying factors driving the correlations between the variables. This helps eliminate redundant or irrelevant variables, simplifying the analysis and making it easier to interpret the results. By identifying the most important factors, researchers can focus their attention on the variables that are most likely to have an impact on the outcome of interest. Therefore, factor analysis is an essential technique for researchers when preparing data for analysis.

Factor analysis has been extensively used in previous research in various data mining applications. It has been used to identify the underlying factors that contribute to the success of online businesses, to determine the key factors that influence customer satisfaction and loyalty in e-commerce, and to identify the factors that affect students' academic performance. Additionally, factor analysis has been used to reduce the dimensionality of data in areas such as marketing, finance, and healthcare. The results of factor analysis can also be used as input for other data mining techniques, such as cluster analysis, decision trees, and neural networks. Therefore, factor analysis plays an important role in the pre-processing phase of data mining and is a useful tool for understanding the underlying structure of data.

2.5 Classification Methods

Prior observations with known class labels, such as whether a customer accepted or declined an offer, are used in classification procedures. These models were used to categorise the audience into different groups. Machine learning and statistical methodologies such as neural networks, support vector machines, and decision trees are employed in classification algorithms (classifiers). Table 2.1 shows the results of a literature review on response modelling.

Table 2.1

Target Marketing using Response Modelling Past Works

Author/s	**Year**	**Title**	**Method**
Capatina, Kachour, Lichy, Micu & Codignola	2020	"Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations"	Neural Network
Perlich, Dalessandro, Raeder, Stitelman & Provost.	2015	"Machine learning for targeted display advertising: Transfer learning in action"	CART and CHAID
Zhang, Ji, Huang & Liu	2019	"Sitcom-star-based clothing retrieval for video advertising: a deep learning framework"	Neural Network
Coenen,Swinnen, Vanhoof &Wets	2000	"The improvement of response modelling: Combining rule-induction and case-based reasoning"	Rule-induction and Case-based Reasoning C5 Algorithm

Deichmann, Eshghi & Teebagy Haughton, Sayek	2002	"Application of multiple adaptive splines (MARS) in direct response modeling"	Multiple Adaptive Regression Splines, Logistic Regression
Lo, Cornforth & Chiong	2015	"Effects of training datasets on both the extreme learning machine and support vector machine for target audience identification on twitter"	Support Vector Machine (SVM)
Yu & Cho	2006	"Constructing response model using ensemble based on feature subset selection"	Support Vector Machine (SVM) & Neural Network (MLP)

As shown in the Table 2.1, many models using various methodologies were created to select targets in commercial applications. Response modelling has traditionally been done using statistical approaches, mostly regression techniques. Because of its simplicity, explainability, and accessibility, logistic regression is popular among the researcher (Lo, Cornforth & Chiong, 2015). Decision trees and neural networks are two machine-learning methods that have recently been suggested. Zhang, Ji, Huang, and Liu (2019) conducted a comparison of the response lifts of two commonly utilized decision tree algorithms, namely Classification and Regression Tree (CART) and Chi-Square Automatic Interaction Detector (CHAID). Despite the fact that the two models' tree-generating mechanisms are different, there was no significant variation in the response-lift perspective. A Naive Bayes response model and a C4.5 response model was evaluated by Ling and Li (1998). They conducted experiments using the ADA-boost algorithm on three marketing problems, including, insurance product campaign, loan promotion and bonus programme (Freund & Schapire, 1996) to improve each base model's performance. All of the experiments were created to address issues that can occur throughout the response modelling process, such as class imbalance and the justification of evaluating measures.

2.5.1 Logistic Regression

By generating a set of logistic functions to predict the probability of belonging to a specific output class, a logistic regression model can categorise examples into the most likely group (Parlar & Acaravci, 2017). To implement the classification task for a two-class problem, we can create a binary classifier. Given a set of data points $D = \{(\mathbf{x}_{i,}\ y_i)\}_{i=1}^{N}$, with input data $\mathbf{x}_i \in R^n$ and a corresponding binary target variable $y_i \in \{(0,1)\}$, a logistic regression model would estimate the probability $P\ (y = 1|\ \mathbf{x})$ as

$$P(y=1|\mathbf{x}) = \frac{1}{1+e^{-(w_0+w^T x)}}$$

(2.2)

where **w** is the parameter vector and w_0 is the intercept. Maximum likelihood methods are used to estimate the parameters of equation (1). The main advantage of this method is that it does not require a normal distribution for an input variable, allowing it to cope with qualitative indicators.

Wrapping techniques such as forward selection, backward elimination, and stepwise selection can all be used to include input variables during model training. Forward selection starts with no candidate variables in the model and adds variables using a model comparison criterion until the entry significance level or stopping criterion is met, whereas backward elimination starts with all candidate input variables in the training model and removes them using a model comparison criterion until the stay significance level or stopping criterion is met. The forward model is used to start the stepwise approach, however variables that are already in the model may be removed. This method is repeated until either the stepwise halting conditions or the stay significance level are met.

2.5.2 Decision Tree - J48 Algorithm

The divide-and-conquer principle is used in the construction of a tree structure to solve a problem in the decision tree approach. Customer segmentation and response modelling have both benefited from the use of decision tree. J48 algorithm and random forest are two techniques that have been proposed in the literature for constructing a decision tree model. By choosing the algorithm, the selection of specific methods should be done to look for patterns. For example, in considering accuracy compared to understanding, better methods to use are neural networks and decision trees. Using data mining method in this step, the data mining algorithm may be used repeatedly to obtain the appropriate results. J48 algorithm in decision tree, predictions are made with several if-then conditions that are similar to control statements in various programming languages. In decision tree, its structure consists of root, branch, and leaf nodes (Maulana & Meriska, 2020).

J48 was utilised in this study to build a target marketing decision tree model that used lift as an assessment metric. A decision tree's advantage is that it may provide explanatory guidelines for decision stakeholders, making the decision model's deployment and human resource training easier (Hamed & Ahmed, 2016).

2.5.3 Artificial Neural Networks

Many studies indicate that artificial neural networks (NN) have a superior capacity to accomplish data mining tasks such as classification, clustering, numeric prediction, and optimization. NN has recently been used in marketing decision-making studies, such as response modelling and customer attrition analysis (Zhang, 2018; Abbas, 2015; Aliabadi & Berenji, 2013). NNs are mathematical representations that are prompted by the human brain system's function. The Multilayer Perceptron is one of the many types of network topologies that have been utilised to create networks (MLP). MLP is one of the most often utilised designs for implementing data mining tasks. The MLP (Multi-Layer Perceptron) model is composed of an input layer, one or more hidden layers, and an output layer. Each layer contains multiple neurons, which receive input signals from the previous layer and apply various transfer functions to compute their outputs, which are then passed on to the next layer. By utilizing these transfer functions, NNs (neural networks) are capable of modeling non-linear relationships within the dataset, making them effective for solving non-linear classification problems. Figure 2.3 depicts an MLP with one hidden layer and one output unit, which is often useful for binary classification problems due to the logistic transfer function's output being constrained to the [0, 1] range.

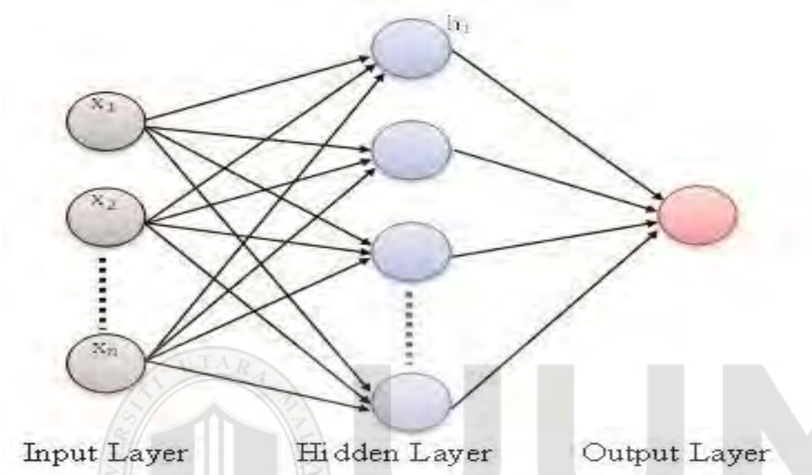


Figure 2.3 An Example of Neural Network Architecture

2.5.4 Support Vector Machines (SVM)

In recent years, the Support Vector Machine (SVM) has become a popular machine learning method for pattern recognition in classification and numeric prediction tasks due to its numerous benefits. Several studies, including Li (2019), Ahmed and Rajaleximi (2019), and Zhu, Qiu, Ergu, Ying, and Liu (2019), have employed this revolutionary approach. The SVM does not require data to follow any particular distribution; however, statistical methods mostly need to do. The SVM algorithm creates a classifier hyperplane based on a limited number of important examples, known as support vectors, which are selected using upper and lower bounds. The SVM applies the principle of structural risk minimization to construct the classifier, instead of empirical risk minimization, which ensures a higher level of robustness. Additionally, the SVM can facilitate the creation of a hyperplane to distinguish between different class instances by employing non-linear kernel functions that map the original data space to a high-dimensional data space. When using the SVM, small samples are allowed. According to previous studies, SVM outperforms a variety of benchmark models, including traditional statistical models (Lo, Cornforth & Chiong, 2015). Only a few researches have looked at SVM model performance in customer targeting and response modelling in marketing applications.

2.6 Classifier Performance and Evaluation Metric

Evaluation measures play a vital role in guiding the classifier modeling as well as in assessing the classification performance (Gu, Cai, Zhu, & Huang, 2008). Confusion matrix is a detailed table arrangement that is used to visualize the algorithm performance (Wang, 2011). The matrix is showed in columns and rows. The cases in each column are from a predicted class, while the cases in the row are from an actual class. Confusion matrix are usually used in supervised learning. However, in the case of unsupervised learning another matric called a matching matric is usually used.

A variety of criteria have been used to assess the efficiency of imbalance dataset. The confusion matrix as presented in Table 2.2, are used in the majority of investigations (Tang & Zhang, 2009). This section presents the main evaluation metrics for class imbalance learning.

Table 2.2

Confusion Matrix

Predicted	**Actual**	
	Positive Class	**Negative Class**
Positive Class	TP	FP
Negative Class	FN	TN

2.6.1 Accuracy

In order to assess predictive accuracy and consistency of the outcomes of the training and testing, the dataset was evaluated using evaluation metrics. A critical criterion for evaluating a classification algorithm's success on a particular dataset is the use of a meaningful measure of general stability (Brodersen, Ong, Stephan, & Buhmann, 2010). To evaluate the performance of the classifiers, 3 popular metrics were used as suggested by Chicco and Jurman,(2020) and Weng and Poon (2008).

One of the commonly used evaluation metrics in traditional applications is accuracy. It is calculated by determining the total number of correctly predicted "High risk" (true positive- true High risk) and "Low risk" (true negative- true Low risk) categories and dividing it by the total number of categories. Equation 2.3 can be used to collect the accuracy formula.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$$

(2.3)

However, because accuracy evaluation is insufficient in imbalanced datasets, many practitioners have noticed that the recall of the minority class is frequently zero, resulting in no classification rules being developed for the minority class (Stefanowski & Wilk, 2006; Tang & Zhang, 2009)

Specifically, from Table 2.1, it is likely to obtain four performance indicators that assess classification quality for both the positive and the negative classes separately (López, Fernández, Torres, & Herrera, 2012):

1. **True positive rate,** $TP_{rate} = \frac{TP}{TP + FN}$

   (2.4)

   - It is the current percentage of positive instances that are appropriately classified as positive.

2. **True negative rate,** $TN_{rate} = \frac{TN}{FP + TN}$

   (2.5)

   - It is the current percentage of negative instances that are appropriately classified as positive.

3. **False positive rate,** $FP_{rate} = \frac{FP}{FP + TN}$

   (2.6)

   - It is the current percentage of negative instances that have been misclassified as positive.

4. **False negative rate,** $FN_{rate} = \frac{FN}{TP + FN}$

   (2.7)

- It is the current percentage of positive cases that are mistakenly categorized as negative.

2.6.2 Sensitivity

The sensitivity rate is the proportion of accurately predicted "High risk" (true positive- true High risk) to the total number of successfully anticipated "High risk" and mistakenly predicted "Low risk" predictions (false negative- false Low risk). The formula of the rate can be understood as the correctly predicted rate “High risk” over the total of actual “High risk”. The sensitivity rate is calculated using Equation 2.8 as follows:

$$Sensitivity = \frac{TP}{TP+FN} \quad (2.8)$$

2.6.3 Specificity

The specificity rate is the proportion of correctly predicted "Low risk" to the total number of "Low risk" that actually exists. This is clarified in Equation 2.9 as follows:

$$Specificity = \frac{TN}{TN+FP} \quad (2.9)$$

In order to assess predictive accuracy and consistency of the outcomes was evaluated using evaluation metrics. A critical criterion for evaluating a classification algorithm's success on a particular dataset is the use of a meaningful measure of general stability (Brodersen, Ong, Stephan, & Buhmann, 2010).

2.7 Conclusion

Chapter 2 of this thesis provides an extensive review of the literature on customer response modeling in target marketing, knowledge discovery in database process, classifier performance and evaluation metrics. The chapter starts with a historical overview of the topic, followed by an in-depth analysis of the current state of the literature. It highlights the essential aspects of the existing body of knowledge related to customer response modeling in target marketing, focusing on the use of different techniques such as logistic regression, decision trees, and neural networks.

The chapter also discusses the process of knowledge discovery in databases and its applications in target marketing. It reviews the literature on classifier performance and evaluation metrics, including accuracy, specificity and sensitivity, which are commonly used in assessing the performance of classification models. The strengths and weaknesses of previous studies are evaluated, and gaps in the existing knowledge are identified, such as the need for more research on the application of data mining techniques in customer response modeling, factor analysis for variable selection, and the applicability of several evaluation metrics in selecting the best classification models.

In summary, this chapter provides a critical review of the existing literature on customer response modeling in target marketing, highlighting the importance of various techniques and methods used in the process. It also identifies gaps in the literature and suggests potential areas for future research to improve the understanding of customer response modeling in target marketing.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter provides an overview of the research methodology, outlining the steps taken to develop a predictive model for customer response to personal loan products. The primary objective of this study was to utilize two classification models, Logistic Regression and Decision Tree, to develop the predictive model. To reduce the complexity of the model, variable selection was performed using Factor Analysis on the dataset. In order to minimize sampling bias, 5-fold and 10-fold cross-validation techniques were employed to split the data between training and testing for each model. Performance values were then calculated to identify the best model for predicting customer response to personal loan products. The study utilized data cleaning, variable selection, and model development and evaluation, with each step carried out using specific tools such as SPSS, Weka, and R programming. The dataset used for this study contained information from 10,952 customers and 110 variables, obtained from a commercial bank in Malaysia. Overall, this chapter provides a clear and concise overview of the research methodology used in this study to develop the customer response prediction model.

3.2 Research Methodology

The methodology of this thesis is consistent with the methods in the Knowledge Discovery Database (KDD) process used by Fayyad, Shapiro, & Smyth (1996). However, the method of this research was not necessarily realised in the order that is being presented. In fact, the method was somewhat iterative as expected by the KDD methodology. Figure 3.1 depicts the general KDD process employed in this research.

Detailed explanation for each phase in the KDD processes is discussed in the following subsections. Figure 3.2 exhibits the research framework for this study.

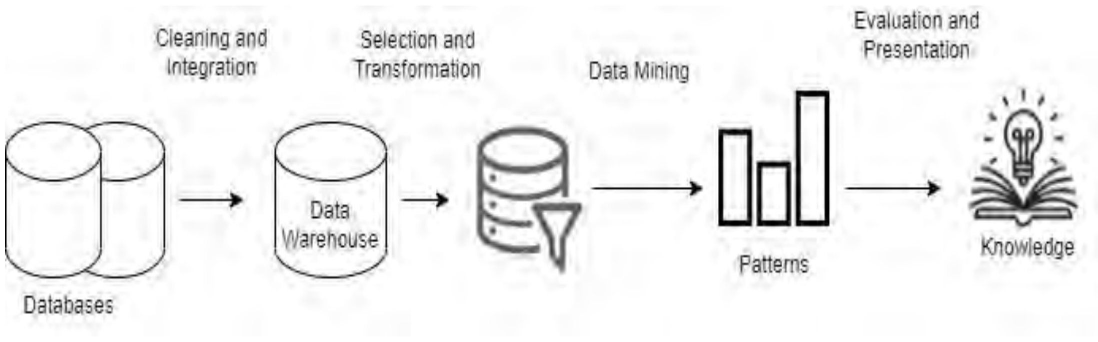


Figure 3.1. Knowledge Discovery Process reproduce from (Han, Kamber and Pei, 2011)

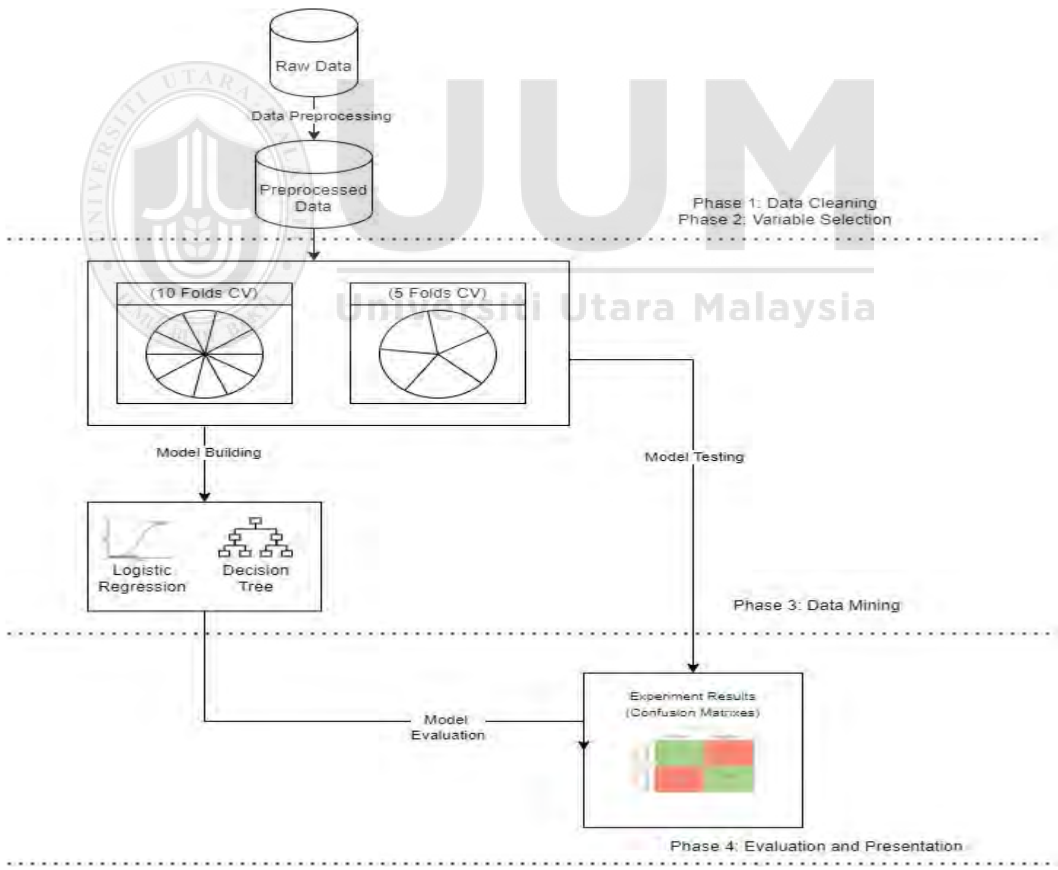


Figure 3.2. Research Framework

3.3 Phase 1: Data Cleaning

Data cleaning is an essential step in target marketing research as inaccurate or inconsistent data can lead to flawed conclusions and ineffective marketing campaigns. It involves the process of detecting, correcting, and removing inaccuracies, inconsistencies, and incomplete data in order to produce quality and consistent data for analysis. One common issue in target marketing research is missing data, where certain data points are not available for some customers. Various data cleaning techniques can be used to address missing data, including imputing missing values based on other available data or removing observations with missing data if imputation is not possible.

