A STUDY INTO THE EFFECTIVENESS OF MACHINE LEARNING AND STATISTICAL MODELLING TO PREDICT CUSTOMER'S PURCHASE INTENTIONS OF ELECTRONICS USING THE TIDYMODELS FRAMEWORK

by

Noel Mbeya

Dissertation submitted to University of Plymouth
in partial fulfilment of the requirements for the degree of

MSc Data Science and Business Analytics

University of Plymouth
Faculty of Science & Engineering

September 2023

Copyright statement

Abstract

The business world faces various challenges including the ability to accurately predict customer's purchase intentions. A business' ability to identify the factors and product features that influence purchase intentions can help tailor a business' strategy to increase their competitive advantage, customer acquisition and sales. The continuous improvement of machine learning technologies presents new tools and opportunities for businesses seeking to capitalise on the vast amount of customer data readily available. This study utilises the Tidymodels framework in R to understand its effectiveness in machine learning tasks. The data consisted of 133 survey responses regarding purchase intentions of the Apple M1 MacBook. Numerous machine learning classification algorithms were tuned and tested on the data including logistic regression, random forest, decision trees, naïve bayes, support vector machines and a neural network. This study discovered the neural network model to be the most accurate classifier with an accuracy score of 76%. The most important features influencing purchase intentions of the M1 MacBook were found to be the number of Apple products the participants own, the age of their current computer and the importance of the M1 laptop chip.

Word count: 14,448

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my dissertation supervisor Dr. Luciana Dalla Valle for her feedback, suggestions, and guidance throughout this study research. I am grateful for her support in this research.

I would also like to thank my family for their moral support and unwavering motivation to help me complete this study. Without them, this study would not have been completed without their words of faith and encouragement.

Finally, I would like to thank the Kaggle user Hunter for collecting this data and sharing it on Kaggle, allowing machine learning tasks to be performed on it.

1 Introduction

Technology has transformed the way customers purchase products on a global scale. The introduction of e-commerce in the late 90s ushered in a new age of selling possibilities for businesses. From innovative websites to mobile phones, the advancements in technology have allowed online shoppers to benefit from the ease and flexibility of e-commerce at the click of a button (Pasquali, 2023). In 2022, retail e-commerce sales were estimated to be $5.7 trillion worldwide (Statista, 2023), with these figures expected to grow in coming years. The growth in e-commerce has created new challenges for professionals seeking to successfully optimise their business practices. The world of e-commerce is highly competitive as businesses utilise different methods to attract and retain customers. Customer insights are one of the driving forces behind achieving success in e-commerce (Wong & Marikannan, 2020). A business' ability to understand customers' preferences and why they purchase products is crucial information that businesses can leverage. Ensuring purchase revenue increases through e-commerce now depends on the technological tools being utilised by businesses. Professionals need to create personalised purchasing experiences to remain competitive (Ståhl et al., 2019).

The Covid pandemic contributed to the surge in people purchasing online compared to traditional methods. Despite this new uptick online shoppers, the trend is expected to continue in the future (Wang, 2021). Consequently, the increase in demand for people "working from home" also led to a surge in global sales for electronic equipment including PCs and laptops (Forbes, 2020). In 2023, the global revenue for the laptop market amounts to $122.80 Billion dollars and is expected to grow annually by 0.75% from 2023 to 2027 (Statista, 2023). Businesses that can

understand their customers purchasing behaviours can capitalise on that knowledge to create a competitive advantage and increase their sales and profit in this highly lucrative market.

In the current technological era of big data, large amounts of actionable information and data is more easily accessible than ever before. The term big data represents the vast amounts of high-dimensional or unstructured data that is continuously produced and is difficult to process using traditional analysis methods (Fan et al, 2014). The evolution in computational power in recent years and the large volumes of data available have streamlined the process of data analysis. The analysis of big data presents two main advantages that can be utilised: creating methods for predicting future outcomes and gaining deeper insights into the relationships between dependent and independent variables (Fan et al, 2014). Big data analysis has shown to be integral across different facets of business including fraud detection in finance, optimisation planning in supply chain management, and aiding in social-influencer marketing (Udell, 2014; Ngai et al., 2011; Downing, 2010). Effectively utilising big data analytics tools in business can lead to more efficient customer segmentation, anticipation of customer behaviour, and understanding of customer trends (Duan & Xiong, 2015). Analytics-driven insights create opportunities for businesses to effectively react to real-time changes occurring in the business landscape, including reactive marketing, new product design, and customer relationship management. To achieve the most value from analytical insights, business analytics must be closely linked with business strategy and organisational processes (Duan & Xiong, 2015).

Machine learning is transforming the e-commerce industry (Policarpo et al., 2021) as this method is one of the big data analytics tools that businesses can employ to predict customers purchase intentions. The surge in data and computational power over the years enabled machine learning and artificial intelligence to be at the forefront of handling complex data analysis tasks (Badillo et al., 2020). Artificial intelligence “uses a computer to model intelligent behaviour with minimal human interaction” (Hamet & Tremblay, 2017). Machine learning is a subset of artificial intelligence and is defined as developing computational algorithms that can learn from experience and build models based on data to make predictions on new observations (Zhou, 2021). The process of a computers learning patterns from data and applying them to unseen data is a concept that has been present since the 1950s (Badillo et al., 2020). However, modern computing advancements enable these processes to be completed almost instantaneously.

Machine learning requires different algorithms to solve data analysis tasks as no algorithm can adapt to every possible task. The algorithm to be implemented can vary depending on multiple factors including the type of problem to be solved, the best fitting model for the task, and the number of variables (Mahesh, 2020). Supervised learning is an estimation method within machine learning that uses algorithms to reach a desired result/output by using existing data as an input (Ozdemir & Turanli, 2021). This method aims to explain the relationships between input and output values as well as creating functions to predict output values based on a set of input values. Classification and regression are among the most used supervised learning methods. Algorithms for classification include logistic regression (LR), naïve bayes (NB), decision tree (DT), random forest (RF), support vector machines (SVM), and neural networks (NN).