In this study, data cleaning was conducted for each variable in the dataset to ensure the accuracy and reliability of the data used for predicting customer response to personal loan products. The dataset used in this study contained information from 10,952 customers and 110 variables. Data cleaning was conducted to remove unwanted variables, resulting in only 35 variables being selected, including one dependent variable. The variables that were discarded consisted of 50% missing values, and therefore were deemed unsuitable for the analysis. This stage was critical in ensuring that only accurate and reliable data was used for predicting customer response to personal loan products, which was the first objective of the research.

Data cleaning is a crucial step in target marketing research methodology, as it helps to ensure the accuracy and reliability of the data used in the analysis. By using accurate and reliable data, researchers can create more effective and targeted marketing campaigns that are more likely to drive sales and improve customer loyalty. Overall, the data cleaning process played an important role in ensuring the quality of the data used for this study, and helped to increase the validity and reliability of the findings. It is therefore recommended that data cleaning be conducted as a standard practice in target marketing research.

3.4 Phase 2: Variable Selection

In order to achieve second objective, the variable selection task is utilize to identify irrelevant features that can be used to predict the class variable. Unnecessary and/or duplicate characteristics would make using the whole dataset as input needlessly complicated. Many data mining algorithms struggle to distinguish between important and irrelevant features (Roiger & Gaetz). Assume that the irrelevant attributes are kept in the dataset. In that situation, these attributes may obstruct the data mining process, lowering the quality of prediction models and causing issues such as overfitting (Cheng et al., 2007). The data mining process will become more complex if a dataset has too many attributes. In solving these issues, the identification of suitable and important attributes for the classification task has to be optimum.

3.4.1 Factor Analysis

In this study, factor analysis was employed to decrease the data's dimensionality. This approach improves comprehensibility while minimizing information loss, enabling researchers to uncover the most significant features of the dataset. Dimensionality reduction is vital in machine learning, particularly when dealing with a large number of features. Factor analysis is a multivariate analysis method that reduces dimensions based on anticipated correlations among interval-scaled variables. The aim is to test as few variables as possible that are both meaningful and independent, and to account for as much of the variables' variance in question as possible.

3.4.1.1 Procedure for Selecting Variables

The following steps are the procedure for selecting variables by utilizing Factor Analysis.

Step 1: Calculating the Correlation Matrix

A correlation matrix can be used to quickly summarise the correlations between variables. The correlation matrix entails computing the bivariate correlations of the individual variables (according to Pearson). The effect of a correlation of 0.500, according to Cohen (1988), might be regarded as substantial. However, the correlation matrix is insufficient when determining whether correlated variables can be explained through a common factor. However, if certain variables do not correlate, or barely minimally correlate, with others, the option of assigning a common component can be ruled out for definite.

Step 2: Suitability of the Correlation Matrix

Based on the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) measure, if the value is more than.60, the data is eligible for factor analysis. The KMO assesses how strongly the variables being examined correlate and assumes values between 0 and 1. According to Backhaus, Erichson, Plinke, and Weiber (2006), the magnitude of the KMO value (as shown in Table 3.1) may be calculated as follows:

Table 3.1

The KMO Value Description

KMO Value	Description
KMO>= 0.9	Marvellous
KMO>= 0.8	Meritorious
KMO>= 0.7	Middling
KMO>= 0.6	Mediocre
KMO>= 0.5	Miserable
KMO < 0.5	Unacceptable

Another relevant criterion for assessing if the data are adequate for factor analysis is Bartlett's Sphericity Test. In addition, anti-image matrices that rely on the decomposition of the two variances can also be utilized to evaluate whether individual variables could be included in the factor analysis. The anti-image matrices measure the variance portion of one variable whether it can explain the correlating variables (image) associated with the inexplicable variance portion (anti-image). The variables should be included in the factor analysis if the anti-image matrix values are low.

Step 3: Determining the Communalities

The communalities describe the variance sections of the individual variables that can be explained by the extracted components (total of the squared loadings of a variable across all factors). There are various methods for extracting the components, each with its own method for determining communalities. Conducting a principal components analysis assumes that the extracted factors fully explain the variance of the variables.

Step 4: Extracting the Factors

A factor analysis involves mathematically computing matrices and vectors to create a new frame of reference (system of coordinates) with fewer dimensions (axes) in which the analyzed data is embedded as deeply as feasible. This entails graphing the variables on the axes. The vectors can be shown as images in a two-dimensional framework (tested variables). The closeness of the linked vectors in the system of coordinates relates to the degree of the correlation between the variables. The narrower the angle of the linked vectors, the higher the connection between two variables.

Step 5: Determining the Factors

There are no set rules for determining the ideal number of components to extract in factor analysis. However, the Kaiser criterion is often used to identify relevant factors, based on whether their eigenvalue - the amount of variance in all the variables that can be explained by the factor - exceeds one. If the eigenvalue is greater than one, it indicates that the factor can explain more variance than a single tested variable in the factor analysis, as explained by Moosbrugger and Engel (2007).

Step 6: Determining the Factor Values and Factor Rotation

After establishing the number of factors and extracting them, the following step calculates the magnitude of factor loadings (correlations between the variables being tested and the factors). Every variable should ideally have a significant correlation with only one component, making the content of the factors easy to comprehend. Loadings of 0.50 or more, on the other hand, are usually regarded as significant.

3.5 Phase 3: Data Mining

Choosing a Data Mining Algorithm

The J48 Decision Tree algorithm and Logistic Regression algorithms were utilized in this study. By using these two mining techniques, the third objective of this research was achieved which is assist to develop the customer response prediction model for personal loan product using logistics regression and decision tree.

3.5.1 k-Folds Cross validation

Cross-validation methods are a crucial component of data mining classification issues. Cross-validation is a method for analysing a model's performance and generalisation potential. The 5-fold and 10-fold cross-validation methods are the two that are most often employed. The size of the dataset and the available computational resources determine which cross-validation method is most suited. In general, 5-fold cross-validation is better suitable for bigger datasets, whereas 10-fold cross-validation is preferable for small to medium-sized datasets. By guaranteeing that the model is tested on data that was not included in its training, cross-validation approaches assist to decrease overfitting and boost the accuracy of the model. The capacity to give a more thorough evaluation of the model's performance and increase the dependability of the results produced justifies the use of 5-folds and 10-folds cross-validation for data mining classification issues.

For this research, a non-exhaustive cross-validation k-fold method was utilized with k values of both 10 and 5. This involved randomly dividing the original sample into k equal-sized sub-samples, with one sub-sample being removed to serve as the validation data for testing, and the remaining k-1 sub-samples being used as training data. This process was repeated k times (in folds), with each sub-sample being used once as validation data. The final estimate was derived by averaging the results of all k folds. To assess the accuracy of the classifications, the study constructed 4 x 4 confusion matrices for each of the 5 and 10 folds. These matrices allowed for the calculation of measurement metrics such as accuracy, sensitivity, and specificity.

3.5.2 Logistic Regression

Regression analysis is an essential component of any data analysis process that aims to describe the relationship between a response variable and explanatory variable/s. The response variable can have multiple discrete potential values. The objective of building regression models is to identify the best-fitting, most concise, and acceptable model that describes the association between a set of independent (predictor/explanatory/output) variables and a dependent (response/input) variable.

Regression models are employed when the dependent variable is dichotomous/binary/binomial (binary logistic regression) or has more than two levels (multinomial logistic regression), and the independent variables can be of any type. The purpose of regression modeling is to select the optimal dependent/input variables within the context of the problem. According to Hosmer & Lemeshow (1989), this involves two fundamental tasks: "1) a basic plan for selecting the variables for the model and 2) a set of methods for assessing the adequacy of the model in terms of its individual variables and overall fit."

This thesis utilizes a logistic regression model to choose the most parsimonious model that describes the data. This is achieved by reducing the number of variables to produce a more generic model. A model with numerous variables will have a significant estimated standard error and rely heavily on the observed data (Hosmer & Lemeshow, 1989). Chapter 4 describes the building of the logistic regression model and variable selection. The logistic regression analysis characterizes the association between the target/class variable and the set of input variables, commonly referred to as covariates in statistical terms (Hosmer & Lemeshow, 1989).

The logistic regression model is defined in Equation 3.1.

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \, , \tag{3.1}$$

where;

y = Natural logarithm of the odds ratio,

$\beta_0, \beta_1, \ldots, \beta_k$ = Coefficients of the input attributes,

ε = Error variable,

Y = Dichotomous class attribute,

$x_1, x_2 \ldots, x_k$ = Input attributes

The coefficients are estimated using a statistical technique called **maximum likelihood estimation.**

The next step is to analyse the importance of the variables in the model after calculating the coefficients and fitting the model. This determines if the model's input variables are significantly connected to the class variable. Finally, the use of logistic regression adds to the classification power of the input and target variables. The data will be used to create and fit a logistic regression model, and the significance of the coefficients will be examined.

3.5.3 Decision Tree – J48

The J48 algorithm is a popular machine learning algorithm based on the C4.5 method by Quilan. It is the upgraded version of the ID3 classification machine learning algorithm. J48 is built in a top-down style. It starts at the top, with the whole training dataset. J48 uses a measure called 'information gain" to choose which attribute to split at each stage (Saravanan & Gayathri, 2018). J48 is a non-linear solution and has interactions between independent variables. The J48 capable to handle collinearity and does better when there is a large set of categorical values in the training data. In addition, J48 does well when the scenario demands an explanation over the decision (Almustafa, 2020).

3.5.3.1 Information Gain

Information gain splitting rule measures the probable decrease in entropy by splitting the instances based on attributes. In general, the probable information gain rule is the change in information entropy. The information gain assumes S is a set of cases, A is an attribute, S_v is the subset of S with $A = v$, and Values (A) is the set of all potential values of A, as clarified by Bahety, (2014).

$$Gain(S.A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

(3.2)

where values A refers to all possible values for instance A, and S_v is the subset of S for which the instance A has value v.

3.6 Phase 4: Evaluation and Presentation

In order to assess predictive accuracy and consistency of the outcomes of the training and testing, the dataset was evaluated using evaluation metrics. A critical criterion for evaluating a classification algorithm's success on a particular dataset is the use of a meaningful measure of general stability. To evaluate the performance of both classifiers 3 popular metrics were used as suggested by Chicco & Jurman,(2020) and Weng & Poon (2008). These metrics were calculated using the confusion matrix (see Table 2.2 (Chapter 2)). Following are the performance measures used in evaluating the performance of the classifiers namely; accuracy, sensitivity and specificity. Detailed explanation of the performance measure is provided in Section 2.6 (see Chapter 2).

3.7 Conclusion

This chapter presents the methodology used in this research, which consists of four phases such as data cleaning, variable selection, data mining, and evaluation and presentation. Each phase is described in detail, with an explanation of the tools and techniques used to achieve the research objectives. The first phase, data cleaning, aimed to eliminate missing data from the dataset with a high percentage of missing values. In the second phase, variable selection, the most significant predictors of customer response to personal loan products were identified using factor analysis to reduce the model's complexity and select the most relevant variables. In the third phase, data mining techniques were used to develop predictive models for customer response using two prominent classification techniques, logistic regression, and decision tree. The fourth phase involved the evaluation of the models using performance metrics such as accuracy, sensitivity, and specificity. The best performing model was selected and presented in the final stage of the research. Overall, this chapter provides a comprehensive overview of the methodology used in this research. Each phase was carried out using appropriate tools and techniques, ensuring that the research objectives were achieved, and the results obtained were accurate and reliable.

CHAPTER FOUR

ANALYSIS AND FINDINGS

4.1 Introduction

The main goal of this study was to predict customer response for personal loan product by using two popular data mining classifiers namely the logistic regression model and decision tree model. 10-folds and 5-folds cross validation techniques were used to minimize the sampling bias in splitting the data between training and testing for each model built. 10-folds and 5-folds was used by previous researchers of performance values were calculated to identify the best model that can help to predict customer response for personal loan.

Section 4.2 briefly presented the descriptive information about the data, followed by variables selected to be included into the classification models in Section 4.3. Section 4.4 discussed the finding by utilizing Logistic Regression model and Section 4.5 dedicated for Decision Tree model results. In Section 4.6, the performance of both models was presented followed by a conclusion section in 4.7.

4.2 Descriptive Analysis

Table 4.1 and Figure 4.1 summarizes the number of genders. Figure 4.1 shows the proportion of genders within the dataset. There is more male which is 54% compared to 46% of female participants.

Table 4.1

Gender

Gender	Total
Female	5060
Male	5892

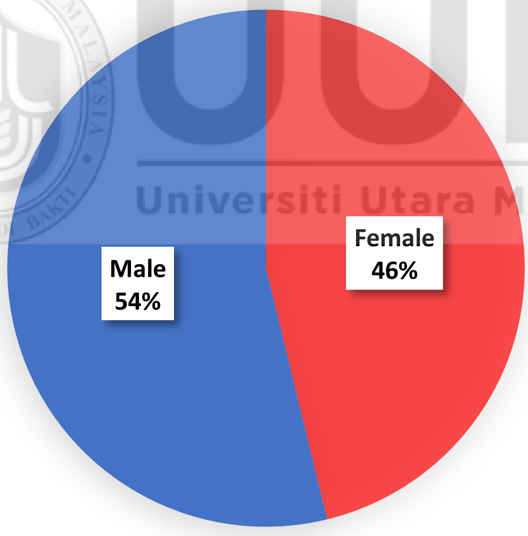


Figure 4.1. Gender

Table 4.2 and Figure 4.2 shows the total number of marital status within the dataset. Most of the customers were single which is 68% and only 32% of the participants were married.

Table 4.2

Marital Status

Marital Status	Total
Married	3557
Single	7395

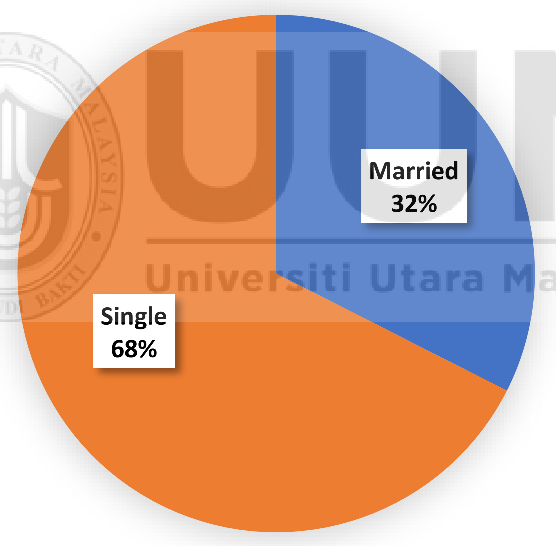


Figure 4.2. Marital Status

Table 4.3 and Figure 4.3 summarizes the number of races of the bank customers. The most bank customers were Malay, 59.99% while Chinese were 23.71%. Moreover, Indian customer were only 8.07% and 8.23% were unknown.

Table 4.3

Race

Race	Total
Malay	6570
Chinese	2597
Unknown	901
Indian	884

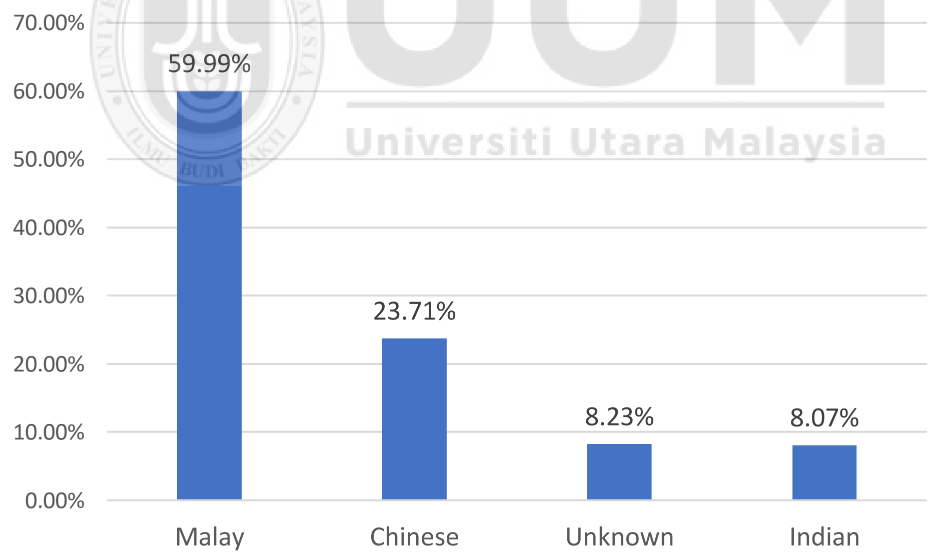


Figure 4.3. Race

Table 4.4 summarizes the number of customers' segment. Figure 4.4 shows the percentage of customer's segment. The most participants have MM as their segment, 71.62%. It was followed by participants who have LMA as their segment which is 22.24%. While only 6.05% and 0.08% of participants have UMA and HNI as their segment respectively.

Table 4.4.

Customer's Segment

Customer's Segment	Total
MM	7844
LMA	2436
UMA	663
HNI	9

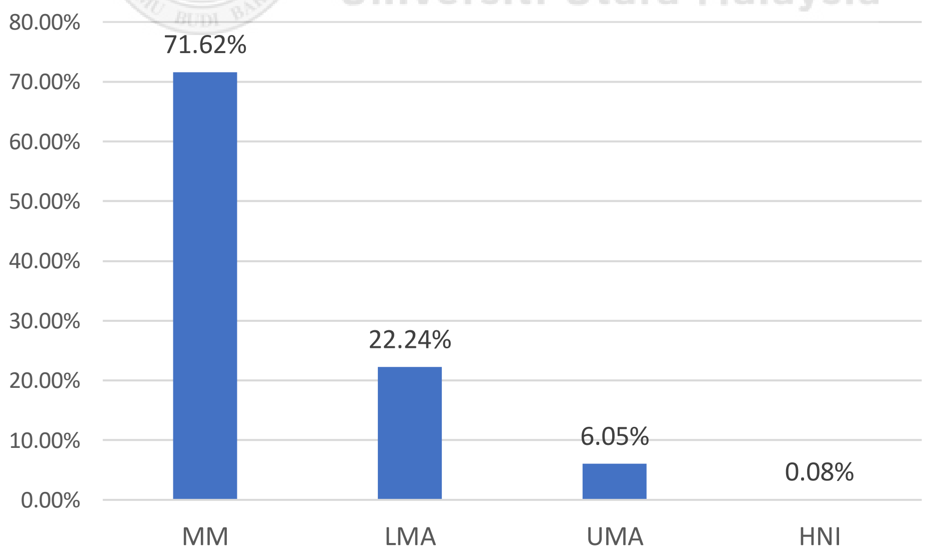


Figure 4.4. Customer's Segment

Table 4.5 summarizes the number of dependants. Figure 4.5 shows the percentage of number of dependants. Most of the participants have 0 dependant, 93.4%. Only 1.49%, 2.07% and 3.04% of the participants have 1, 2 and 3 dependants respectively.

Table 4.5

Number of Dependant

Number of Dependant	Total
0 dependent	10229
1 dependent	163
2 dependents	227
3 dependents	333

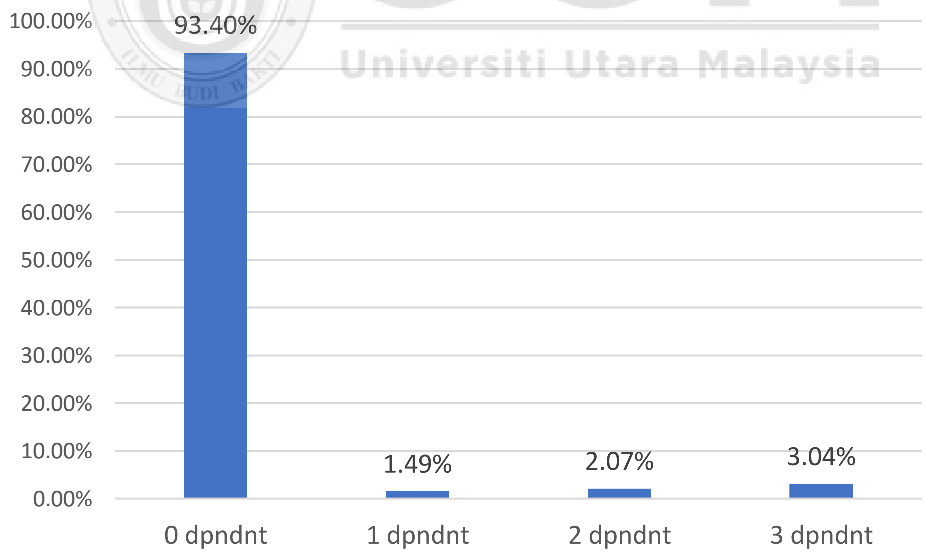


Figure 4.5. Number of Dependant

Table 4.6 summarizes the number of educations of bank customers. Figure 4.6 shows the education of participants. 79.67% of the participants' education are not available for this survey. Yet, participants who have school as their education were 9.97%, grad and post grad as their education were 9.31% and 1.05% respectively.

Table 4.6

Education

Education	Total
Not Available	8725
School	1092
Grad	1020
Post Grad	115

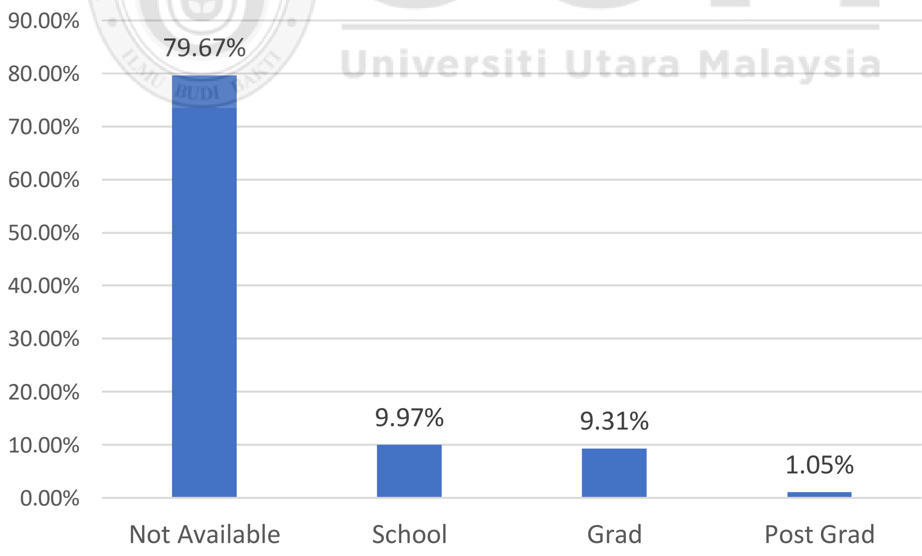


Figure 4.6. Education

Table 4.7 dan Figure 4.7 summarizes the number of occupations of the bank customers. The most customers are non-exec which were 26.1% and 23.45% of the participants were students. The customers who had already retired recorded 3.69% and 4.22% of the participants were doing business.

Table 4.7

Occupation

Occupation	Total
Non exec	2858
Student	2568
Others	1897
Exec and above	1473
Prof govt	728
Housewife	562
Business	462
Retired	404

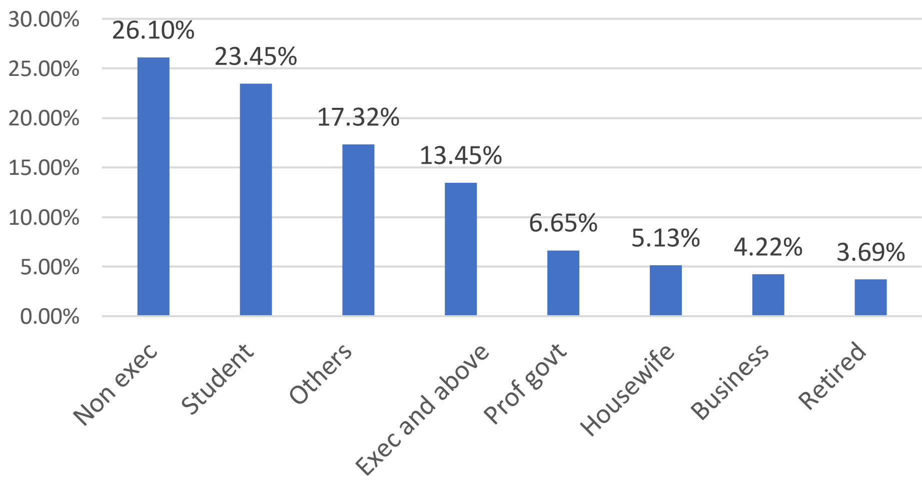


Figure 4.7. Occupation

Table 4.8 and Figure 4.8 summarizes the total number of balances for saving account of the bank customers. Most participants have med number of total balances for saving account which is 29.88%.

Table 4.8

Total Balance for Saving Account

Total Balance for Saving Account	Total
Zero bin	191
Very low	958
Low	2406
Med	3272
High	2571
Very high	1240
Not Applicable	314

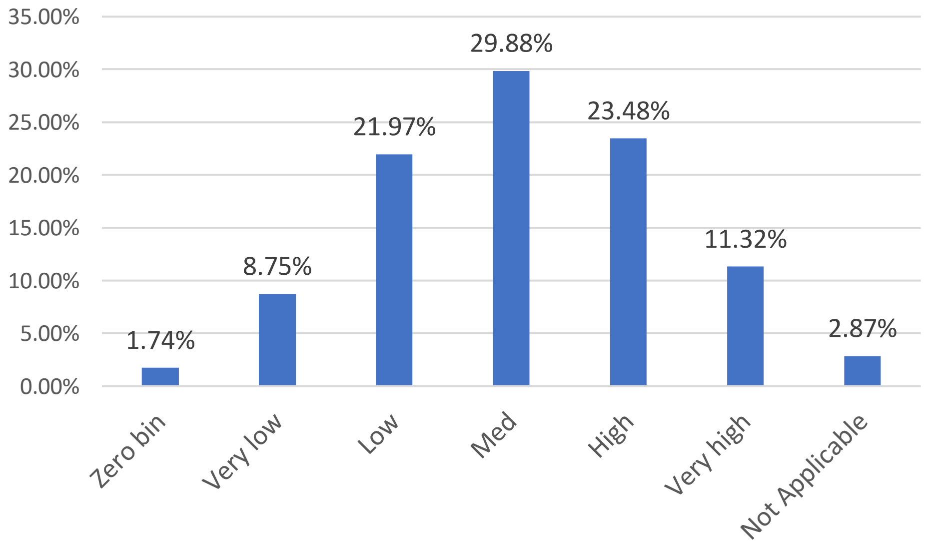


Figure 4.8. Total Balance for Saving Account

Table 4.9 and Figure 4.9 summarizes the total number of CASA Account. There are more participants who have one total number CASA account which is 91% compared to 9% of participants who have GT 1 total number CASA account.

Table 4.9

Total Number of CASA Account

Total Number of CASA Account	**Total**
1	9923
GT 1	1029

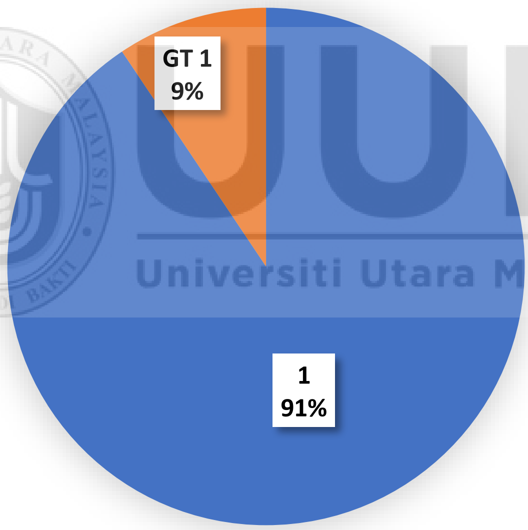


Figure 4.9. Total Number of CASA Account (Current Account and Saving Account)

Table 4.10 and Figure 4.10 summarizes the total number of Principal Credit Card account. There are more participants who have none Principal Credit Card account which is 90% compared to participants who have one and GT 1 total number of Principal Credit Card account.

Table 4.10

Total Number of Principal Credit Card Account

Total Number of Principal Credit Card Account	**Total**
0	9853
1	773
GT 1	326

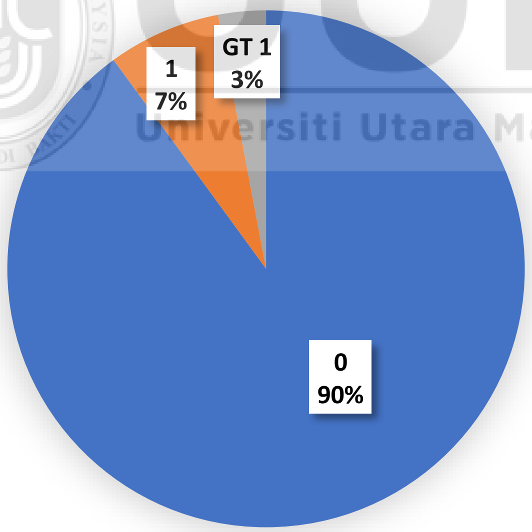


Figure 4.10. Total Number of Principal Credit Card Account

Table 4.11 and Figure 4.11 summarizes the total number of active clicks. There are more participants who have none number of active clicks which is 66% compared to 34% of participants who have one total number of active clicks.

Table 4.11

Total Number of Active Clicks

Total Number of Active Clicks	Total
0	7268
1	3684

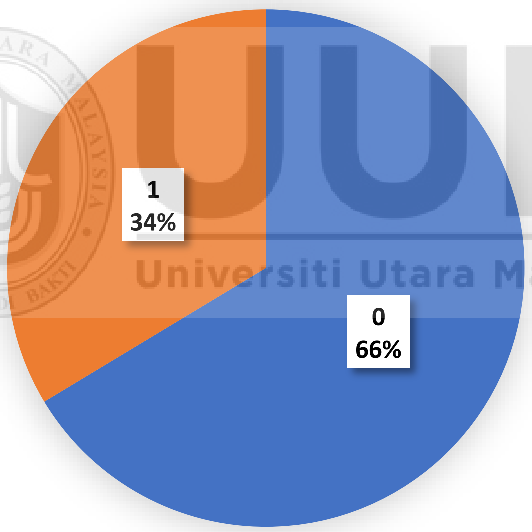


Figure 4.11. Total Number of Active Clicks

Table 4.12 and Figure 4.12 summarizes the total number of debit card. There are more participants who have none number of debit card which is 89% compared to 11% of participants who have one total number of debit card.

Table 4.12

Total Number of Debit Card

Total Number of Debit Card	Total
0	9717
1	1235

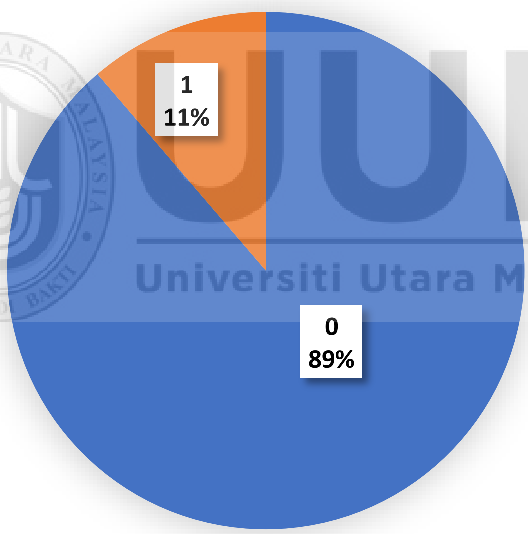


Figure 4.12. Total Number of Debit Card

Table 4.13 and Figure 4.13 summarizes the total number for fixed deposit. There are more participants who have none for fixed deposit which is 95% compared to participants who have one and GT 1 total number for fixed deposit.

Table 4.13

Total Number for Fixed Deposit

Total Number for Fixed Deposit	Total
0	10444
1	220
GT 1	288

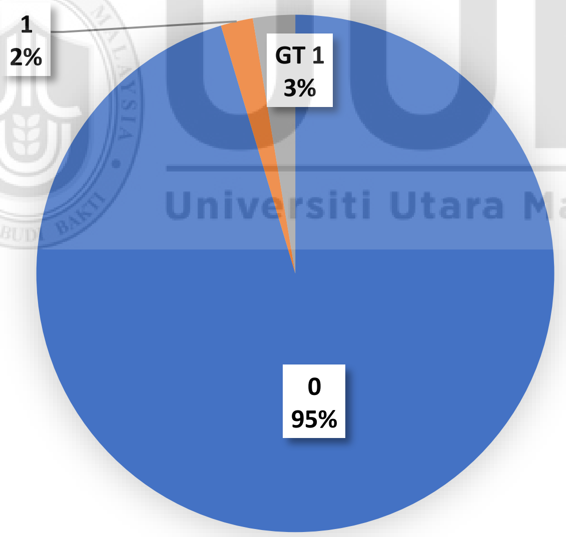


Figure 4.13. Total Number for Fixed Deposit

Table 4.14 and Figure 4.14 summarizes the total number for saving account. There are more participants who have none saving account which is 90% compared to participants who have one and GT 1 total number for saving account.

Table 4.14

Total Number for Saving Account

Total Number for Saving Account	**Total**
0	314
1	9916
GT 1	722

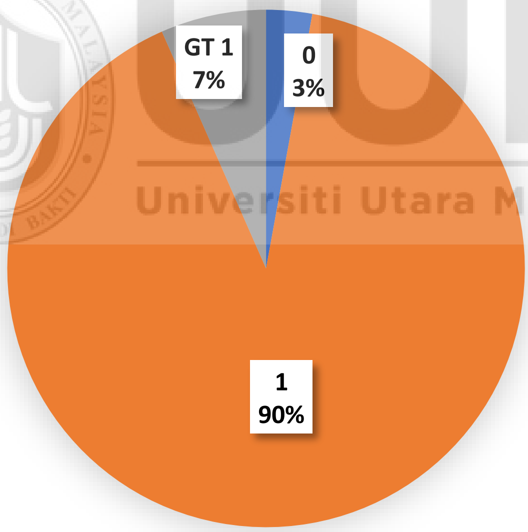


Figure 4.14. Total Number for Saving Account

Table 4.15 and Figure 4.15 summarizes the total product holding. There are more participants who have one total product holding which is 65.67% compared to participants who have 2, 3 and 4 or more total product holding.

Table 4.15

Total Product Holding

Total Product Holding	Total
1	7192
2	2153
3	924
4 or more	683

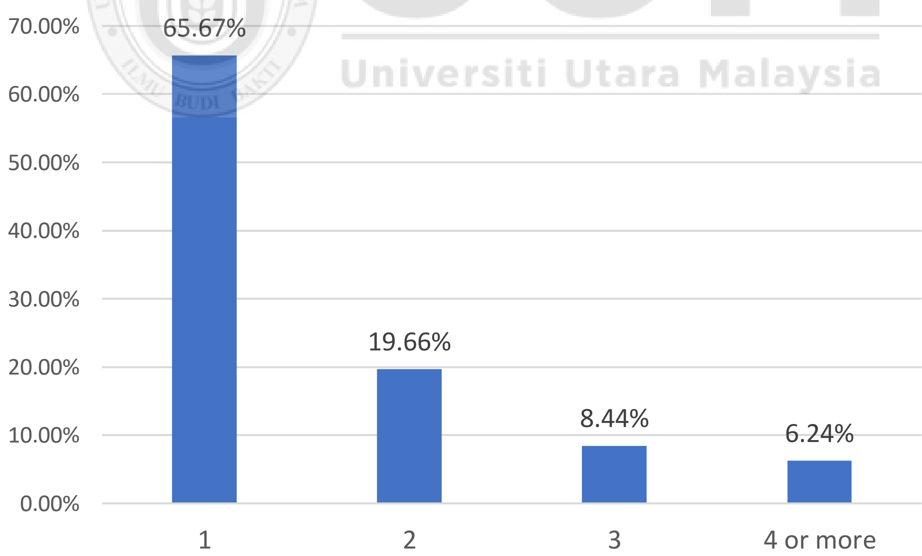


Figure 4.15. Total Product Holding

Table 4.16 and Figure 4.16 summarizes average product holding count for Amanah Saham Bumiputra. There are more participants who have no product holding count for Amanah Saham Bumiputra which is 96% compared to 4% of participants who have product holding count for Amanah Saham Bumiputra.

Table 4.16

Average Product Holding Count for Amanah Saham Bumiputra

Average Product Holding Count for ASB	Total
No	10565
Yes	387

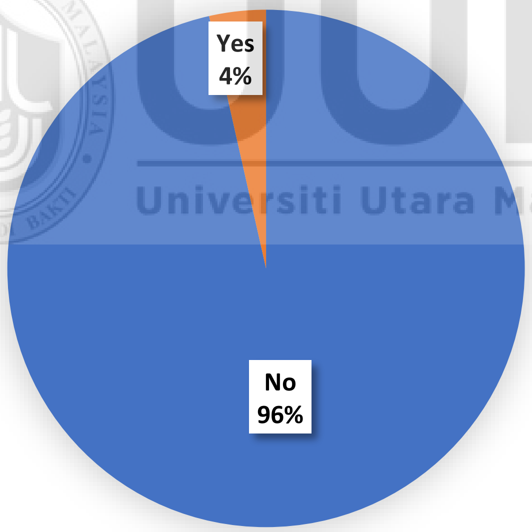


Figure 4.16. Average Product Holding Count for Amanah Saham Bumiputra

Table 4.17 and Figure 4.17 summarizes average product holding count for Business Premises Loan. There are more participants who have no product holding count for Business Premises Loan which is 99.7% compared to 0.3% of participants who have product holding count for Business Premises Loan.

Table 4.17

Average Product Holding Count for Business Premises Loan

Average Product Holding Count for BPL	Total
No	10917
Yes	35

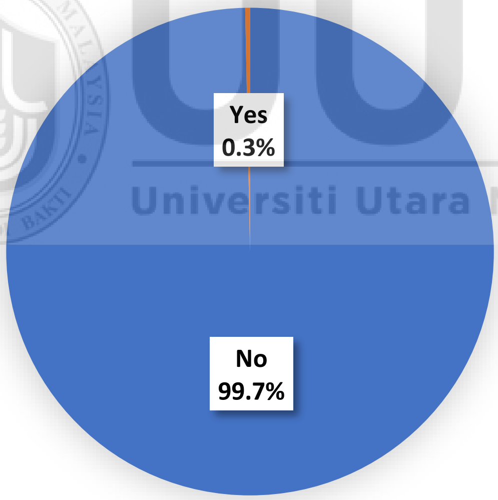


Figure 4.17. Average Product Holding Count for Business Premises Loan

Table 4.18 and Figure 4.18 summarizes average product holding count for Balance Transfer. There are more participants who have no product holding count for Balance Transfer which is 99.9% compared to 0.1% of participants who have product holding count for Balance Transfer.

Table 4.18

Average Product Holding Count for Balance Transfer

Average Product Holding Count for BT	**Total**
No	10940
Yes	12

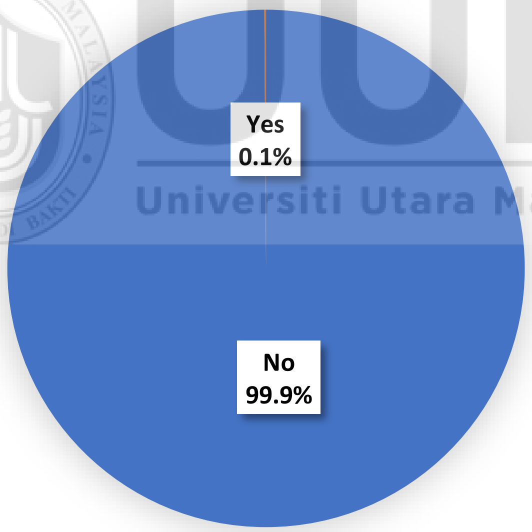


Figure 4.18. Average Product Holding Count for Balance Transfer

Table 4.19 and Figure 4.19 summarizes average product holding count for Current Account. There are more participants who have no product holding count for Current Account which is 94% compared to 6% of participants who have product holding count for Current Account.