Predicting purchase behaviour is an interesting and challenging task that empirical studies have investigated using a range of different techniques to understand customers and develop the most accurate methods for predictions. A business' ability to understand why their consumers purchase their products and which features influence their purchase intentions can have a significant impact on a business' strategy to increase their sales, customer acquisition and competitiveness (Qiu et al, 2015).

Previous studies on purchase intentions have explored this topic from a machine learning and a traditional purchasing and marketing perspective. Studies from a traditional purchasing perspective found factors such as risk, specifically security risk to have a significant impact on customers purchase intentions (Kamalul Ariffin et al., 2018). Trust in a brand, the merchant's perceived reputation and the ease of use of their website have also shown to play a role in consumer's purchasing choices. Additionally, different characteristics of the consumer, product, brand, and demographic have also shown to have an impact on purchase intentions.

Studies exploring purchase intentions from a machine learning perspective aimed to create highly accurate models capable of predicting which new customers would make purchases. Current studies on the topic of purchase intentions utilise machine learning techniques on consumers' online behaviour in the form of clickstream data to develop the most accurate models for predicting purchase intentions (Surjandy et al., 2021). A variety of different models and frameworks have been employed to gain a deeper understanding of what influences consumers to purchase. Utilising machine learning techniques for predicting customers interests, choices and requirements has shown to be advantageous for businesses by enabling them to

react to consumer changes through marketing and operational actions (Chaitanya & Gupta, 2017; Zhao et al., 2017). Furthermore, the implementation of machine learning models allows for businesses to create customer profiles, create customer targeted campaigns, and provided personalised ads to help acquire new customers and retain current customers (Ozdemir & Turanli, 2021). Consequently, identifying why customers purchase your products and capitalising on that information can lead to an increase in income and the development of better strategies.

1.1 Research Problem

E-commerce can significantly benefit from employing big data analysis methods through artificial intelligence and machine learning. As online shipping trends continue to change, understanding and pre-emptively predicting purchase behaviours has become an important task for businesses seeking to automate their marketing strategies to increase revenue and target consumers (Ha et al., 2021). With e-commerce being one of the leading methods for consumer purchases, analysing the online information available about customers and their behaviour can lead to valuable insights for designing marketing campaigns for reaching broader target audiences, promoting greater customer involvement, and achieving higher investment returns (Trivedi et al., 2022).

1.1.1 Research Objectives

This project aims to create an accurate classification machine learning model that can predict whether participants would purchase the Apple M1 laptop or not. This study uses the tidymodels framework within RStudio to gain a deeper analysis. This framework contains a set of packages designed for statistical modelling and machine learning utilising the tidyverse principles (Kuhn & Wickham, 2020). Commonly used

classification algorithms including logistic regression, decision trees, random forests, support vector machines and naïve bayes will be tested. A feed-forward, multilayer perceptron neural network will also be used to test and compare the accuracy with the other models.

The three objectives for this study can be defined as:

1. To determine which features which have the most significant impact on purchasing decisions.
2. To determine which classification model produces the most accurate results amongst the ones that are tested.
3. To determine how effective the tidymodels framework is for creating classification algorithms.

This research paper aims to utilise the tidymodels machine learning framework to analyse consumer purchase intentions. This research is divided into the following sections. Section 2 reviews the current utilised methods and the understanding of consumer purchase intentions in the e-commerce context. Section 3 presents a conceptual framework for the proposed methodology of this study and the algorithms to be utilised. Section 4 provides exploratory analysis of the data and the results of study. In section 5, the results of the study are discussed. Finally, section 6 highlights the conclusions, limitations, and future of research within this topic area.

2 Literature Review

Previous studies on purchase intentions have explored this topic from a machine learning and a traditional purchasing and marketing perspective. Studies in the traditional online purchasing intentions found different factors to influence a consumer's choice before making a purchase. According to Parihar & Yadav (2022) different factors could influence purchasing behaviour including social factors, lifestyle, culture, education, occupation, and previous purchases. Akar & Nasir (2015) conducted a review of online consumer purchase behaviour in extant literature and split consumer purchase behaviour into four main categories: consumer, merchant, website, and product characteristics. Trust, risk, attitudes towards online purchasing, and subjective norms were found to be the main consumer characteristics influencing online purchasing intentions. Trust in vendors or websites was one of the most important factors influencing consumer's purchase behaviour (El Ansary & Roushdy, 2013; Kamtarin, 2012). A lack of trust has a negative impact on purchase intentions as consumers prefer not to shop where websites or vendors are not trustworthy. Perceived risk also has a negative impact on consumer purchase intentions as consumers may feel uncertain about online shopping due to security or privacy concerns (Akar & Nasir's, 2015; Li et al., 2007). Product price, type, perceived quality, and product knowledge have shown to have a significant impact on consumer's purchase intentions (Gatautis et al., 2014). Product information and price is particularly important as the diversity of online shopping allows consumers to explore different options to find the most optimal choice before making a purchase.

Gender has shown to influence purchase intention with most studies finding men to be more likely to purchase online compared to women. Additionally, most studies investigating the importance of age found it had no significant impact on purchase behaviour (Doolin et al., 2005; Thamizhvanan & Xavier, 2013). However, Clemes et al. (2014) found that younger consumers were more likely to shop online due to their increased internet experience; as increased internet experience has shown to increase the likelihood of consumers shopping online (Gong & Maddox, 2011; Saprikis, 2013). More educated consumers with higher incomes were more likely to shop online compared to their counterparts (Thamizhvanan & Xavier, 2013). A website and its services also contribute substantially to a consumer's purchase intentions. A business' website serves as a crucial aspect for attracting new customers through marketing and increasing sales (Parihar & Yadav, 2022). Seckler et al. (2015) explained that websites may increase the business' popularity and build trust in consumers. Factors such as service & after-service quality, online advertisements, payment & delivery have all shown to have a significant impact on purchase intentions (Clemes et al., 2014; Momtaz et al., 2011; Gatautis et al., 2014; Oncioiu, 2014). Additional characteristics such as the reputation of the merchant/brand and the use of social media have shown to have a significant positive impact on consumer purchase intentions, if these characteristics are executed faultlessly by the merchant (Aghdaie et al., 2011; Vinerean et al., 2013).

Overall, extant literature has studied the different characteristics that influence online consumer purchase intentions and consumer, demographic, website, and product characteristics have all shown to have an impact. Understanding these factors to pre-emptively predict customers behaviour patterns can have multiple applications for businesses including increasing sales, identifying high value consumers, and

reallocating resources to improve consumer's experiences (Tufail et al., 2022). Akar & Nasir's (2015) review highlights the importance of online shopping as this method of purchasing continues to attract more customers due to its product availability, convenience, and cheaper prices (Adnan, 2014; Clemes et al., 2014; Vahidehi, 2014). Additionally, this review also explores the need for merchants to increase their utilisation of social media as it is a low-cost powerful tool for sales promotion to attract and reach the maximum number of potential consumers (Chaturvedi & Gupta, 2014).

Policarpo et al.'s (2021) systematic review of machine learning applications in e-commerce found the main applications to be purchase prediction, recommendation systems, fraud detection and discovering relationships between data. Most studies employed machine learning algorithms to predict purchases and to discover the relationship between data with the aim of understanding consumer behaviour to serve as the foundation for developing strategic objectives (Policarpo et al., 2021). Machine learning methods can perform analysis that can be used as a basis for strategic decisions to increase profitability (Policarpo et al., 2021). Predicting purchase behaviour aims to increase profits by attempting to ensure the acquisition of potential customers (Mokryn et al., 2019).

Studies on the topic of purchase behaviour in a machine learning context are mostly comparative studies working on the same or similar datasets to create the most accurate classification model for purchase predictions and understanding the data. Classification uses a machine learning algorithm to train and test the attribute inputs from a dataset to predict a categorical attribute known as a class label. As a result, this allows the classification model to be measured in how accurate it is at predicting

purchase intentions. Empirical studies have employed a range of different classification algorithms that vary in accuracy when predicting purchase intentions. The novel changes between studies are comprised of the methods used when preprocessing e.g., feature selection and addressing class imbalance issues. Additionally, the theoretical frameworks and methods of interpreting model results differ throughout the studies.

Most e-commerce data used to predict purchase behaviour is available in the form of clickstream data which is defined as a track record of user's online behaviour while browsing the web or mobile applications to understand visitor traffic in a singular session (Bucklin and Sismeiro, 2009). This can include information on the amount of time a user spends while visiting specific pages, the number of pages they view or the number of clicks on an advertising banner (Moe and Fader, 2004). The application of clickstream datasets has been researched across different applications of business including customer profiling, segmentation, and purchase behaviour. Clickstream data presents an alternative perspective in understanding the factors that influence customer's purchase intentions. For example, bounce rate defines the number of customers who click off a website after visiting only one page. A high bounce rate can indicate a business need to improve their landing page to retain more customers. Additionally, learning more about customers through clickstream data such as the traffic type, region, and webpage visits can unveil the potential new customers that can be targeted through marketing to increase purchase revenue. However, one of the drawbacks of clickstream data is that it provides only one form of implicit feedback and does not directly represent or explain every decision leading to a purchase (Wen et al., 2023). Clickstream data can produce millions of records and requires vigorous systems to track and store the information about users.

Additionally, businesses collecting clickstream data must ensure their data security is robust to safeguard the individual privacy of their consumers, to adhere to General Data Protection Regulation (GDPR) standards and to prevent the risk of re-identification attacks (Vamosi et al., 2022). Nevertheless, clickstream data can be analysed to further understand and predict customers' purchase behaviours.

Before implementing machine learning algorithms on the data, studies have employed preprocessing steps to ensure the model training and testing produces optimal results. Preprocessing steps include normalisation, feature selection and amending class imbalance in the data. Normalisation involves transforming the data onto a similar scale to allow for effective comparisons between features as this can improve model performance. Multiple feature selection methods have been employed in extant literature and feature selection has shown to be crucial in machine learning tasks as it improves accuracy by eliminating features that lack in the ability to provide meaningful insights. Feature selection is possible through the embedded method of using random forest to find the variables with the most importance towards the target class. Filter and wrapper methods of feature selection are some of the most used in research. Wrapper methods use machine learning models and evaluate model performance by adding or removing features to create a subset until the model's predictive ability cannot be improved further. This method generates the optimal subset of features to generate the best predictive results (Trivedi et al., 2022). Filter methods are possible through correlation, mutual information (MI) and minimum redundancy maximum relevance (mRMR). Correlations aims to visualise the features with the highest correlation to each other to reduce multicollinearity. MI measures the mutual dependence between two variables even when there is a nonlinear relationship between them. mRMR selects

a subset of features with the highest relevance to the target variable while also having the lowest redundancy. mRMR has shown to produce the highest model metrics while reducing the number of features needed for the model (Sakar et al., 2018). Feature selection can aid in significantly reducing training times, especially filter methods as they do not utilise a machine learning model. Addressing the class imbalance issue in a dataset is possible through sampling methods. Additionally, The Balanced Bagging Classifier is another method of dealing with class imbalance issues in a dataset. This ensemble method can effectively reduce overfitting, handle samples with different feature weights and easily adapt to imbalanced data leading to improved accuracy and model robustness (Chen et al., 2023). Sampling methods such as Synthetic minority oversampling technique (SMOTE), under sampling, oversampling and Random Over Sample Examples (ROSE) can create balanced datasets by oversampling the minority class or under sampling the majority class.