Table 4.19

Average Product Holding Count for Current Account

Average Product Holding Count for CA	**Total**
No	10337
Yes	615

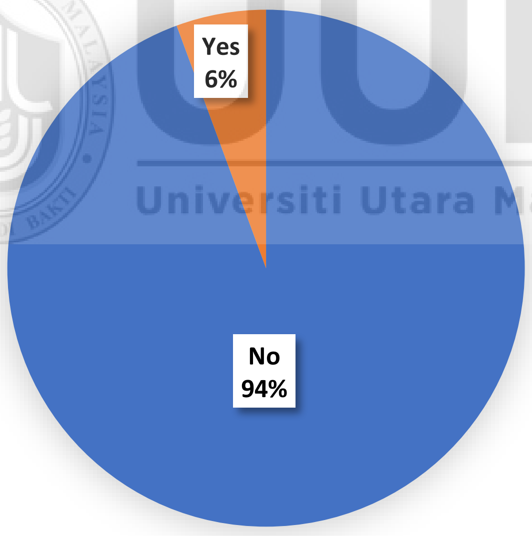


Figure 4.19. Average Product Holding Count for Current Account

Table 4.20 and Figure 4.20 summarizes average product holding count for Credit Card (classic). There are more participants who have no product holding count for Credit Card (classic) which is 99% compared to 1% of participants who have product holding count for Credit Card (classic).

Table 4.20

Average Product Holding Count for Credit Card (classic)

Average Product Holding Count for CC (classic)	Total
No	10855
Yes	97

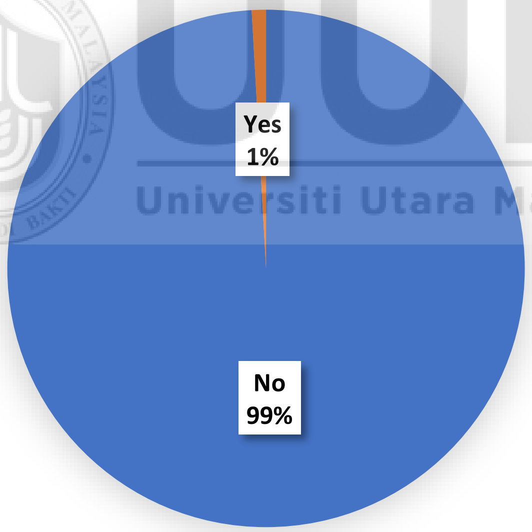


Figure 4.20. Average Product Holding Count for Credit Card (Classic)

Table 4.21 and Figure 4.21 summarizes average product holding count for Credit Card (any type). There are more participants who have no product holding count for Credit Card (any type) which is 90% compared to 10% of participants who have product holding count for Credit Card (any type).

Table 4.21

Average Product Holding Count for Credit Card (any type)

Average Product Holding Count for CC (any type)	**Total**
No	9853
Yes	1099

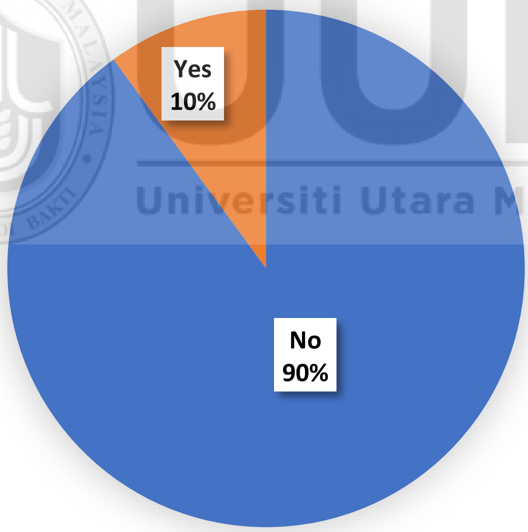


Figure 4.21. Average Product Holding Count for Credit Card (any type)

Table 4.22 and Figure 4.22 summarizes average product holding count for Credit Card (gold). There are more participants who have no product holding count for Credit Card (gold) which is 97% compared to 3% of participants who have product holding count for Credit Card (gold).

Table 4.22

Average Product Holding Count for Credit Card (gold)

Average Product Holding Count for CC (gold)	Total
No	10618
Yes	334

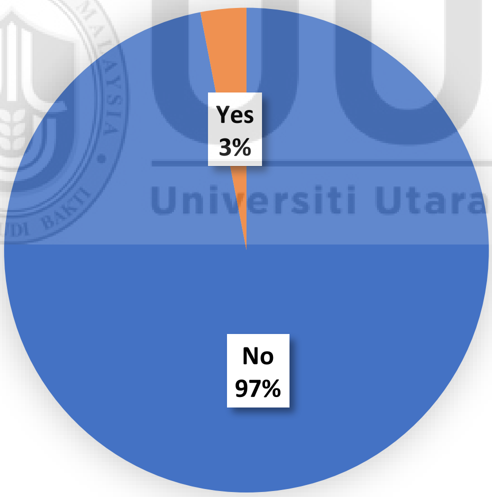


Figure 4.22. Average Product Holding Count for Credit Card (gold)

Table 4.23 and Figure 4.23 summarizes average product holding count for Credit Card (platinum). There are more participants who have no product holding count for Credit Card (platinum) which is 97% compared to 3% of participants who have product holding count for Credit Card (platinum).

Table 4.23

Average Product Holding Count for Credit Card (platinum)

Average Product Holding Count for CC (platinum)	**Total**
No	10585
Yes	367

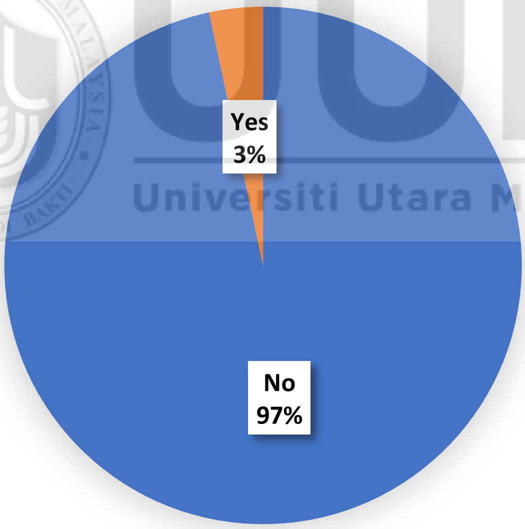


Figure 4.23. Average Product Holding Count for Credit Card (platinum)

Table 4.24 and Figure 4.24 summarizes average product holding count for Cashlite. There are more participants who have no product holding count for Cashlite which is 99% compared to 1% of participants who have product holding count for Cashlite.

Table 4.24

Average Product Holding Count for Cashlite

Average Product Holding Count for Cashlite	Total
No	10839
Yes	113

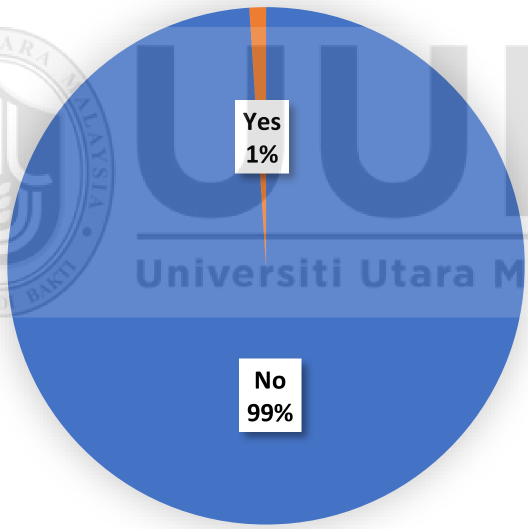


Figure 4.24. Average Product Holding Count for Cashlite

Table 4.25 and Figure 4.25 summarizes average product holding count for Insurance (credit life). There are more participants who have no product holding count for Insurance (credit life) which is 98% compared to 2% of participants who have product holding count for Insurance (credit life).

Table 4.25

Average Product Holding Count for Insurance (credit life)

Average Product Holding Count for Insurance (credit life)	Total
No	10746
Yes	206

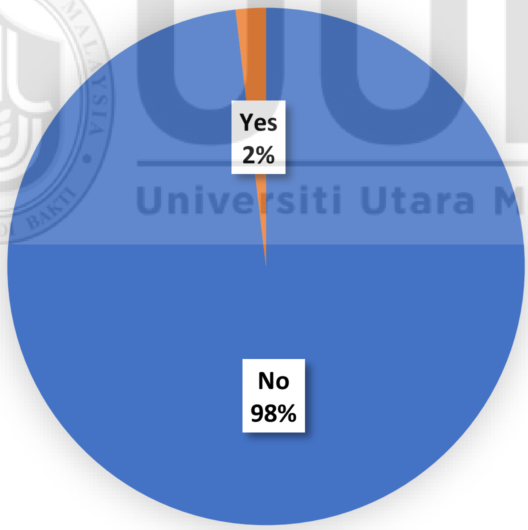


Figure 4.25. Average Product Holding Count for Insurance (credit life)

Table 4.26 and Figure 4.26 summarizes average product holding count for Dual Currency Investment. There are more participants who have no product holding count for Dual Currency Investment which is 100%.

Table 4.26

Average Product Holding Count for Dual Currency Investment

Average Product Holding Count for Dual Currency Investment	**Total**
No	10950
Yes	2

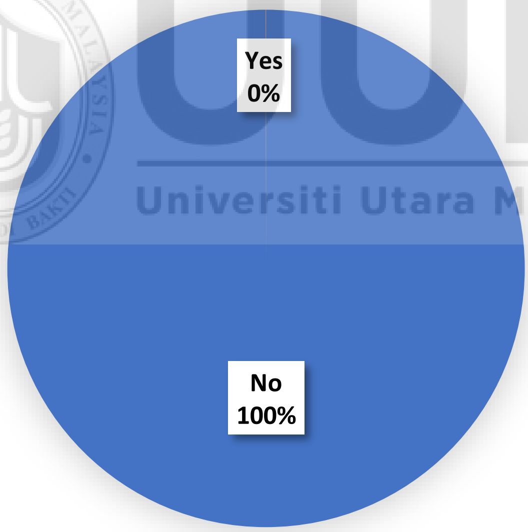


Figure 4.26. Average Product Holding Count for Dual Currency Investment

Table 4.27 and Figure 4.27 summarizes average product holding count for Foreign Currency Current Account. There are more participants who have no product holding count for Foreign Currency Current Account which is 99.8% compared to 0.2% of participants who have product holding count for Foreign Currency Current Account.

Table 4.27

Average Product Holding Count for Foreign Currency Current Account

Average Product Holding Count for Foreign Currency Current Account	**Total**
No	10931
Yes	21

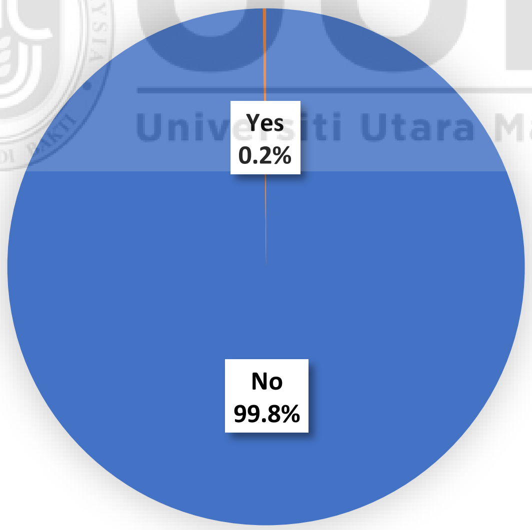


Figure 4.27. Average Product Holding Count for Foreign Currency Current Account

Table 4.28 and Figure 4.28 summarizes average product holding count for Foreign Currency Fixed Deposit. There are more participants who have no product holding count for Foreign Currency Fixed Deposit which is 99.9% compared to 0.1% of participants who have product holding count for Foreign Currency Fixed Deposit.

Table 4.28

Average Product Holding Count for Foreign Currency Fixed Deposit

Average Product Holding Count for Foreign Currency Fixed Deposit	**Total**
No	10943
Yes	9

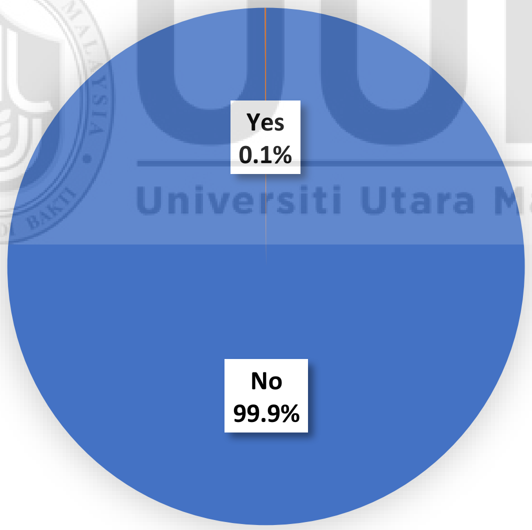


Figure 4.28. Average Product Holding Count for Foreign Currency Fixed Deposit

Table 4.29 and Figure 4.29 summarizes average product holding count for Fixed Deposit. There are more participants who have no product holding count for Fixed Deposit which is 95% compared to 5% of participants who have product holding count for Fixed Deposit.

Table 4.29

Average Product Holding Count for Fixed Deposit

Average Product Holding Count for Fixed Deposit	Total
No	10444
Yes	508

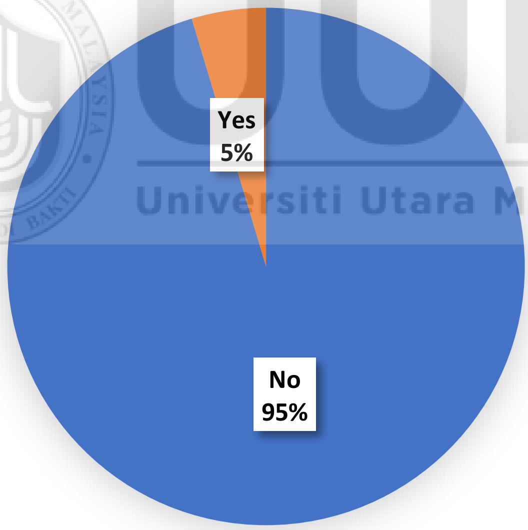


Figure 4.29. Average Product Holding Count for Fixed Deposit

Table 4.30 and Figure 4.30 summarizes average product holding count for Gold Investment Account. There are more participants who have no product holding count for Gold Investment Account which is 99% compared to 1% of participants who have product holding count for Gold Investment Account.

Table 4.30

Average Product Holding Count for Gold Investment Account

Average Product Holding Count for Gold Investment Account	**Total**
No	10860
Yes	92

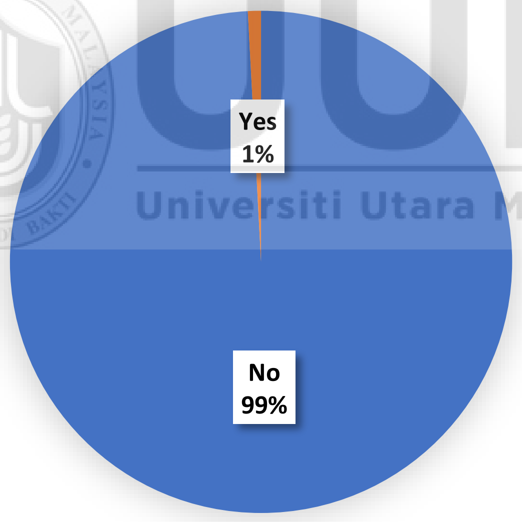


Figure 4.30. Average Product Holding Count for Gold Investment Account

Table 4.31 and Figure 4.31 summarizes average product holding count for Housing Loan. Figure 31 shows the average product holding count for Housing Loan of the participants. There are more participants who have no product holding count for Housing Loan which is 97% compared to 3% of participants who have product holding count for Housing Loan.

Table 4.31

Average Product Holding Count for Housing Loan

Average Product Holding Count for Housing Loan	**Total**
No	10575
Yes	377

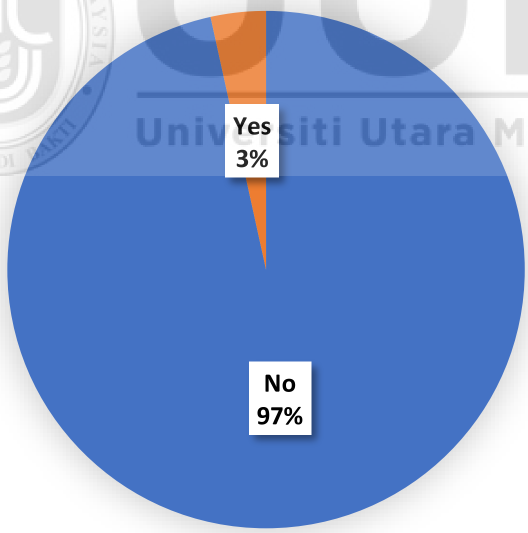


Figure 4.31. Average Product Holding Count for Housing Loan

Table 4.32 and Figure 4.32 summarizes average product holding count for Hire Purchase. There are more participants who have no product holding count for Hire Purchase which is 96% compared to 4% of participants who have product holding count for Hire Purchase.

Table 4.32

Average Product Holding Count for Hire Purchase

Average Product Holding Count for Hire Purchase	**Total**
No	10519
Yes	433

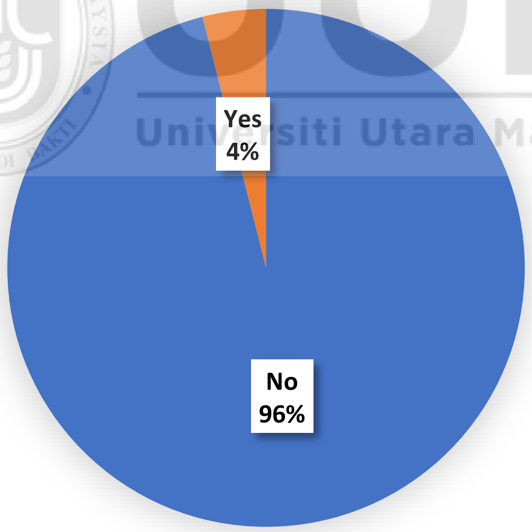


Figure 4.32. Average Product Holding Count for Hire Purchase

Table 4.33 and Figure 4.33 summarizes average product holding count for Insurance (any type). There are more participants who have no product holding count for Insurance (any type) which is 88% compared to 12% of participants who have product holding count for Insurance (any type).

Table 4.33

Average Product Holding Count for Insurance (any type)

Average Product Holding Count for Insurance (any type)	**Total**
No	9681
Yes	1271

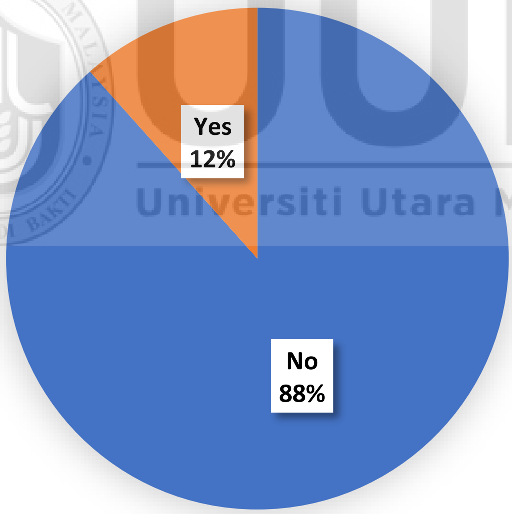


Figure 4.33. Average Product Holding Count for Insurance (any type)

Table 4.34 ad Figure 4.34 summarizes average product holding count for Insurance (life). There are more participants who have no product holding count for Insurance (life) which is 98% compared to 2% of participants who have product holding count for Insurance (life).

Table 4.34

Average Product Holding Count for Insurance (life)

Average Product Holding Count for Insurance (life)	Total
No	10682
Yes	270

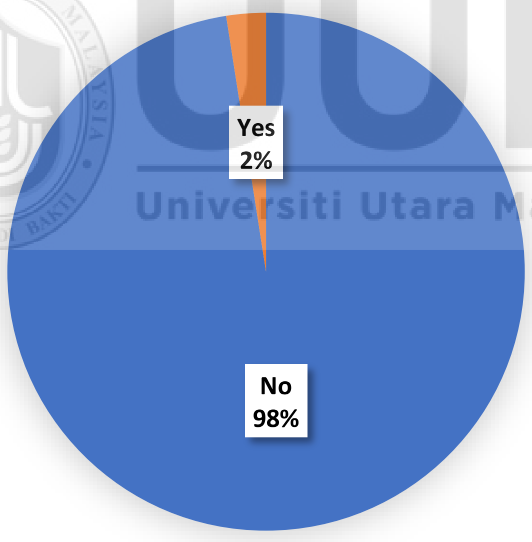


Figure 4.34. Average Product Holding Count for Insurance (life)

Table 4.35 and Figure 4.35 summarizes average product holding count for Insurance (non-life). There are more participants who have no product holding count for Insurance (non-life) which is 97% compared to 3% of participants who have product holding count for Insurance (non-life).