Various methods of analysing e-commerce data without machine learning exist including math programming, greedy programming and statistical analysis (Policarpo et al., 2021). However, studies on the topic of consumer purchase behaviours have employed a variety of machine learning models and frameworks including logistic regression (Shao & Li, 2014), Bayesian logistic regression (Li & Kannan, 2014), game-theory (Berman, 2018) and hidden Markov models (Abhishek et al., 2015). Psychological and cognitive theories such as the theory of planned behaviour and the technology acceptance model have been investigated to further understand the behaviours that drive consumer's purchasing behaviours (Borres et al., 2023; Lu et al., 2021). Additionally, different types of datasets have been analysed by machine learning models including surveys in the case of Borres et al. (2023). Their study showcased the impact of analysing survey data to predict purchase behaviour of

consumers. Not only can the questions in a survey be tailored to the precise research questions being investigated, the task of inferring the real-world applications of the data is significantly reduced. This overcomes an issue with clickstream data as analysing a consumer's behaviour on webpages does not always translate into genuine purchase intent. On the other hand, collecting survey data can be time consuming, costly, and may lead to biases and sampling issues. Most machine learning studies conduct their work utilising a variety of different machine learning libraries including TensorFlow, Weka, Scikit-learn and MATLAB. Policarpo et al. (2021) stated the future of machine learning tasks lies in serverless cloud computing platforms. Machine learning in cloud computing allows for cheaper and faster task executions as well as providing the option to scale the data being utilised.

Empirical studies have tested multiple classification algorithms on clickstream data. Bing & Yuliang's (2016) study compared the use of a C4.5 decision tree against a naïve bayes classifier on 15,000 web server sessions from a Chinese holdings company to determine the most accurate in prediction algorithm. Their results showed that the decision tree algorithm performed better in accuracy by 10% compared to naïve bayes and the decision tree was more effective in dealing with a larger scale of data. In addition, the decision tree was able to avoid any issues that derived from the sample distribution as decision trees do not rely on prior probability of the sample compared to naïve bayes. Unfortunately, this study did not apply any feature selection methods or employ a sampling technique to address the class imbalance in the dataset.

Ozdemir & Turanli's (2021) study tested logistic regression, support vector machines and naïve bayes algorithms on google analytics data of 100,000 consumers from a Turkish e-commerce site to predict purchase behaviours. Their results showed that logistic regression's overall model accuracy outperformed support vector machines and naïve bayes. However, logistic regression had the lowest precision of the three algorithms when predicting the correct number of customers who made purchases. Consequently, their study stated that support vector machines performed the best when considering the overall model accuracy and precision of correctly predicted customers who made a purchase. On the other hand, this study did not apply any feature selection tools or take steps to address any imbalances in the dataset. Ozdemir & Turanli (2021) stated that future studies on purchasing intentions should focus on using different data sizes to create a more accurate model.

Along with commonly used classification algorithms, ensemble and bagging methods have been tested in multiple studies and have shown to produce consistently high results. Kabir et al.'s (2019) study used web server information from 12,330 session from a sportswear website. This dataset was obtained from the UCI repository and is a famous dataset used to predict purchase intentions. Kabir et al. (2019) found random forests paired with the ensemble method of gradient boosting to be the most accurate model for predicting the purchase intentions of online shoppers. This ensemble method outperformed traditionally used classification algorithms such as a decision tree, support vector machines and naïve bayes. However, this study did not address the class imbalance in the dataset or apply feature selection. Zhao et al. (2016) found the ensemble method of random forest to be the more accurate than logistic regression, Naïve bayes, and support vector regression (SVR) when predicting purchase intentions from clickstream data of 19,500 online catalogue

customers. Additionally, Martinez et al.'s (2020) study predicted purchase intentions on a dataset of 10,136 customers of a B2B manufacturing company and found gradient tree boosting to be more accurate and have a higher Area Under the Curve (AUC) value than a logistic lasso regression and extreme learning machine algorithms. Ensemble methods such as boosting and bagging can be paired with other algorithms to improve model accuracy as ensemble methods aim to improve a weaker learner into a strong learner (Kabir et al, 2019). Consequently, random forest have become increasingly popular due to its optimal performance in classification tasks and its ease of use compared to neural networks (Policarpo et al., 2021)

The study conducted by Noviantoro & Huang (2021) aimed to test different classification algorithms on the same UCI dataset as Kabir et al. (2019). This study used the wrapper feature selection method to improve the overall model performance. The wrapper method of feature selection utilises a machine learning algorithm to identify the optimal subset of input features to improve the model's predictive ability. This study discovered that a neural network classification algorithm had the highest accuracy and F score when classifying purchase intentions from data of online shoppers. Decision trees followed second from neural network in accuracy and F score. Random forests and neural network we considered as the top supervised learning algorithms for their study based on their Receiver Operating Characteristic curve (ROC) analysis. Noviantoro & Huang's (2021) study produced impressive metrics however, the study did not employ any methods to address the class imbalance in the dataset. Suchacka & Stemplewski's (2017) study found their neural network model to be 99.6% accurate while only using 7 attributes to predict purchase intentions. Their study used information collected from web server logs of an online bookstore and the dataset consisted of information from 33,354 sessions.

Suchacka & Stemplewski (2017) stated that neural networks are a promising algorithm that can potentially be improved further by utilising an increased number of attributes to predict purchase intentions.

Kurwanian et al. (2020) conducted a comparative study to assess the impact of different applied methods on model accuracy. Data level methods such as Particle Swarm Optimisation (PSO) feature selection, ten-fold cross validation and class-rebalancing with SMOTE were tested on the dataset as class imbalance is one of the main issues in machine learning and has shown to have a negative impact on the performance of machine learning algorithms (Kurwanian et al., 2020). Additionally, the data level methods were coupled with algorithm level approaches by incorporating AdaBoost with the different classifiers. This study used the UCI repository dataset of online shopper behaviour to predict purchase intentions and tested C4.5, NN, RF and SVM classifiers. This study was split into three separate experiments. The first experiment assessed the performance metrics of the four classification models on the unchanged and imbalanced UCI dataset. The second experiment incorporated ten-fold cross validation, oversampling of the dataset using SMOTE and the AdaBoost ensemble method. The third experiment was identical to the second however, PSO feature selection was incorporated along with the respective other methods. PSO is an optimisation technique used to identify the best subset of features capable of achieving the highest model evaluation metrics. Their study found the second experiment to be better than all other experiments on all evaluation metrics. Following previous studies by Cheng et al. (2018) and Rana et al. (2015), Kurwanian et al. used AUC as their main model evaluation metric as they were working with a class imbalanced dataset. AUC evaluates the classifier's ability to distinguish between positive and negative classes across different thresholds and