Table 4.35

Average Product Holding Count for Insurance (non-life)

Average product holding count for Insurance (non-life)	Total
No	10629
Yes	323

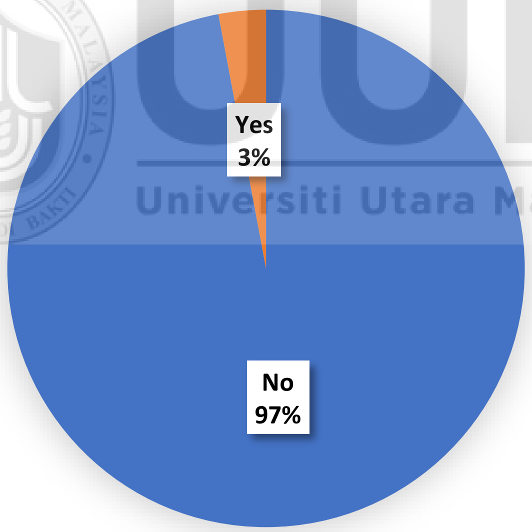


Figure 4.35. Average Product Holding Count for Insurance (non-life)

Table 4.36 and Figure 4.36 summarizes average product holding count for Overdraft. There are more participants who have no product holding count for Overdraft which is 99.6% compared to 0.4% of participants who have product holding count for Overdraft.

Table 4.36

Average Product Holding Count for Overdraft

Average Product Holding Count for Overdraft	Total
No	10910
Yes	42

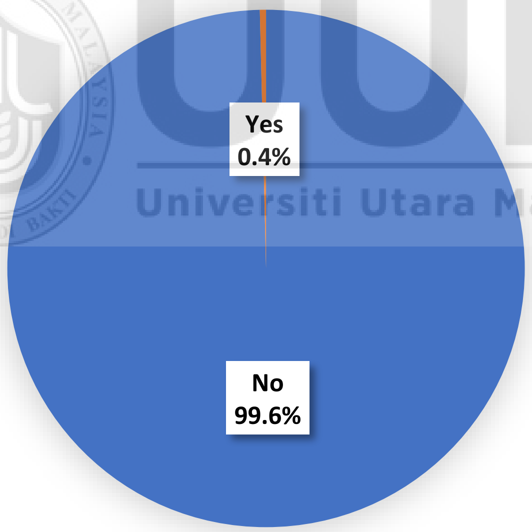


Figure 4.36. Average Product Holding Count for Overdraft

Table 4.37 and Figure 4.37 summarizes average product holding count for Personal Financing. There are more participants who have no product holding count for Personal Financing which is 98% compared to 2% of participants who have product holding count for Personal Financing.

Table 4.37

Average Product Holding Count for Personal Financing

Average Product Holding Count for Personal Financing	Total
No	10724
Yes	228

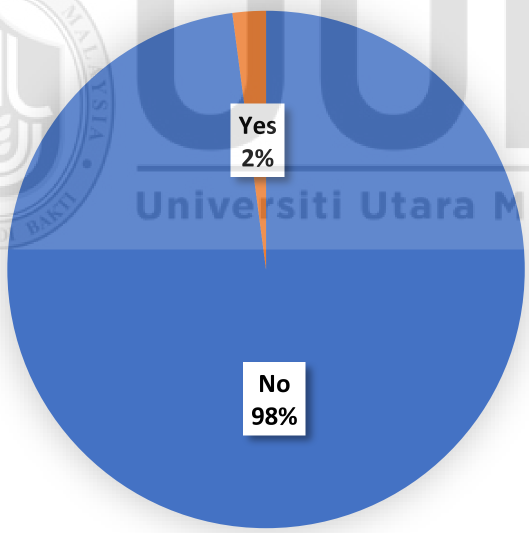


Figure 4.37. Average Product Holding Count for Personal Financing

Table 4.38 and Figure 4.38 summarizes average product holding count for Profit Sharing Specific Investment Account. There are more participants who have no product holding count for Profit Sharing Specific Investment Account which is 99% compared to 1% of participants who have product holding count for Profit Sharing Specific Investment Account.

Table 4.38

Average Product Holding Count for Profit Sharing Specific Investment Account

Average Product Holding Count for PSSIA	Total
No	10868
Yes	84

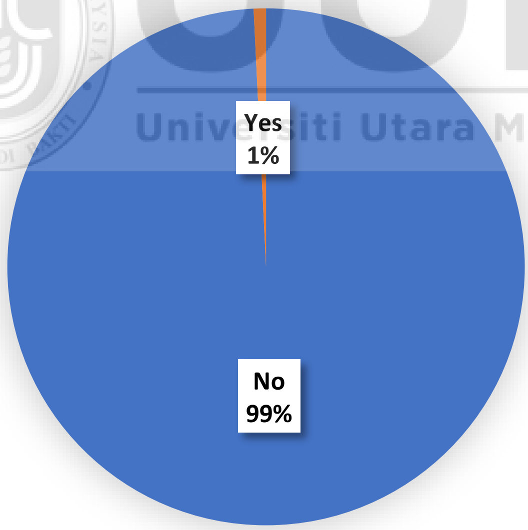


Figure 4.38. Average Product Holding Count for Profit Sharing Specific Investment Account

Table 4.39 and Figure 4.39 summarizes average product holding count for Saving Account. There are more participants who have no product holding count for Saving Account which is 97% compared to 3% of participants who have product holding count for Saving Account.

Table 4.39

Average Product Holding Count for Saving Account

Average Product Holding Count for Saving Account	**Total**
No	314
Yes	10638

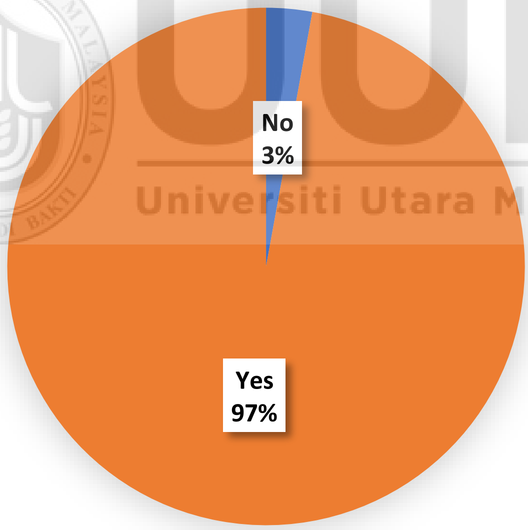


Figure 4.39. Average Product Holding Count for Saving Account

Table 4.40 and Figure 4.40 summarizes average product holding count for Structure Product. There are more participants who have no product holding count for Structure Product which is 99.8% compared to 0.2% of participants who have product holding count for Structure Product.

Table 4.40

Average Product Holding Count for Structure Product

Average Product Holding Count for Structure Product	Total
No	10929
Yes	23

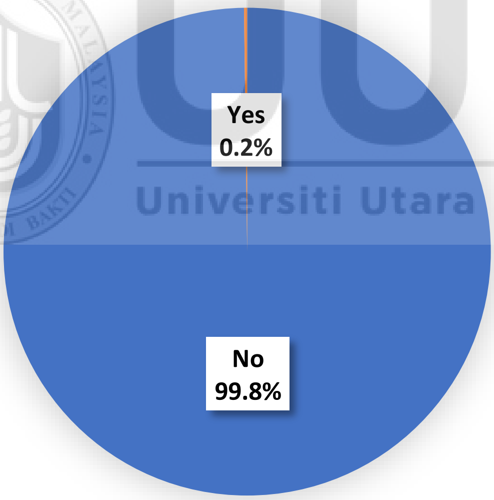


Figure 4.40. Average Product Holding Count for Structure Product

Table 4.41 and Figure 4.41 summarizes average product holding count for Term Loan. There are more participants who have no product holding count for Term Loan which is 99.9% compared to 0.1% of participants who have product holding count for Term Loan.

Table 4.41

Average Product Holding Count for Term Loan

Average Product Holding Count for Term Loan	Total
No	10937
Yes	15

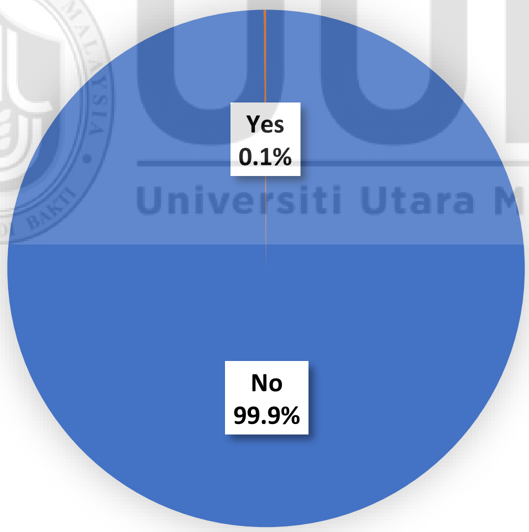


Figure 4.41. Average Product Holding Count for Term Loan

Table 4.42 and Figure 4.42 summarizes average product holding count for Unit Trust. There are more participants who have no product holding count for Unit Trust which is 99% compared to 1% of participants who have product holding count for Unit Trust.

Table 4.42

Average Product Holding Count for Unit Trust

Average Product Holding Count for Unit Trust	Total
No	10875
Yes	77

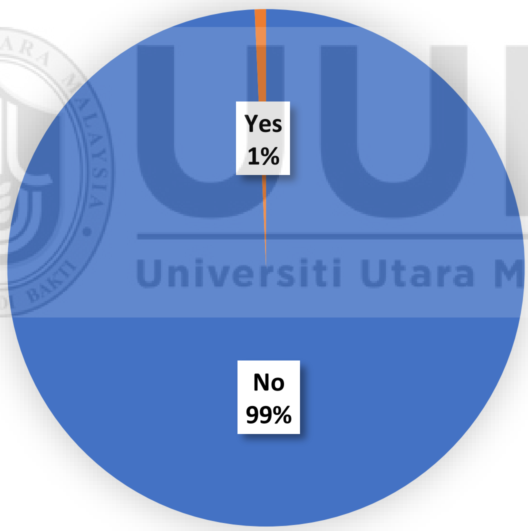


Figure 4.42. Average Product Holding Count for Unit Trust

Table 4.43 and Figure 4.43 summarizes the numbers of payroll customer. There are more customers who have no payroll which is 75% compared to 25% of customers who have payroll.

Table 4.43

Payroll Customer

Payroll Customer	Total
No	8170
Yes	2782

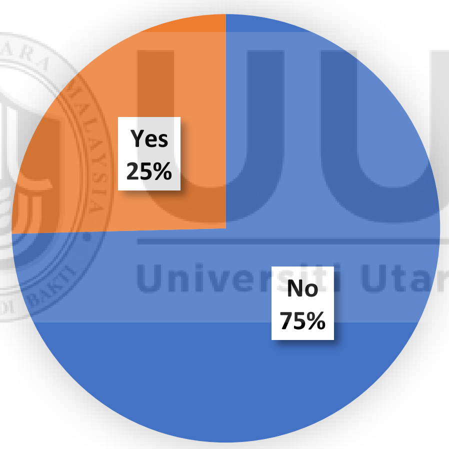


Figure 4.43. Payroll Customer

Table 4.44 and Figure 4.44 summarizes the numbers of Prime customer. Figure 4.44 shows the percentage of prime customer. 100% of the customers are not prime customer.

Table 4.44

Prime Customer

Prime Customer	Total
No	10952
Yes	0

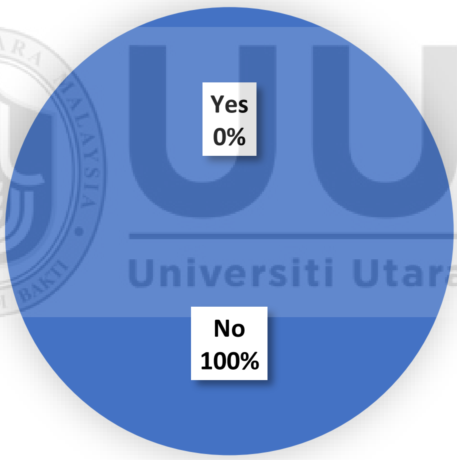


Figure 4.44. Prime Customer

Table 4.45 summarizes the total ATM transaction of participants. Figure 4.45 shows the total ATM transaction of participants. Most bank customers have med number of total ATM transaction which is 24.61%. It was followed by 22.57% of participants who have zero bin number of total ATM transaction.

Table 4.45

Total ATM Transaction

Total ATM Transaction	**Total**
Zero bin	2003
Very low	1955
Low	2695
Med	903
High	924
Very high	2472

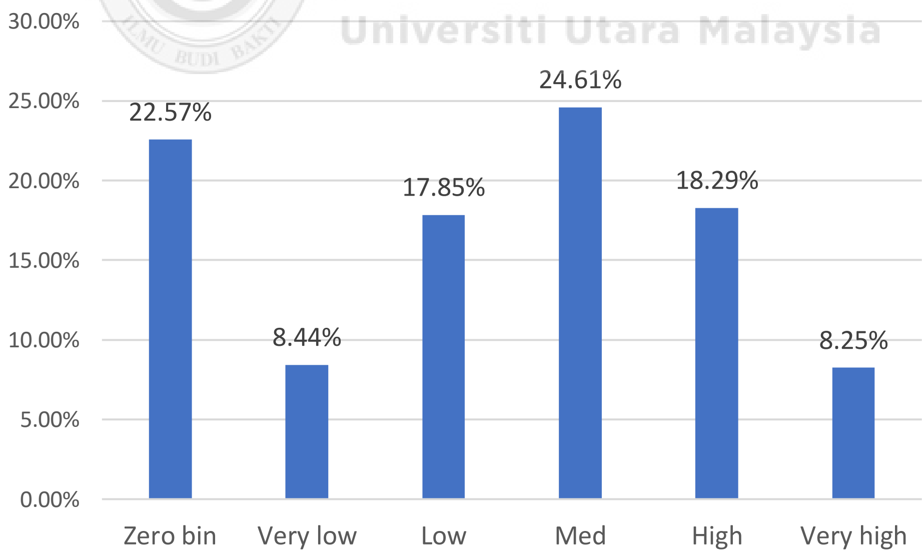


Figure 4.45. Total ATM Transaction

Table 4.46 summarizes the total Branch transaction of participants. Figure 4.46 shows the total Branch transaction of participants. Most bank customers have zero bin number of total branch transaction which is 47.41%.

Table 4.46

Total Branch Transaction

Total Branch Transaction	**Total**
Zero bin	5192
Very low	1233
Low	794
Med	1681
High	1446
Very high	606

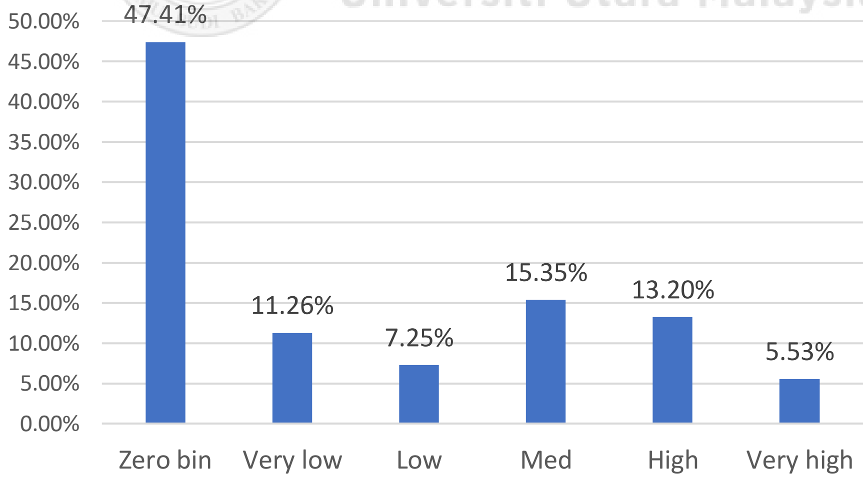


Figure 4.46. Total Branch Transaction

Table 4.47 summarizes the Branch debit transaction last 1 year of participants. Figure 4.47 shows the percentage of Branch debit transaction last 1 year of participants. Most participants have zero bin number of Branch debit transaction last 1 year which is 59.13%.

Table 4.47

Branch Debit Transaction Last 1 Year

Branch Debit Transaction Last 1 Year	Total
Zero bin	6476
Very low	813
Low	663
Med	1447
High	1096
Very high	457

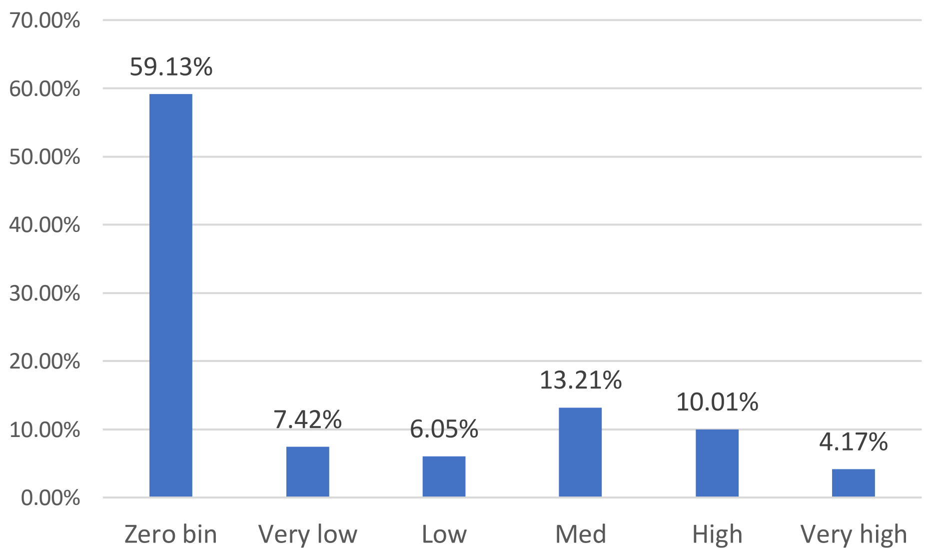


Figure 4.47. Branch Debit Transaction Last 1 Year

Table 4.48 summarizes the Branch credit transaction last 1 year of participants. Figure 4.48 shows the percentage of Branch credit transaction last 1 year of participants. Most participants have zero bin number of Branch credit transaction last 1 year which is 68.50%.

Table 4.48

Branch Credit Transaction Last 1 Year

Branch Credit Transaction Last 1 Year	**Total**
Zero bin	7502
Very low	1482
Med	877
High	762
Very high	329

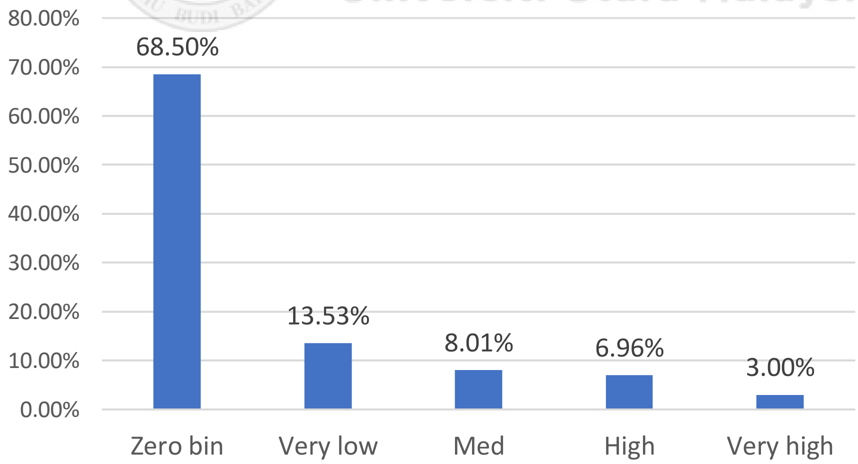


Figure 4.48. Branch Credit Transaction Last 1 Year

Table 4.49 summarizes the total number of online transaction of participants. Figure 4.49 shows the percentage of online transaction of participants. Most participants have zero bin number of online transactions which is 68.42%.