AUC score is not affected by the class distribution. Consequently, using AUC as the evaluation metric allows for a more reliable measure of the classifier's performance and improves comparison across different studies (Cheng et al., 2018). Kurwanian et al. (2020) found random forests coupled with the AdaBoost ensemble method to be the most accurate model with an AUC score of 96%. The future research for this study suggested using different classification algorithms to assess the impact on accuracy and AUC. This study employed feature selection methods and addresses the class imbalance issues. As a result, this study showed an improvement in evaluation metrics compared to previous studies. Conversely, the experiment that produced the best results did not include feature selection.

Wong & Marikanna's (2020) study tested decision tree, random forest, neural network, and support vector machine models on 56,000 customer records from a Brazilian e-commerce retailer. The authors used a decision tree for feature selection and found that utilising the 5 most important features influencing the target variable significantly reduced computational time while maintaining high performance across all evaluation metrics. Additionally, normalising the skewed dataset has shown to improve the computational speed during model training in a previous study (Han et al., 2011). However, normalising the data showed to have no effect on computational speeds on the model evaluation metrics in this case. SMOTE, Under-sampling, Oversampling and ROSE were applied to the data to address the class imbalance issues in the minority class of the target variable. Although SMOTE is one of the most used techniques, their results showed that all four class imbalance techniques had no effect improving the F1 score and specificity when tested with a decision tree classifier. The results of this study found random forest to perform the best amongst the other classifiers in accuracy, sensitivity, and specificity. The results of the study

did show that random forest had a long computation time however, this was resolved by reducing the numbers of cross validations the model was trained on. As a result, the model retained its high levels of accuracy and specificity while significantly reducing computation time. Additionally, the study found decision trees to have a slightly lower accuracy however, the decision tree model was able to generate results in seconds. This study also showed the importance of computational speed when dealing with large datasets, an important aspect of machine learning that data scientists need to consider in real-world applications.

Lee et al.'s (2021) study aimed to understand which machine learning model and sampling method was the most suitable for predicting consumer purchase behaviour. Additionally, the authors sought to interpret the impact of the results in the context of e-commerce and find a solution to the black box problem in machine learning. The black box problem is the result of the researcher being unable to see and understand exactly how the model arrives at specific decisions or which features it used for making predictions (Yale et al., 2017). This study employed classification tree, NN, K-nearest neighbour (KNN), LR, SVM, RF, GBM and XGB on google analytics data from 374,749 customers of a merchandise store. SMOTE Oversampling was found produce the most optimal metrics when addressing the class imbalance issue in the dataset. To interpret the impact of each variable on purchase intentions, the SHAP explainable machine learning framework was employed and the contribution of each feature on the model output is allocated based on its marginal contribution (Shapley, 1953). Understanding the relationship between non-linear data is possible through the attribution analysis of Shapley additive explanation (SHAP). This method can map high-dimensional feature spaces to lower dimensions, which helps in modelling non-linear relationships. This method is compatible with all models and is based on

Shapley values in game theory. SHAP can quantify the importance of input features on the model's predicted output (Lundberg & Lee, 2017). Their results found the ensemble method of extreme gradient boosting to be the most optimal algorithm for predicting purchase behaviour. Like other studies on the topic, Lee et al.'s (2021) study further reinforced ensemble methods being superior predictive algorithms. Additionally, the use of explainable machine learning provided deeper insights into impact of consumer's choices; an aspect typically overlooked in machine learning studies on this topic. The SHAP model was able to reveal that page views and duration had the biggest positive impact on purchase behaviour. The use of SHAP explainable machine learning highlighted the importance of learning which individual factors have the most significant impact on purchase intentions. As a result, SHAP revealing that page views and duration have a significant impact on purchase intention can help businesses implement website design improvements to entice consumers to stay on their sites longer, increasing the chance of more sales conversions.

Hanami & Muzakki (2021) conducted research on the same UCI dataset and expanded on previous research by employing hyperparameter tuning on LR, KNN, DT and RF algorithms. Hyperparameter tuning involves adjusting different combinations of model parameters to optimise the model's performance. Hyperparameter tuning has shown to increase model accuracy despite the significantly longer training times (Hanami & Muzakki, 2021). Feature selection was also employed and showed to have a small increase in accuracy for all models. This study found random forests to produce the most accurate results and improved on the previous studies by a small margin.

Sakar et al.'s (2018) study tested different classification algorithms to make real-time predictions of consumer purchase intentions, coupled with a long short-term memory recurrent neural network to predict the likelihood of the consumer not making a purchase and abandoning the webpage. RF, SVM and NN were tested on 12,330 consumer records from a google analytics dataset. This study used SMOTE oversampling to address the class imbalance issue in the dataset and used filter-based feature selection methods. Sakar et al. (2018) used correlation, mutual information (MI) and minimum redundancy maximum relevance (mRMR) filter methods in their research. The results of this study showed that mRMR feature selection paired with the NN algorithm produced the best evaluation metrics when compared to the other tested algorithms. Page value was shown to have the highest feature ranking when predicting purchase intentions.

More recent studies on consumer purchase intentions have sought to understand this topic by utilising less commonly used mathematical methods and psychology-based cognitive biases to assist in result interpretation. Chen et al.'s (2023) study aimed to showcase the importance of model explanation in the domain of purchase intentions in search advertising. Chen et al. explained that consumers' behaviour decisions are influenced by factors such as emotions, cognition, and society. This is known as the anchoring effect in psychology and can cause biases in consumers' purchase behaviours (Simon, 1955). The influence of anchoring effects can help further understand consumers' purchase behaviours. Additionally, the adoption of explainable machine learning techniques aids in solving the black box problem associated with machine learning and improves our ability to predict and explain consumer behaviour (Chen et al., 2023). Using SHAP, explainable machine learning can illustrate the correlation between behaviour patterns and purchase decisions.