Table 4.49

Online Transaction

Online Transaction	Total
Zero bin	7493
Very low	374
Low	841
Med	1127
High	789
Very high	328

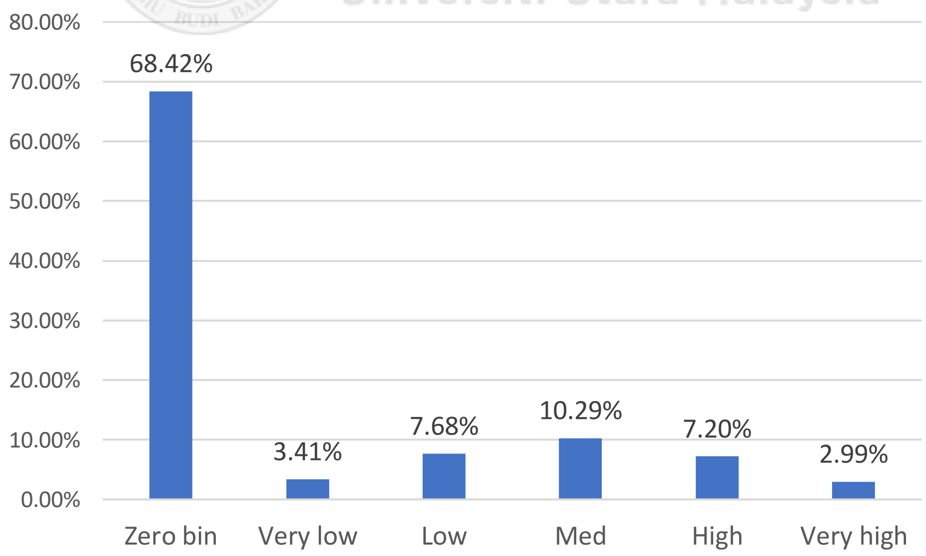


Figure 4.49. Online Transaction

Table 4.50 summarizes the total number of Ratio of Online Transaction vs (Branch + ATM) transaction of participants. Figure 4.50 shows the percentage of Ratio of Online Transaction vs (Branch + ATM) transaction of participants. Most participants have zero bin number of Ratio of Online Transaction vs (Branch + ATM) transaction which is 56.14%.

Table 4.50

Ratio of Online Transaction vs. (Branch + ATM) Transaction

Ratio of Online Transaction vs. (Branch + ATM) Transaction	**Total**
Zero bin	6149
Less than 0	1381
Very low	381
Low	829
Med	1131
High	777
Very high	304

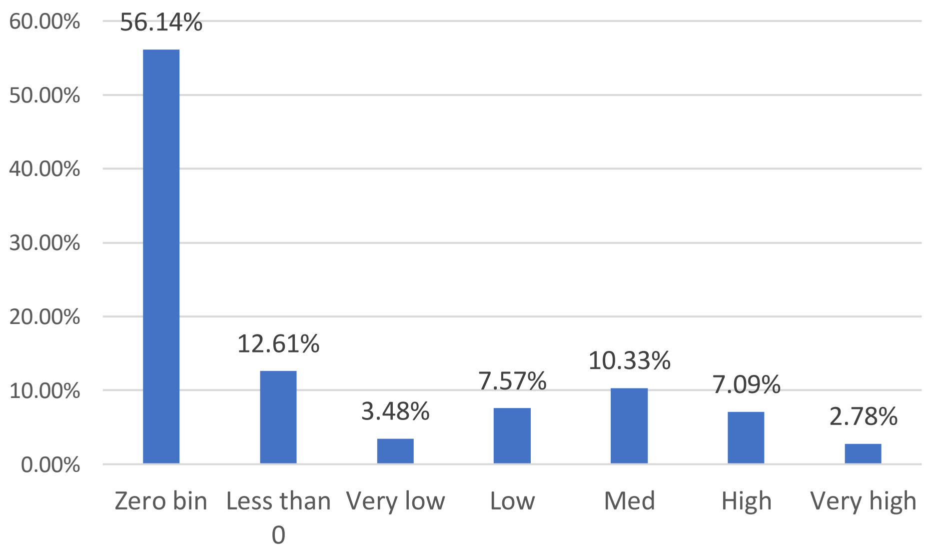


Figure 4.50. Ratio of Online Transaction vs. (Branch + ATM) Transaction

Table 4.51 summarizes the total number of Customer MOB with bank. Figure 4.51 shows the percentage of Customer MOB with bank. Most customers have med number of MOB with bank which is 33.78%.

Table 4.51

Customer MOB with Bank

Customer MOB with Bank	**Total**
Very low	1158
Low	2504
Med	3700
High	2520
Very high	1070

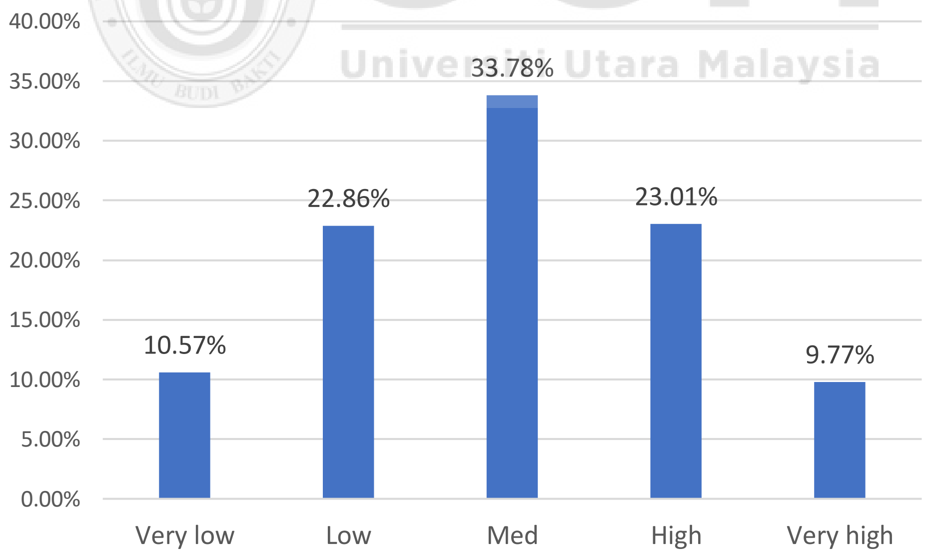


Figure 4.51. Customer MOB with Bank

Table 4.52 summarizes the total number of average amount ATM transaction debit last 1 year of participants. Figure 4.52 shows the percentage of average amount ATM transaction debit last 1 year of participants. Most participants have med number of average amount ATM transaction debit last 1 year which is 25.43%. It was followed by 22.80% of the participants who have zero bin number of average amount ATM transaction debit last 1 year.

Table 4.52

Average Amount ATM Transaction Debit Last 1 Year

Average Amount ATM Transaction Debit Last 1 Year	**Total**
Zero bin	2497
Very low	831
Low	1839
Med	2785
High	2124
Very high	876

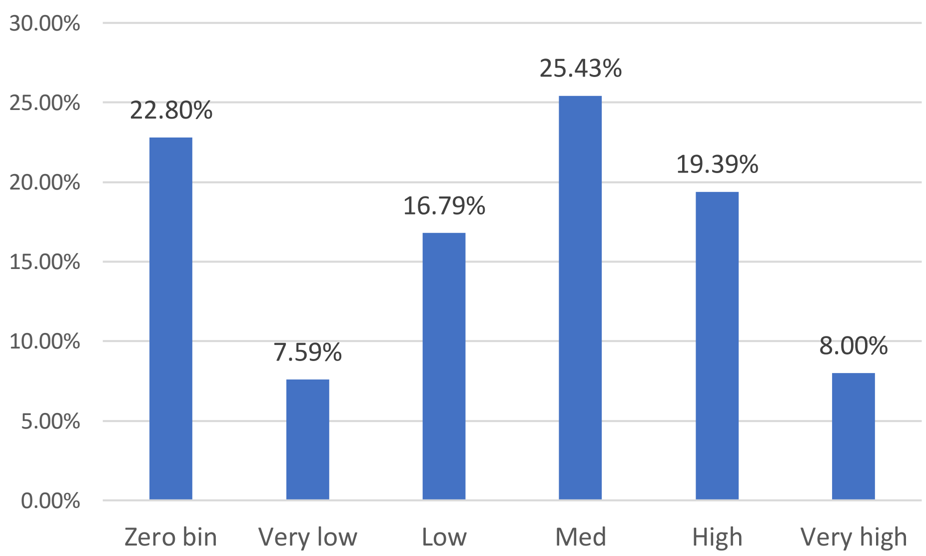


Figure 4.52. Average Amount ATM Transaction Debit Last 1 Year

Table 4.53 summarizes the total number of average amount Online transaction debit last 1 year of participants. Figure 4.53 shows the percentage of average amount Online transaction debit last 1 year of participants. Most participants have zero bin number of average amount Online transaction debit last 1 year which is 68.44%.

Table 4.53

Average Amount Online Transaction Debit Last 1 Year

Average Amount Online Transaction Debit Last 1 Year	**Total**
Zero bin	7496
Very low	339
Low	796
Med	1103
High	829
Very high	389

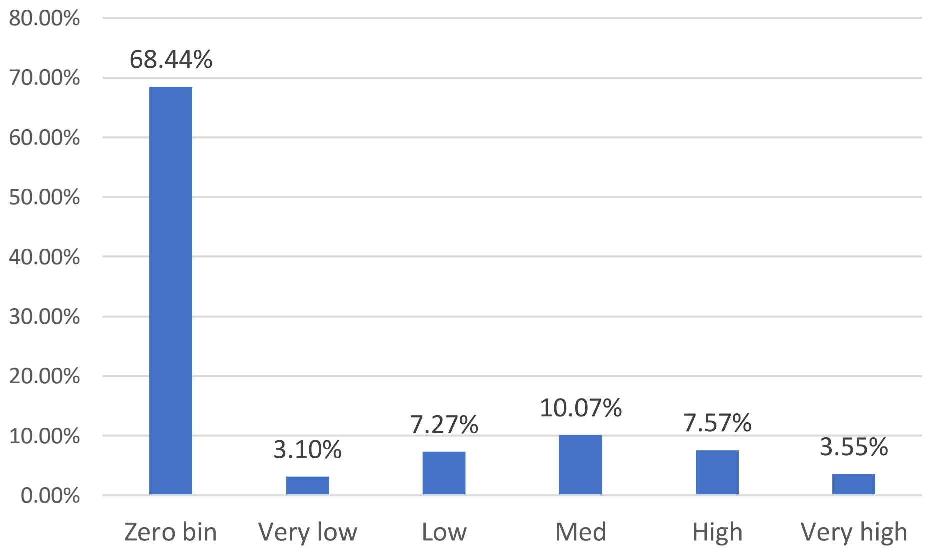


Figure 4.53. Average Amount Online Transaction Debit Last 1 Year

Table 4.54 summarizes the number of participants that consist of different age group. Figure 4.54 shows the percentage of age groups of the participant in the survey. Most participants come from the med number of age which is 31.21%.

Table 4.54

Customer's Age

Customer's Age	**Total**
Very low	1149
Low	2737
Med	3418
High	2647
Very high	1001

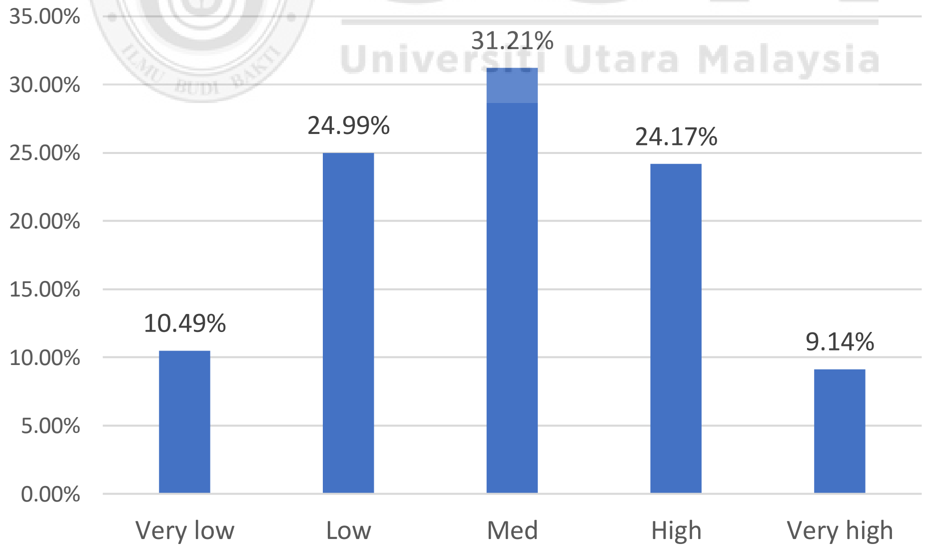


Figure 4.54. Customer's Age

Table 4.55 summarizes the total number of transaction (any type) of participants. Figure 4.55 shows the percentage of number of transaction (any type) of participants. Most participants have med number of total numbers of transaction (any type) which is 32.88%. It was followed by 24.29% of participants who have high number of total numbers of transaction (any type).

Table 4.55

Total Number of Transaction (Any Type)

Total Number of Transaction (Any Type)	**Total**
Zero bin	10
Very low	1126
Low	2374
Med	3601
High	2660
Very high	1181

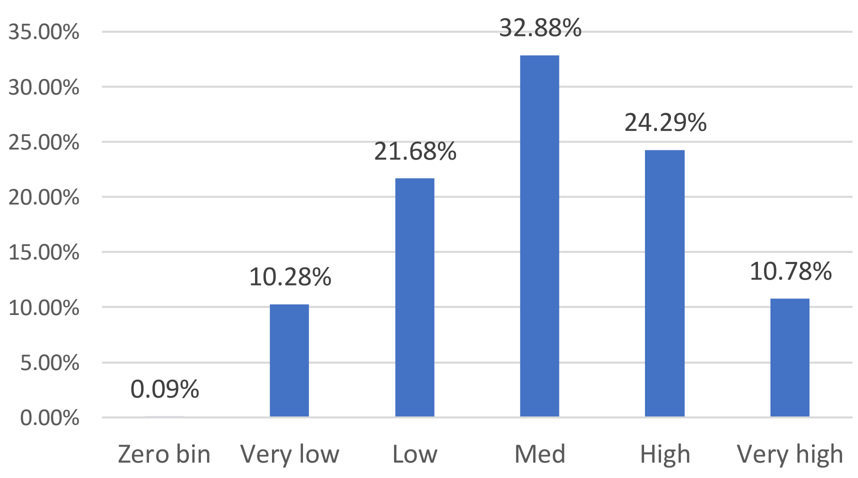


Figure 4.55. Total Number of Transaction (Any Type)

Table 4.56 summarizes the total number of Ratio of Branch Transaction vs (Online + ATM) transaction of participants. Figure 4.56 shows the percentage of Ratio of Branch Transaction vs (Online + ATM) transaction of participants. Most participants have zero bin number of Ratio of Branch Transaction vs (Online + ATM) transaction which is 35.14%.

Table 4.56

Ratio of Branch Transaction vs. (Online + ATM) Transaction

Ratio of Branch Transaction vs. (Online + ATM) Transaction	**Total**
Zero bin	3848
Less than 0	2227
Very low	492
Low	1070
Med	1595
High	1209
Very high	511

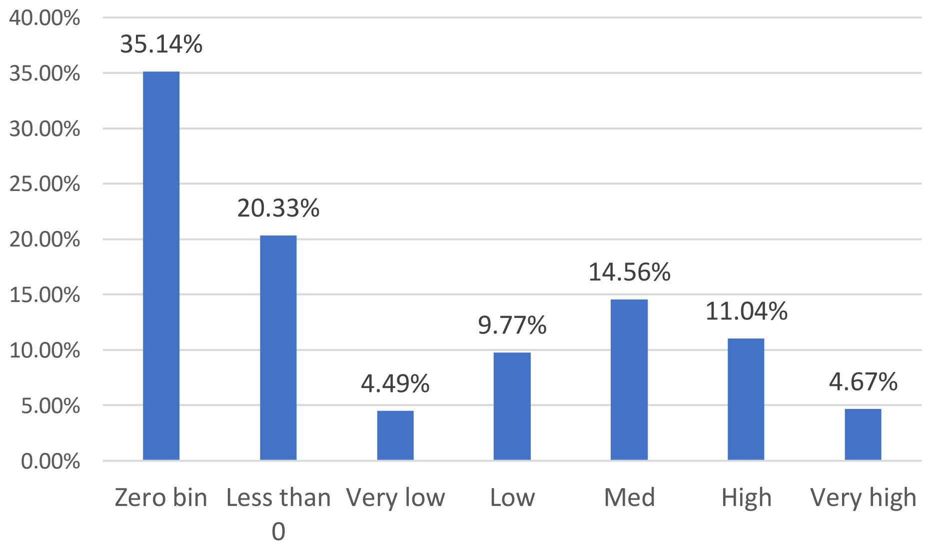


Figure 4.56. Ratio of Branch Transaction vs. (Online + ATM) Transaction

Table 4.57 summarizes the total number of Ratio of Inflow vs. Outflow of participants. Figure 4.57 shows the percentage of Ratio of Inflow vs Outflow of participants. Most participants have med number of Ratio of Inflow vs Outflow which is 30.84%. It was followed by 22.87% of the participants who have high number of Ratio of Inflow vs Outflow.

Table 4.57

Ratio of Inflow vs. Outflow

Ratio of Inflow vs. Outflow	Total
Zero bin	470
Less than 0	181
Very low	973
Low	2345
Med	3378
High	2505
Very high	1100

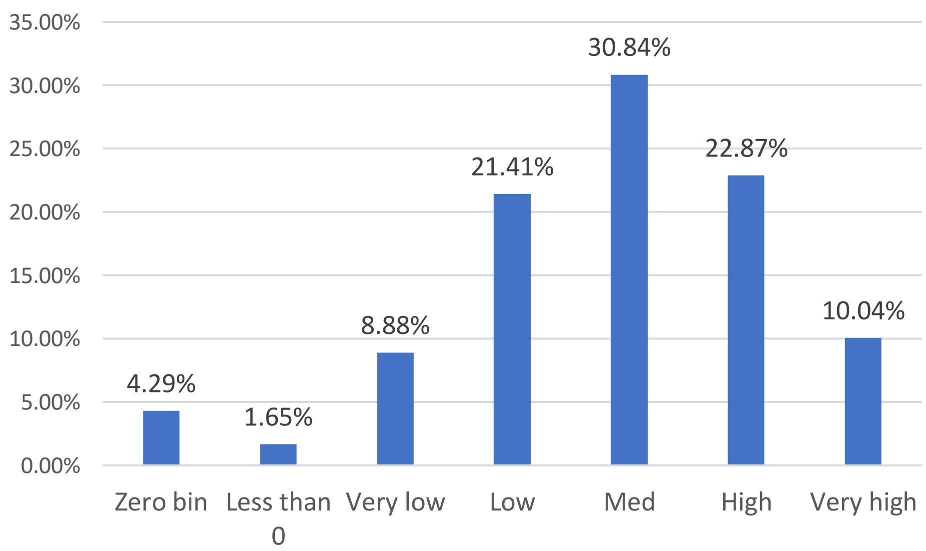


Figure 4.57. Ratio of Inflow vs. Outflow

4.3 Variables Selection

4.3.1 Factor Analysis

In order to identify a set of variables that influence the customer response for personal loan product, factor analysis is used to reduce the dimensionality of big data sets by reducing a large collection of variables into a smaller one that retains the majority of the information from the larger set. Since this study has more than 35 variables (including one dependent variable), factor analysis helps to discard those variables known to be redundant.

4.3.1.1 KMO and Barlett's Test

From the Table 4.58, it can be seen that KMO value is clearly stated to be 0.707 which is greater than 0.6. Therefore, KMO value is well within the acceptable limits and thus, implying that PCA is appropriate for this data.

Besides, as can be seen also, *p*-value for Barlett's Test of Sphericity is 0.000, less than $\alpha = 0.05$. Thus, the study rejected the null hypothesis and concluded that the variables were not identify / not independent and unrelated to each other. Hence, the data is suitable for reduction.