The study was split into two stages, the first stage predicted customer purchase intentions on 6,224,279 clickstream records from a Chinese retailer consisting of information related to the products, merchants, and users. The study conducted multicollinearity and stability analysis as well as employing five-fold-cross-validation and the BalancedBaggingClassifier algorithm to deal with class imbalance in the data. LR, AdaBoost (ADA), eXtreme Gradient Boosting (XGB), MLP, NB, and RF models were implemented, and the authors found random forest algorithm to have the most optimal performance across the tested evaluation metrics. As a result, the RF model was utilised in the second stage and input into the SHAP explainable framework to visualise the impact of input feature on the model's output. The results from the SHAP framework found product information had the most significant impact on purchase behaviour (e.g., price, sales levels, display priority etc.) while user information showed to have a lower impact (e.g., age and gender).

Chen et al.'s (2023) study showed the importance of being able to interpret feature importance to tackle the common issues such as the black box problem in machine learning. Additionally, explainable machine learning techniques such as the SHAP framework have shown to be beneficial in identifying patterns and open a new research direction in predicting consumer purchase behaviour. These models can also assist businesses in real-world applications by serving as justification for strategic decisions.

Trivedi et al.'s (2022) study expanded on previous studies by implementing rigorous feature selection, model analysis tools, and an improved decision tree classifier coupled with the stacking ensemble method. The new classifier used a combination of C5.0, RF and SVM classifiers to predict consumer's purchase intentions. This was

made possible by training the three classifiers on different subsets of the data and using those outputs to train the stacked meta classifier which used C5.0 as the president to make the final classification decision. Stacking is one of the less commonly used ensemble methods however, it is considered superior to traditional ensemble methods such as bagging and boosting as it combines the strengths of multiple diverse models. Nevertheless, stacking requires more computational resources and precise parameter tuning to avoid overfitting. Trivedi et al. (2022) analysed multiple studies that utilised the UCI purchase intentions dataset and found that most of these studies were homogenous with little innovation, did not use appropriate feature selection, did not address the class imbalance issue, or use further model evaluation metrics like Cohen's Kappa or implement more train-test partitions. Cohen's Kappa statistically measures the agreement in classification tasks while also considering the agreement occurring by chance. Additionally, this study aimed to predict purchase intentions and target the customers with lower purchase intentions with personalised ads to increase sales, improve customer acquisition and retention (Trivedi et al., 2022). The wrapper feature selection method of random forest greedy search was implemented along with under sampling to deal with the class imbalance in the dataset. The models were tested on 12,330 sessions from the UCI purchase intention dataset. Evaluation of the models' performance was compared on multiple train-test partitions incorporating the dataset with and without feature selection. The results showed that the improved stacked decision tree model outperformed the other models on most metrics including the kappa statistic for model evaluation and only had a slightly lower accuracy than random forests. This study showed that ensemble stacking methods, feature selection and multiple train-test partitions can improve overall accuracy, when compared to extant literature on

the same dataset. The model also produced the highest specificity when tested on a balanced dataset enabling it to correctly make prediction in the minority class.

Machine learning plays a vital role across multiple fields in the modern day. However, some authors criticise the methods and lack of transferable results utilised by researchers in this field. Wagstaff's (2012) article on machine learning laid out the main areas of concern surrounding publications and research methods in machine learning. Despite the article being published as machine learning was increasing in popularity, some of the issues mentioned are still prevalent in modern research on the topic. For example, most publications perform experiments on UCI archive and synthetic datasets. Although experimenting on familiar datasets allows for easy comparisons with other studies, experiments vary in their methodology and implementations, nullifying the impact of direct comparisons.

Wagstaff (2012) also stated that most of the machine learning researchers rarely make meaningful interpretations of the results as they usually do not have domain knowledge of the datasets they are experimenting on. Additionally, researchers implementing different machine learning models rarely need to make further interpretations past the evaluation metrics of their models. Evaluation metrics provide no useful information regarding the generalisation or impact on the domain specific problem being addressed. Conclusively, Wagstaff (2012) highlighted the need for deeper interpretation of the domain-related impact of machine learning research. Data scientists interpreting results, collaborating with domain experts, and publishing their findings to the relevant domain communities will bridge the gap and reduce the disassociation between machine learning contributions and their impact on real-world problems. Many studies on purchase intentions discussed in this

section utilise the UCI dataset consisting of google analytics data. However, a lack of domain knowledge is evident as very few interpret the results past the machine learning context or provide real-world recommendations that can be implemented based on their findings.

Empirical studies have produced a variety of results when classifying purchase intentions. However, numerous studies in the e-commerce domain have found Random Forests to be one of the most consistently accurate classification algorithms (Hamami & Muzakki, 2021; Kurniawan et. al, 2020). Ensemble methods such as boosting and stacking have also shown to produce consistently high results (Trivedi et al., 2022; Lee et al., 2021; Martinez et al., 2020; Kabir et al., 2019). Wong & Marikanna (2020) stated that most studies on this topic found random forest to be the highest performing algorithm compared to other classifiers. This is likely because random forests create multiple decision trees that can be paired with ensemble and hyperparameter tuning methods to further increase their accuracy (Hamami & Muzakki, 2021). In contrast, decision trees and support vector machines have shown to be good classifiers while naïve bayes had a consistently lower accuracy than other algorithms. Neural networks are less commonly used, most likely due to their complexity however, Noviantoro & Huang (2021) and Suchacka & Stemplewski's (2017) studies found neural networks to be the most accurate algorithm for predicting purchase intentions and Suchacka & Stemplewski stated that neural networks should be explored further in future research of this topic. Additionally, Borres et al. (2023) also stated that deep learning neural networks can generate numerous accurate predictions, making them highly beneficial in predicting human consumer behaviour. The studies on this topic have shown that most algorithms perform well for classification tasks and model performance predominantly depends

on the dataset and preprocessing methods employed throughout the research. The use of explainable machine learning through frameworks such as SHAP is emerging as a powerful tool in gaining a deeper insight into consumer's purchasing behaviours. Going beyond model metrics and being able to statistically explain the importance of attributes on purchase intentions serves as a powerful justification tool for guiding business' strategic decisions to help increase their profits and customer acquisition.

3 Methodology/Procedure

To investigation customer's purchase intentions, a classification algorithm will be used to predict which customers are likely to purchase the product. Tidymodels is a framework for machine learning and statistical modelling within RStudio and consists of a set of specialised packages. This framework streamlines the machine learning and statistical modelling process by creating a consistent workflow from preprocessing up to model evaluation. Multiple models can be created, tuned, and evaluated simultaneously. Most of the processes are automated removing the need for manual coding input. As a result, this framework serves as a valuable tool for non-expert individuals and businesses seeking to carry out machine learning analysis on data of all sizes.

Empirical studies have used online data such as google analytics to make predictions for purchase intentions. this study will analyse survey responses as this this type of data can gain a deeper insight into attributes influencing purchase intentions that are not easily interpreted from google analytics data. For example, customers' trust towards a brand has shown to have a significantly positive impact on a customer's purchase intentions (Aghdaie et al., 2011; Momtaz et al., 2011) however, this is not directly measurable through online data unless specifically stated and collected. Additionally, privacy guidelines that all e-commerce businesses must adhere to serves as a partial barrier for collecting consumer information that is useful to the business. Studies have shown that collecting data on user's experiences through surveys, feedback or social media can help to improve customer involvement by identifying the motivations that attract consumers to products and implementing strategies to capitalise on the consumer insights (Lim et

al., 2012). Product feedback contributes significantly to improving consumer engagement as feedback relays the consumer's needs and feelings towards a product. Consequently, product feedback serves as a medium to understand the motivations to help retain consumers. With the correct analysis, consumer feedback can be utilised to improve the user experience and make the appropriate product and service changes to retain and attract new consumers (Zhao et al., 2021)

The dataset to be used in this project was found on Kaggle (2021) and the responses were collected months after the latest Apple M1 MacBook was released in 2020. The survey aimed to collect attitudes towards the laptop and whether the respondents would purchase it. Apple stated their innovative new M1 computer chip enabled their laptops to have the world's fastest CPU in low-power silicon, best CPU performance per watt and fastest integrated graphics in a personal computer (Apple, 2020). The laptops were available in different sizes with notable price differences. Additionally, they were computationally powerful, and a consumer's budget and computing needs would need to be consideration before purchasing the laptops.

The dataset contains 133 responses from a survey sent to different tech-savvy groups over a two-week period. The dataset has 23 total attributes including the respondent's choice as to whether they would purchase the M1 laptop. The questions in the survey were split into three main sections: general laptop questions, product specific questions and demographic questions.

<u>The variable names, general laptop questions and measures are as follows:</u>

- **trust_apple -** Do you trust the Apple brand: (Yes, No)
- **Interest_computers -** Level of interest in computers: (1 Not interested - 5 Very interested)

- **age_computer -** Age of your current computer: (0 is less than one year - 6 years or more)
- **user_pc or mac -** Type of computer you currently use: (0 PC, 1 Apple, 2 Hp or Other)
- **appleproducts_count -** Count of Apple products you own: (0 - 10 or more)
- **familiarity_m1 -** Familiarity with brand/M1: (Yes, No)

The variable names, Apple M1 laptop features questions and measures are as follows:

- **f_batterylife** - Importance of battery life: (1 Not important - 5 is very important)
- **f_price** - Cheaper price: (1 Not important - 5 is very important)
- **f_size** - Thinner size of the computer: (1 Not important - 5 is very important)
- **f_multitasking** - Improved multitasking power: (1 Not important - 5 is very important)
- **f_noise** - Less noisy: (1 Not important - 5 is very important)
- **f_performance** - Improved performance: (1 Not important - 5 is very important)
- **f_neural** - Neural engine: (1 Not important - 5 is very important)
- **f_synergy** - How important is a seamless experience: (1 Not important - 5 is very important)
- **f_performanceloss** - A small loss in performance: (1 Not important - 5 is very important)
- **m1_consideration** - M1 Chip into account in the selection process of buying a new Apple computer: (1 Not important - 5 is very important)
- **m1_purchase** - Would you buy one of the new Apple M1 Macs: (Yes, No)

The demographic questions are as follows:

- **Gender, Age group, Income group, Status and Domain**

The conceptual framework displayed in Figure 1 outlines the proposed steps for this research. The framework for this research can be split into three main steps.