Table 4.58

KMO and Barlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.707
Bartlett's Test of Sphericity	Approx. Chi-Square	303802.201
	df	1540
	Sig.	0.000

4.3.1.2 Anti-Image Correlation

It is feasible to determine whether particular variables should be excluded from the analysis using the anti-image matrices. From the partial correlations, pairs of variables with large partial correlations share variances with one another but not with the remaining variables; this is problematic. Kaiser's Measures of Sampling Adequacy (MSA) will tell, for each variable, how much of this problem exists. The smaller the MSA, the greater the problem. Based on the rule of thumb, an MSA more than 0.9 is marvelous, 0.8 meritorious, 0.7 middling, 0.6 mediocre and 0.5 miserable (Glen, 2021). Variables with small MSA's should be deleted. From the Table 4.59, it shows the list of variables remained after other variables that is less than 0.6 were rejected.

Table 4.59

Anti-Image Correlation

Variable Name	**Value**
Total balance for Saving Account	0.840
Education	0.854
Occupation	0.819
Total number of Principal Credit Card Account	0.724
Total number of Active Clicks	0.961
Total number of Debit Card	0.833
Total number for Fixed Deposit	0.630
Total Product Holding	0.693
Average product holding count for Business Premises Loan	0.804
Average product holding count for Balance Transfer	0.697
Average product holding count for Current Account	0.651

Average product holding count for Cash Lite	0.918
Average product holding count for Foreign Currency Current Account	0.629
Average product holding count for Foreign Currency Fixed Deposit	0.607
Average product holding count for Fixed Deposit	0.625
Average product holding count for Gold Investment Account	0.793
Average product holding count for Housing Loan	0.703
Average product holding count for Overdraft	0.845
Average product holding count for Saving Account	0.725
Average product holding count for Term Loan	0.687
Prime Customer	0.604
Average product holding count for Structure Product	0.727
Average product holding count for Unit Trust	0.825
Total ATM transaction	0.742
Total Branch transaction	0.742
Branch Debit transaction last 1 Year	0.817
Online transaction	0.813
Customer MOB with Bank	0.797
Average amount ATM transaction debit last 1 Year	0.778
Average amount Online transaction debit last 1 Year	0.881
Branch Credit transaction last 1 Year	0.715
Customer's age	0.733
Number of Dependants	0.765
Marital Status	0.796

4.3.1.3 Communalities

The communalities describe the variance portions of the individual variables that can be explained by the extracted factors. The concept of the extraction is the same as the concept of R^2, which means how much the variability explains the model. If the value is low, the variable is not good to explain the analysis. Thus, any variable with a low extraction value will be deleted. Based on variables remaining in Table 4.59, the extraction values are shown below. From Table 4.60, the highlighted values have an extraction value less than 0.5. It is considered low and should be deleted from the model. The other 29 variables remain since it has a value of more than 0.5.

Table 4.60

Communalities

Variable Name	Value
Total balance for Saving Account	0.617
Education	0.582
Occupation	0.532
Total number of Principal Credit Card Account	0.847
Total number of Active Clicks	0.687
Total number of Debit Card	0.533
Total number for Fixed Deposit	0.953
Total Product Holding	0.625
Average product holding count for Business Premises Loan	0.224
Average product holding count for Balance Transfer	0.52
Average product holding count for Current Account	0.847
Average product holding count for Cash Lite	0.431

Average product holding count for Foreign Currency Current Account	0.66
Average product holding count for Foreign Currency Fixed Deposit	0.683
Average product holding count for Fixed Deposit	0.95
Average product holding count for Gold Investment Account	0.302
Average product holding count for Housing Loan	0.632
Average product holding count for Overdraft	0.591
Average product holding count for Saving Account	0.857
Average product holding count for Term Loan	0.519
Prime Customer	0.525
Average product holding count for Structure Product	0.493
Average product holding count for Unit Trust	0.469
Total ATM transaction	0.896
Total Branch transaction	0.9
Branch Debit transaction last 1 Year	0.808
Online transaction	0.88
Customer MOB with Bank	0.504
Average amount ATM transaction debit last 1 Year	0.862
Average amount Online transaction debit last 1 Year	0.858
Branch Credit transaction last 1 Year	0.679
Customer's age	0.739
Number of Dependant	0.565
Marital Status	0.594

4.3.1.4 Total Variance Explained

There are no generally applicable rules for selecting the optimal number of components to extract. According to the Kaiser criterion, factors are considered relevant if their eigenvalue (the proportion of the total variance of all variables that can be explained by the factor) is more than one, explaining more of the variation than a single tested variable included in the principal component analysis.

From Table 4.61, the column labelled "Total" shows the eigenvalues of the factors. In this data, there are 20 factors whose eigenvalues are greater than one. Based on the Kaiser criterion, 20 factors need to be extracted. These 20 factors together explain 66.5% of the total.

Table 4.61

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	7.125	12.723	12.723	7.125	12.723	12.723
2	3.842	6.86	19.583	3.842	6.86	19.583
3	2.77	4.946	24.529	2.77	4.946	24.529
4	2.278	4.067	28.596	2.278	4.067	28.596
5	2.155	3.848	32.444	2.155	3.848	32.444
6	2.06	3.679	36.123	2.06	3.679	36.123
7	1.73	3.089	39.211	1.73	3.089	39.211
8	1.511	2.698	41.91	1.511	2.698	41.91

9	1.431	2.555	44.465	1.431	2.555	44.465
10	1.326	2.368	46.833	1.326	2.368	46.833
11	1.225	2.188	49.021	1.225	2.188	49.021
12	1.193	2.13	51.151	1.193	2.13	51.151
13	1.177	2.102	53.253	1.177	2.102	53.253
14	1.131	2.02	55.274	1.131	2.02	55.274
15	1.106	1.975	57.248	1.106	1.975	57.248
16	1.071	1.912	59.161	1.071	1.912	59.161
17	1.052	1.878	61.039	1.052	1.878	61.039
18	1.034	1.847	62.885	1.034	1.847	62.885
19	1.017	1.815	64.701	1.017	1.815	64.701
20	1.004	1.792	66.493	1.004	1.792	66.493
21	0.979	1.747	68.24			
22	0.969	1.731	69.971			
23	0.958	1.711	71.683			
24	0.933	1.666	73.349			
25	0.927	1.655	75.004			
26	0.905	1.615	76.619			
27	0.887	1.584	78.204			
28	0.864	1.544	79.747			
29	0.837	1.494	81.241			
30	0.827	1.476	82.718			
31	0.812	1.449	84.167			
32	0.783	1.399	85.566			

33	0.762	1.361	86.927			
34	0.746	1.332	88.259			
35	0.71	1.267	89.526			
36	0.578	1.033	90.559			
37	0.555	0.991	91.55			
38	0.532	0.95	92.5			
39	0.517	0.924	93.424			
40	0.475	0.848	94.272			
41	0.454	0.81	95.082			
42	0.42	0.75	95.833			
43	0.368	0.658	96.49			
44	0.361	0.645	97.135			
45	0.337	0.602	97.737			
46	0.309	0.552	98.289			
47	0.216	0.386	98.674			
48	0.208	0.372	99.046			
49	0.138	0.246	99.292			
50	0.089	0.159	99.452			
51	0.08	0.143	99.595			
52	0.052	0.093	99.688			
53	0.05	0.089	99.776			
54	0.047	0.083	99.86			
55	0.043	0.076	99.936			
56	0.036	0.064	100			

4.3.1.5 Rotated Component Matrix

The rotated component matrix, often known as the loadings, is the primary output of principal components analysis. It includes estimates of the correlations between each variable and the estimated components. This data applies the most often used Varimax approach, which results in a reduction of variables with large multiple loadings.

The variables were grouped into components according to their value. Since in the previous section, it found out that 20 components need to be extracted. The variables were being grouped into 20 components according to their correlation value. Table 4.62 shows the items loading on the 20 components with 29 variables. From the table, "online activities" would work as the general term for the first factor because it has high loadings with Total number of active clicks, Total Branch transaction, Branch debit transaction last 1year, Online Transaction and Average amount Online transaction debit last 1 year. Whereas for second factor, "banking machine activities" would work as the general term. Besides, fourth factor would be "demographic" and sixth factor would be "fixed deposit".

Table 4.62

Summarization of Rotated Component Matrix

Component	Variables	Correlation
1	Total Number of Active Clicks	0.768
	Total Branch Transaction	0.85
	Branch Debit Transaction Last 1 Year	0.857
	Online Transaction	0.892
	Average Amount Online Transaction Debit Last 1 Year	0.876

2	Total ATM Transaction	0.879
	Average Amount ATM Transaction Debit Last 1 Year	0.871
3	Total Number of Principal Credit Card Account	0.867
4	Occupation	0.676
	Customer's Age	0.845
	Marital Status	0.681
5	Average Product Holding Count for Current Account	0.84
6	Total Number for Fixed Deposit	0.954
	Average Product Holding Count for Fixed Deposit	0.956
7	Average Product Holding Count for Housing Loan	0.766
8	Education	0.686
	Number of Dependant	0.661
9	Average Product Holding Count for Saving Account	0.161
10	Total Balance for Saving Account	0.515
11	Average Product Holding Count for Foreign Currency Current Account	0.445
12	Total Number of Debit Card	0.471
13	Customer MOB with Bank	0.151
14	Average Product Holding Count for Term Loan	0.664
15	Branch Credit Transaction Last 1 Year	0.418
16	Average Product Holding Count for Foreign Currency Fixed Deposit	0.815
17	Total Product Holding	0.418
18	Average Product Holding Count for Balance Transfer	0.696
19	Prime Customer	0.711
20	Average Product Holding Count for Overdraft	0.638

4.4 Logistic Regression

Logistic Regression is used to develop customer response model to predict whether a customer is responding to personal loan product. Table 4.63 shows the list of dependent variables (categorical) and the other predicted variables.

Table 4.63

List of Variables

Variables	
Average Product Holding Count for Personal Financing	Y
Total Number of Active Clicks	X1
Total Branch Transaction	X2
Branch Debit Transaction Last 1 Year	X3
Online Transaction	X4
Average Amount Online Transaction Debit Last 1 Year	X5
Total ATM Transaction	X6
Average Amount ATM Transaction Debit Last 1 Year	X7
Total Number of Principal Credit Card Account	X8
Occupation	X9
Customer's Age	X10
Marital Status	X11
Average Product Holding Count for Current Account	X12
Total Number for Fixed Deposit	X13
Average Product Holding Count for Fixed Deposit	X14
Average Product Holding Count for Housing Loan	X15

Education	X16
Number of Dependents	X17
Average Product Holding Count for Saving Account	X18
Total Balance for Saving Account	X19
Average Product Holding Count for Foreign Currency Current Account	X20
Total Number of Debit Card	X21
Customer MOB with Bank	X22
Average Product Holding Count for Term Loan	X23
Branch Credit Transaction Last 1 Year	X24
Average Product Holding Count for Foreign Currency Fixed Deposit	X25
Total Product Holding	X26
Average Product Holding Count for Balance Transfer	X27
Prime Customer	X28
Average Product Holding Count for Overdraft	X29

4.4.1 Hosmer and Lemeshow Test

The Hosmer-Lemeshow test is a statistical test evaluating the logistic regression model's goodness of fit. A high Chi-squared value (with a small p-value of < 0.05) suggests poor fit, whereas a low Chi-squared value (with a larger p-value closer to 1) indicates good logistic regression model fit. Based on the Table 4.64, the p-value is 0.965 and thus, suggest that the model is a good fit to the data.

Table 4.64

Hosmer and Lemeshow Test

Chi-square	df	Sig.
2.417	8	0.965

4.4.2 Cross Validation Technique

There are several cross-validation techniques in determining how well a statistical analysis generalizes to a different set of data. The study utilized k-fold cross validation, which is considered to be one of the most effective techniques for assessing the performance of a machine learning model through cross validation. Precision is a measure of how accurate our model is when the prediction is correct. Whereas, Recall indicates how well our model correctly predicts positive classes. Table 4.65 shows the value of precision and recall for 5 and 10-fold. As can be seen it can be concluded that when it predicts that an individual has response to personal loan product, it is correct around 98.5% and 98.6% of the time respectively. While, for all the individuals who actually have personal finance, 99.5% and 99.6% are correctly identified as having personal finance respectively. Table 4.65 also shows the accuracy of the resampling cross validated on 5 and 10-fold. The model accuracy of the outcome for both folds is 98.1%.

Table 4.65

Summary Model

	Accuracy	Precision	Recall
5-fold	98.1	98.5	99.5
10-fold	98.1	98.6	99.6

4.4.3 Parameter Estimation

The logit is the natural logarithm (ln) of odds of Y. The estimated equation of simple logistic model has the form:

Log odds: $\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X$

(4.1)

Odds: $p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

(4.2)

Based on the model, after assessing the significant variables (p-value < 0.05), the estimated equation of the logistic model:

$\ln(\frac{p}{1-p})$ =-24.25 + 1.131Total Balance for Saving Account(low) + 1.075Education (not stated) + 1.288Occupation(non_exec) - 1.287Total number of Principal Credit Card account(1) -0.822Total number of Principal Credit Card account(GT1) - 1.994Total number of Debit Card - 3.549Total number for Fixed Deposit(1) - 2.035Average Product Holding Count for Current Account(Yes) - 2.773Average Product Holding Count for Housing Loan(Yes) + 0.481Total ATM transaction(med) - 1.680Branch debit transaction last 1 year(low) - 1.233Branch debit transaction last 1 year(med) - 3.033Branch debit transaction last 1 year(zero bin) + 0.695Customer MOB with bank(low) + 0.879Customer MOB with bank(med) - 1.768Customer MOB with bank(very high) + 1.331Average amount Online transaction debit last 1 year(very low) + 0.852Branch credit transaction last 1 year(very low) - 1.259Customer's age(low) - 1.313Customer's age(very low) - 1.230Number of dependent(3 dpndnt)

Table 4.66

Parameter Estimation

	Estimate	**S.E.**	**z value**	**Pr(>|z|)**
SA_BAL_Totallow	1.131	0.266	4.246	0.000
D_EDUnot stated	1.075	0.249	4.313	0.000
D_OCCnon_exec	1.288	0.612	2.104	0.035
CC_PRIN_CNT_Rec1	-1.287	0.273	-4.720	0.000
CC_PRIN_CNT_RecGT1	-0.822	0.358	-2.295	0.022
DC_CNT_Rec	-1.994	0.255	-7.815	0.000
FD_CNT_Rec1	-3.549	1.068	-3.323	0.001
AVPH_CA_CNTYes	-2.035	0.544	-3.740	0.000
AVPH_HL_CNTYes	-2.773	0.365	-7.587	0.000
ATM_txnmed	0.481	0..245	1.964	0.050
Branch_txn_Db_L1Ylow	-1.680	0.698	-2.409	0.016
Branch_txn_Db_L1Ymed	-1.233	0.470	-2.627	0.009
Branch_txn_Db_L1Yzero bin	-3.033	0.938	-3.234	0.001
Tenurelow	0.695	0.296	2.347	0.019
Tenuremed	0.879	0.232	3.791	0.000
Tenurevery high	-1.768	0.557	-3.177	0.001
A_Amt_Online_txn_Db_L1Yvery low	1.331	0.622	2.138	0.032
Branch_txn_Cr_L1Yvery low	0.852	0.382	2.229	0.026
CD_AGElow	-1.259	0.341	-3.693	0.000
CD_AGEvery low	-1.313	0.576	-2.279	0.023
Depen_CNT_rec3 dpndnt	-1.230	0.531	-2.315	0.021

4.4.4 Interpretation of Coefficients

The exponential function of the regression coefficient (e^{β}) is the odds ratio associated with a one unit increase in the exposure. Table 1.19 shows the odds ratio and interpretation of the contribution of individual inputs in a model.

Table 4.67

Interpretation of Coefficients

X	e^{β}	(e^{β})-1	**% change in odds**	**Interpretation**
SA_BAL_Totallow	3.099	2.099	209.9	The odds of having personal finance for an individual who has low total balance for saving account is 3.099 times higher than those who have high total.
D_EDUnot stated	2.930	1.930	193.0	The odds of having personal finance for an individual who not stated their education is 2.930 times higher than those who grad.
D_OCCnon_exec	3.626	2.626	262.6	The odds of having personal finance for an individual who has non_exec as their occupation is 3.626 times higher than those who has a business.

CC_PRIN_CNT_Rec1	0.276	-0.724	72.4	The odds of having personal finance for an individual who has one total number of principal credit card accounts is 0.276 times higher than those who have none.
CC_PRIN_CNT_RecGT1	0.440	-0.560	56.0	The odds of having personal finance for an individual who has GT one total number of principal credit card accounts is 0.440 times higher than those who have none.
DC_CNT_Rec	0.136	-0.864	86.4	1-unit change in total number of debit card will decrease the odds of having personal finance by 86.4%.
FD_CNT_Rec1	0.029	-0.971	97.1	An individual who has one total number for fixed deposit vs individuals who have none decrease the odds of having personal finance by 97.1%.
AVPH_CA_CNTYes	0.131	-0.869	86.9	The odds of having personal finance for an individual who have product holding count for

				current account is 0.131 times higher than those who have none.
AVPH_HL_CNTYes	0.062	-0.938	93.8	The odds of having personal finance for an individual who have product holding count for housing loan is 0.062 times higher than those who have none.
ATM_txnmed	1.618	0.618	61.8	The odds of having personal finance for an individual who have med total ATM transaction is 1.618 times higher than those who have high total ATM transaction.
Branch_txn_Db_L1Ylow	0.186	-0.814	81.4	The odds of having personal finance for an individual who have low branch debit transaction last 1 year is 0.186 times higher than those who have high transaction.
Branch_txn_Db_L1Ymed	0.291	-0.709	70.9	The odds of having personal finance for an individual who have med branch debit transaction last 1 year is 0.291 times higher than those who have high transaction.

Branch_txn_Db_L1 Yzero bin	0.048	-0.952	95.1	The odds of having personal finance for an individual who have zero bin branch debit transaction last 1 year is 0.048 times higher than those who have high transaction.
Tenurelow	2.004	1.004	100.4	The odds of having personal finance for an individual who have low MOB with bank is 2.004 times higher than those who have high MOB.
Tenuremed	2.408	1.408	140.8	The odds of having personal finance for an individual who have med MOB with bank is 2.408 times higher than those who have high MOB.
Tenurevery high	0.171	-0.829	82.9	The odds of having personal finance for an individual who have very high MOB with bank is 0.171 times higher than those who have high MOB.
A_Amt_Online_txn _Db_ L1Yvery low	3.785	2.785	278.5	The odds of having personal finance for an individual who have very low average amount online transaction debit last 1

				year is 3.785 times higher than those who have high transaction.
Branch_txn_Cr_ L1Yvery low	2.344	1.344	134.4	The odds of having personal finance for an individual who have very low branch credit transaction last 1 year is 2.344 times higher than those who have high transaction.
CD_AGElow	0.284	-0.716	71.6	The odds of having personal finance for customers in low age is 0.284 times higher than those who have a high age.
CD_AGEvery low	0.269	-0.731	73.1	The odds of having personal finance for customers in very low age is 0.269 times higher than those who have a high age.
Depen_CNT_rec3 dpndnt	0.292	-0.708	70.8	The odds of having personal finance for customers who have three dependents is 0.292 times higher than those who have zero dependents.

4.5 Decision Tree

Decision tree is used to develop customer response model to predict whether a customer is responding to personal loan product. J48 algorithm is used for the prediction task. WEKA software is used to develop J48 for customer response prediction model. Decision tree is a non-parametric method that have good application in business domain. The list of dependent variables (categorical) and the other predicted variables are same as Table 4.63.

4.5.1 J48 Results

The work utilizes the J48 algorithm, which is a decision tree algorithm that uses expected class values to create decision nodes. It can handle missing or incomplete attribute values, as well as variable attribute costs. Table 4.68 displays the confusion matrix for the J48 algorithm in this work, indicating that the model accurately predicts outcomes in 98.2% of cases. Therefore, the model demonstrates strong predictive capabilities.

Table 4.68

Confusion Matrix for J48

Observed		Predicted		
		Average Product Holding Count for Personal Loan		**Percentage Correct**
		No	**Yes**	
Average Product Holding Count for Personal Loan	**No**	10684	40	99.6
	Yes	158	70	30.7
Overall Percentage				98.2

Table 4.69 shows the value of precision and recall of J48 algorithm in decision tree. As can be seen it can be concluded that when it predicts that an individual has a personal loan product, it is correct around 63.64% of the time. While, for all the individuals who actually have personal loan product, 30.7% are correctly identified as having personal loan product.

Table 4.69

Precision and Recall for J48

Measurement	Value
Precision	$\frac{\text{True positive}}{\text{Total predicted positive}}$ = 70 / 70+ 40 = 0.636 ≈ 63.6%
Recall	$\frac{\text{True positive}}{\text{Total actual positive}}$ = 70 / 70+158 = 0.307 ≈ 30.7%

Figure 4.58 shows the tree that developed by J48 algorithm. Table 4.70 summarizes the characteristics of the tree developed by J48. The size of the tree has 95 instances and there are 69 leaves in the decision model. The data was modelled using WEKA software as WEKA provide more decision tree options (Olson, Delen & Meng, 2012).

Although different models including logistic regression and J48 are conducted on the same dataset, both models are performing quite well in term of accuracies. However, logistic regression slightly better than J48. Logistic regression has the highest accuracy (98.4%) and precision (36.4%). Table 4.71 shows the rules generated by J48 algorithm. The J48 model generates 19 variables with 54 rows of rules.

Table 4.70

Characteristic of Tree

Parameter	Numbers of Instances
Number of Leaves	69
Size of the tree	95

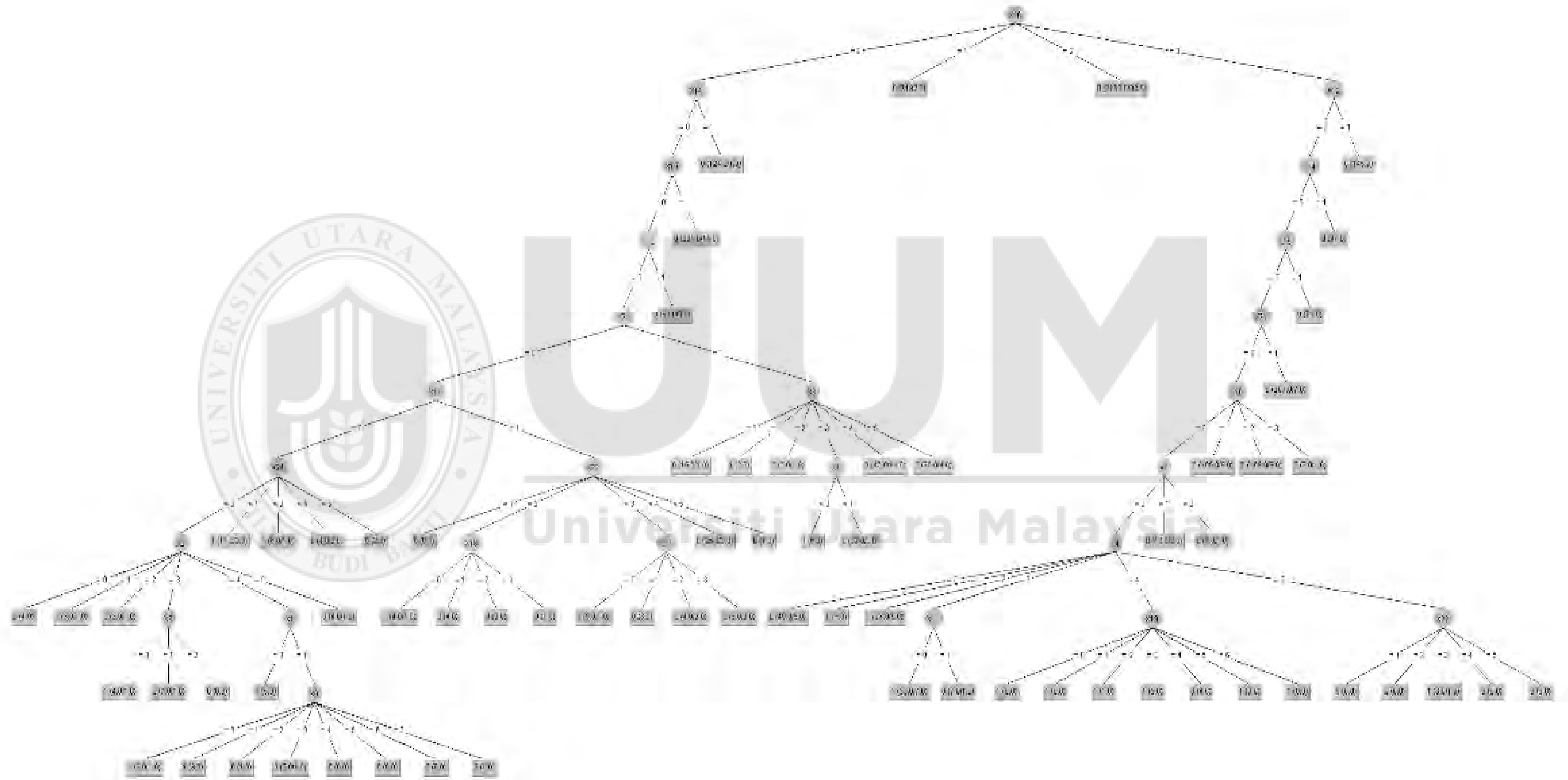


Figure 4.58. Decision Tree Model

Table 4.71

Rules Generated by Decision Tree – J48

Rules										Safe
IF X26 = 1										No(1)
IF X26 = 2										No(0.99)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=3	IF X11=0		Yes(0.83)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=2			Yes(0.79)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=1				Yes(0.82)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=3	IF X17=0			Yes(0.9)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=3				Yes(0.86)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=3	IF X1=0				Yes(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=4	IF X1 = 0		Yes(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=4	IF X19=3		Yes(1)
IF X26 = 0	IF X14 = 0	IF X15 = 1								No(0.94)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=3	IF X8=0		Yes(0.8)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=2	IF X16=0			Yes(0.8)
IF X26 = 3	IF X12 = 1									No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=1			Yes(0.75)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=4	IF X1 = 1	IF X9 = 0	Yes(0.75)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=4	IF X1 = 1	IF X9 = 1	Yes(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=5	IF X22=3		Yes(0.75)
IF X26 = 0	IF X14 = 1									No(0.99)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=2					No(0.92)

IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=1					No(0.93)
IF X26 = 3	IF X12 = 0	IF X14=1								No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=1					Yes(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=1							No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 1							No(0.93)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=0			No(0.84)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=4	IF X19=2		Yes(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=4					No(0.77)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 1				No(0.95)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=3	IF X1=1				No(0.91)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=4				No(0.89)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=5					No(0.86)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=0					No(0.94)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=3	IF X17=3			No(0.73)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=3	IF X8=1		No(0.88)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=5				No(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 2				No(0.88)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=1			No(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=3	IF X11=1		No(0.88)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=4	IF X1 = 1	IF X9 = 3	No(0.83)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=1	IF X3=2					No(0.83)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=0			No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=5			No(0.8)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=2	IF X16=1			No(1)

IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=4	IF X19=4		No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=4				No(0.67)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=3	IF X17=2			No(0.67)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=2			No(0.75)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=3	IF X17=1			No(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=3					No(0.75)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=5				No(1)
IF X26 = 3	IF X12 = 0	IF X14=0	IF X15=0	IF X21=0	IF X16=0	IF X8 = 0	IF X4=5	IF X22=5		No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=0	IF X24=0	IF X2=4	IF X1 = 1	IF X9 = 2	No(1)
IF X26 = 0	IF X14 = 0	IF X15 = 0	IF X12 = 0	IF X21=0	IF X11=1	IF X22=2	IF X16=3			No(1)

4.6 Models Comparison

Table 4.72 shows the summary model for Logistics Regression and J48 algorithms in 5 folds and 10 folds validation. The accuracy of J48 is 97.9% and 98.2% for 5 folds and 10 folds respectively. Logistic Regression has 98.1% of accuracy in 5 folds and 10 folds validation respectively. The precisions of J48 are 50 % and 63.6% for 5 folds and 10 folds respectively whereas the precisions of random forest are 75 % and 83.3% for 5 folds and 10 folds respectively. J48 has 19.3% of recall rate in 5 folds and 30.7% of recall rate in 10 folds. On the other hand, Logistic Regression has 99.5% and 99.6% of recall rate in 5 folds and 10 folds respectively. Overall, both models including logistic regression and J48 has similar accuracy which is around 98%. However, J48 perform slightly better compare Logistic Regression especially comes to 10 folds of cross validation. For 5 folds cross validations, logistic regression performs the best. In most of the time, decision tree should perform the best in business case scenario whereas logistic regression performs better when the noise in the data is less. Logistic regression performs quite good in this dataset due to the pre-processing step before building the logistic regression model. Thus, noise variable is reduced and logistic regression model perform quite good in the case.

Table 4.72

Model Comparison between Logistics Regression and Decision Tree -J48

	Cross validation	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
Logistic Regression	5	98.1	98.5	99.5
	10	98.1	98.6	99.6
Decision Tree - J48	5	97.9	50	19.3
	10	98.2	63.6	30.7

4.7 Conclusion

The findings and analysis of the Decision Tree J48 and Logistics Regression algorithms suggest that they are effective in developing a response model with high accuracy. However, to further improve this research, it is recommended to increase the amount of data used to determine its impact on algorithm accuracy. Additionally, incorporating other algorithms on the variables can help expand the scope of the results, potentially benefitting other fields such as science and economics.

CHAPTER FIVE

CONCLUSION

5.1 Introduction

Based on result and discussion section, it can be concluded that the Decision Tree J48 and Logistic Regression algorithms can be used to evaluate the performances of different classifiers in predicting customer response to personal loan product. The outcomes of each research objective are detailed in this chapter, as well as research contributions, research limitations, and directions for future work linked to this research.

5.2 Objectives of the Study: Revisited

The main goal of this study is to develop customer response prediction model for personal loan product by applying logistic regression and decision tree techniques. This goal was attained by achieving four objectives outlined in Section 1.2. The accomplishment of each goal is highlighted in the following sections.

Objective 1: To prepare suitable data for customer response prediction of personal loan product

The research was started with the investigation of the availability of data for customer response prediction of personal loan product from a commercial bank in Malaysia. The first objective was met by conducting systematic reviews, as detailed in Chapter 3. The dataset consists of customer profile (demographic. Financial etc.), customer profile of customers already holding certain banking product and product information (balances, transaction etc.). Additionally, the needs of preprocessing task were revealed to ensure clean and quality data were fed in developing the prediction model.

Objective 2: To identify a set of variables that influence customers response to personal loan product holding using feature selection techniques

This objective was accomplished through the application of variable selection techniques. Specifically, in this thesis, factor analysis was employed to identify the most significant variables to be included in the prediction model, as detailed in Chapters 3 and 4. As a result, a total of 29 variables were selected for inclusion in the prediction models, ensuring that only the most relevant variables were used in the analysis.

Objective 3: To develop the customer response prediction model for personal loan product using logistics regression and decision tree

The development of the logistic regression model and decision tree model, as discussed and reported in Sections 4.3 and 4.4, respectively, was used to achieve the third objective. Cross-validation approaches of 5-fold and 10-folds were utilized to attain objective number 3, which reduced bias by using most of the data for fitting and significantly reduced variance by using most of the data in the validation set.

Objective 4: To evaluate the performances of different classifiers in predicting customer response to personal loan product

Evaluation metrices are employed to examine the performance of the classification models, thus accomplishing the fourth objective. The best model was selected based on the highest accuracy, precision and recall values. This study has successfully identified the most important variables in predicting customer response for marketing of personal loan product. Furthermore, the suggested technique highlighted the various characteristics of these two classifiers in predicting customer response.

5.3 Contribution of the Study

The findings from this research provide significant contributions to the theoretical understanding of targeted marketing analysis using data mining techniques. Specifically, this study applied factor analysis for variable selection, and logistics regression and decision tree for customer response prediction modeling. The application of factor analysis for variable selection addresses the issue of having too many variables in targeted marketing analysis, which can lead to a complex and less reliable model. By identifying the significant variables for prediction tasks, factor analysis helps to reduce the complexity of the model and improve its accuracy. The use of logistics regression and decision tree as prediction models contributes to the theoretical understanding of targeted marketing analysis. These models have been proven to be reliable and effective in predicting customer responses to a particular product or service, and their application in this study further validates their effectiveness in the field of targeted marketing analysis.

In addition to its theoretical contributions, this research work also provides practical implications for practitioners in the field of targeted marketing. By applying factor analysis for variable selection and logistics regression and decision tree for customer response prediction modeling, this study offers a practical approach for identifying significant variables and predicting customer responses. The application of these techniques in real-world scenarios can help practitioners to better understand their customers and make more informed decisions in product development, marketing strategy, and customer relationship management. Moreover, the findings of this study can be applied to various industries, including finance, healthcare, retail, and e-commerce, where targeted marketing is crucial for business success. Therefore, this research work can serve as a valuable resource for practitioners seeking to improve their targeted marketing efforts and achieve better business outcomes.

5.4 Limitations and Future Works

Despite the research's significant contributions, there are some limitations in the technique and approach that demand more research and investigation. To begin with, the framework given in this thesis has shown how the knowledge discovery database (KDD) process may be used to create a more trustworthy classifier. For statistic-based variable selection, factor analysis was utilized as an example. However, no claim is being made that they are the most effective techniques to employ. As a result, different statistic-based metrics and/or variable selection strategies might be evaluated for inclusion in the suggested framework. It will be interesting to investigate different statistical approaches for variable selection in future work.

Secondly, this study focused solely on one commercial bank in Malaysia, and the targeted product is only a personal loan product. Even though this study has successfully developed a classifier to predict customer response, it should not be considered a complete solution. The results presented in the previous chapter indicate that the classification model may work well on one specific dataset with certain characteristics of the variable selection technique.

Finally, only logistic regression and decision tree (J48) models were covered by the classifiers given in this study. Both classifiers can perform well on the dataset, as seen by the results. However, instead of utilizing the given classifiers, numerous additional classifiers for classification issues should be considered. Several of these strategies, such as those described in (Han et al., 2012), have been demonstrated to be successful and have been widely used in the fields of science, economics, and social sciences.

5.5 Conclusion

The main objective of this study was to develop a predictive model for customer response to personal loan products using Logistic Regression and Decision Tree techniques. To achieve this goal, four specific research objectives were established, and specific methods were used. By applying these techniques, this study demonstrated the effectiveness and originality of using Logistic Regression and Decision Tree techniques in predicting customer response. The results of this study showed that both techniques are reliable and accurate for developing response models, making this research valuable for future researchers. The successful outcome of this study highlights the importance of using these techniques to improve the accuracy of customer response prediction models.

REFERENCES

Abbas, S. (2015). Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset. *Int. J. Comput. Appl.*

Ahmed, M. S. I., & Rajaleximi, P. R. (2019). An empirical study on credit scoring and credit scorecard for financial institutions *Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET)* 8 275–9.

Aliabadi, A., & Berenji, H. (2013). Hybrid model of customer response modelling through combination of neural networks and data pre-processing. In *2013 IEEE International Conference on Fuzzy Systems (FUZZ)*, 2013. pp. 1–4.

Amin, R. K., Indwiarti, & Sibaroni, Y. (2015). Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (case study: bank pasar of Yogyakarta special region) *The 3rd Int. Conf. on Information and Communication Technol. (ICoICT).* pp 75–80.

Aslam, U., Aziz, H. I. T., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience.* 16 pp 3483–8.

Breeden, J. L. (2020). A survey of machine learning in credit risk.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* 16, 321-357.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. http://doi.org/10.1186/s12864-019-6413-7.

Delen, D., Cogdell, D., & Kasap, N. (2012). A Comparative Analysis of Data Mining Methods in Predicting NCAA Bowl Outcomes. *International Journal of*

Forecasting 28 (2): 543– 552.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence* 17 (3): 37– 54.

Fayyad, U. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM* 39 (11): 27–34.

Firman M.M., & Defriani, M. (2020). *Journal Penelitian Ilmu Komputer, System Embedded & Logic,* P-ISSN: 2303-3304, e-ISSN: 2620-3553, Vol. 8 (1): 39-48 (March 2020).

Geng, L., & Hamilton, H. (2006). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys* 38 (3): 9–32.

Ghatasheh, N. (2014). Business analytics using random forest trees for credit risk prediction: A comparison study *Int. Journal of Advanced Science and Technol.* 72 pp 19–30.

Hamid, A. J., & Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining Machine Learning and Applications: *An Int. Journal (MLAIJ).* 3 pp 1–9.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques, *3rd ed. Waltham, MA: Morgan Kaufmann.*

Hormozi, A. M. (2004). Data Mining: A Competitive Weapon for Banking and Retail Industries. *Information Systems Management 21* (2): 62.

Jency, X. F., Sumathi, V. P., & Sri, J. S. (2018). An exploratory data analysis for loan prediction based on nature of the clients *Int. Journal of Recent Technol. and Engineering (IJRTE).* 7 pp 176–9.

Lawi, A., Velayaty, A. A., & Zainuddin, Z. (2018). On identifying potential direct marketing consumers using adaptive boosted support vector machine. *Proc. 2017 4th Int. Conf. Comput. Appl. Inf. Process. Technol. CAIPT 2017*, vol. 2018-Janua. pp. 1–4.

Levin, N., & Zahavi, J. (2001). Predictive modeling using segmentation. *Journal of Interactive Marketing, 15*, 2 - 22.

Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M., & Clark, R. A. (2004). Data Mining Techniques for Cancer Detection Using Serum Proteomic Profiling. *Artificial Intelligence in Medicine 32* (2): 71–83.

Li, Y. (2019). Credit risk prediction based on machine learning methods *The 14th Int. Conf. on Computer Science & Education (ICCSE)*. pp 1011–3.

Madane, N., & Nanda, S. (2019). Loan prediction analysis using decision tree *Journal of The Gujarat Research Society*. 21 pp 214–2.

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (1993). Systems for Knowledge Discovery in Databases. *IEEE Transactions on Knowledge and Data Engineering* 5 (6): 903–913.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. *Expert Systems with Applications 36* (2– 2): 2592–2602.

Pan, Y., & Tang, Z. (2014). Ensemble methods in bank direct marketing in *2014 11th International Conference on Service Systems and Service Management (ICSSSM) 2014*. pp. 1–5.

Parlar, T., & Acaravci, S. K. (2017). Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data. *Int. J. Econ. Financ. Issues*, vol. 7, no. 2, pp. 692–696.

Rais, A. N. (2019). Integrasi SMOTE Dan Ensemble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing. *J. Inform*., vol. 6, no. 2, pp. 278–285.

Shan, L., Lin, L., Sun, C., & Wang, X. (2016). Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. *Electronic Commerce Research and Applications,* 16, 30-42.

Shepard, D. (ed.), (1995), The New Direct Marketing, New York, Irwin.

Shih, Y. Y., & Liu, C. Y. (2003). A method for customer lifetime value ranking – Combining the analytic hierarchy process and clustering analysis. *Journal of Database Marketing & Customer Strategy Management*, *11*, 159-172.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), 4-11.

Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Inf. Fusion*, vol. 54, no. December 2018. pp. 128–144.

Supriya, P., Pavani, M., Saisushma, N., Kumari, N. V., & Vikas, K. (2019). Loan prediction by using machine learning models *Int. Journal of Engineering and Techniques*. 5 pp 144-8.

Ullah, A., Mohmand, M.I., Hussain, H., Johar, S., Khan, I., Ahmad, S., Mahmoud, H.A., & Huda, S. (2023). Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time. *Sensors (Basel, Switzerland), 23*.

Weng, C. G., & Poon, J. (2008). A new evaluation measure for imbalanced datasets. *Conferences in Research and Practice in Information Technology Series*, *87*(81373883), 27–32.

Zhang, J. (2018). Analysis of Neural Network on Bank Marketing Data Dataset Pre-processing. *Coll. Comput. Sci.*

Zhou, W., & G. Kapoor. (2011). Detecting Evolutionary Financial Statement Fraud. Decision Support Systems 50 (3): 570–575.

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm, *The 7th Int. Conf. on Information Technol. and Quantitative Management (ITQM)*. 162 pp 503–13.