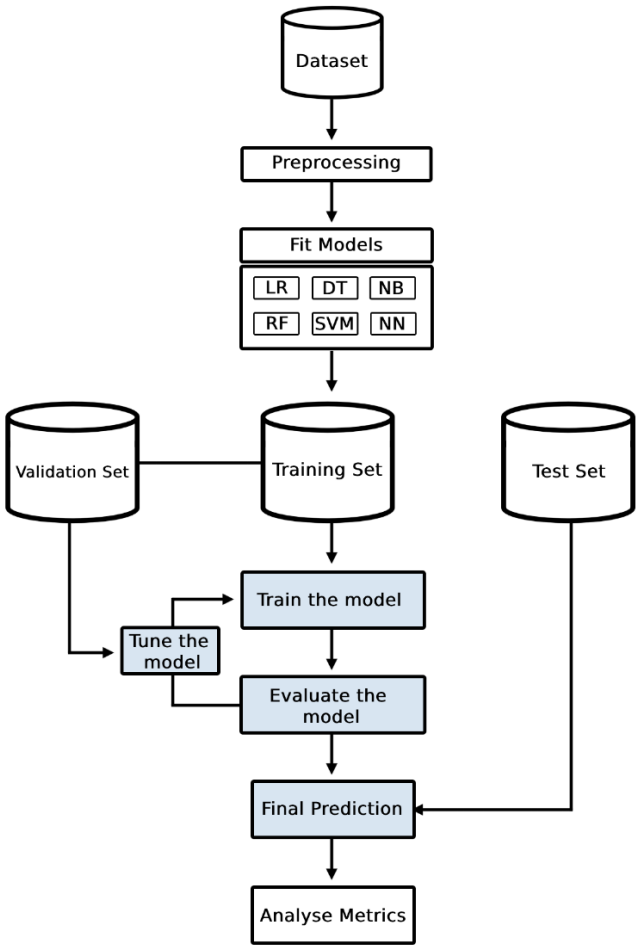

Figure 1. Proposed Study Framework

3.1 Step 1. Pre-Processing

3.1.1 Data cleaning and transformation

Once the data is imported into R, it will be explored to understand the unique values in each column and the structures of the different variables. The data will then be checked for any errors or NULL values before data transformation. Transforming the data includes renaming columns to be shorter and easier to read. Boxplots of each variable will be created to assess the distributions and check for any outliers in the dataset. Identified outliers in the data will be removed or amended using feature

engineering techniques such as aggregation. The removal of outliers ensures the models are training accurately on relevant data points to learn the underlying patterns in the data. Additionally, the removal of outliers is crucial as interpolation on these data points can increase the overall number of outliers in the data which can lead to skewed results and overfitting, further reducing model performance. The variables containing character data types will be replaced with numeric data and label encoding will be implemented. The predictors in the dataset will then be normalised to ensure the variables are on a comparable scale to reduce any chance of biases influencing model performance.

3.1.2 Feature selection

Feature selection will be conducted through a hybrid approach by combining the filter method of Pearson correlation and the embedded method of random forest variable importance. Feature selection is an important step of preprocessing to ensure the data being used for predictions has significance in predicting the target variable. A Pearson correlation matrix is used to further understand the relationship between individual variables in the data. This method measures correlation on a scale of 1 to -1 with correlation numbers > 0.5 having a strong positive correlation and correlation numbers < -0.5 having a strong negative correlation between the features. Multicollinearity exists when independent variables have a linear relationship between each other (Alin, 2010). Although this is commonly an issue to be addressed in regression tasks, studies have shown multicollinearity to have negative effects on algorithms used in classification and may lead to overfitting (Kiang, 2003). The variables with high correlation will be removed to reduce multicollinearity.

Furthermore, a concise dataset with less features helps to reduce model complexity and improves model performance and comprehensibility.

Once the variables with high correlation are removed, variable importance through random forest will be employed to understand which variables have a significant impact on the target variable of M1 laptop purchases. A variable importance plot will visualise the variables with the most significant impact on the accuracy and Gini impurity of the model. Mean decrease accuracy shows the impact of different features on improving overall model accuracy, with a higher score representing the highest importance. Mean decrease Gini shows which features are most effective at reducing model impurity when that feature is used to split nodes in the decision tree.

3.1.3 Data preparation

Once the data is ready, the *rsample* package will be utilised to split the data into a training and test set. The train/test partition will be a 75% (training) 25% (test) split. The *recipes* package will be used to specify the target variable in the data and add any further preprocessing steps such as dealing with any class imbalance issues in the dataset. Different approaches are available when dealing with an imbalanced dataset including under-sampling, over-sampling, SMOTE and ROSE (Liu, 2022). Up-sampling is useful when classes are imbalanced to reduce the bias of models that may favour the majority class. Once the recipe is completed, it will be added to a workflow ready to be fitted to different created machine learning models.

3.2 Step 2. Training

Cross validation will be implemented using 10-fold cross validation to create 10 different iterations of the training set. As a result, the models can be tested on the validation sets to assess model performance and implement model tuning to create

models with the highest performance. The *parsnip* package within tidymodels allows for quick and seamless creation of multiple machine learning models to be fit to the data. The models to be used are logistic regression, decision tree, random forest, naïve bayes, support vector machines and neural networks.

Logistic regression

Logistic regression is one of the most used algorithms in business for predicting customer churn, sales, and events with a dichotomous outcome (Yale et al., 2017). It is a traditional statistical analysis method that is used to understand inter-relationships between variables (Ozdemir & Turanli, 2021). Logistic regression (1) estimates the probability of an outcome based on a non-linear sigmoid function and presents the results as logits and odds ratios rather than probabilities (Hagger-Johnson, 2014). Transforming probability to log odds allows the range restriction issue in the model to be eliminated by using the range of $(-\infty, +\infty)$ as opposed to (0,1). This method uses maximum likelihood estimation on the transformed logit dependent variable with respect to the independent variables. As a result, logistic regression can estimate the probability of the classification events occurring (Dangeti, 2017).

$$ln(odds) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \quad (1)$$

$ln(odds)$= The log odds/logit

x = The value of the predictor variable

β_0 = The intercept

β_1 = The slope

p = the probability of the outcome occurring

Consequently, logistic regression serves as a commonly used model for classifying categorial outcomes as linear regression is not suitable. Linear regression assumes the residuals are normally distributed, in the case of binary outcomes, this assumption is not viable rendering linear regression ineffective for classification (Hagger-Johnson, 2014). This method is fast, simple and can produce accurate results on datasets of all different sizes. Although logistic regression is an established algorithm, the algorithm assumes the predictor variables are independent from each other in their relationship to the target outcome. This assumption is not always reflective of the real business environment (Yale et al., 2017). Additionally, this method can struggle when dealing with an imbalanced dataset as this can produce biased results favouring the majority class over the minority class. Classification of predicting purchase intentions utilises binary logistic regression as two categories are used; whether customers are likely to buy a product or not (Ozdemir & Turanli, 2021). Logistic regression has only a few parameters that can be changed however, it will be utilised as it is a powerful model capable of generating accurate results and is easy to implement.

Decision Tree

A decision tree is a flowchart-like structure that recursively splits data until the model is left with decision outputs known as leaf nodes. Within a decision tree, the top node is known as the root node and each branch off that node represents a decision rule with each leaf node representing the classification result.

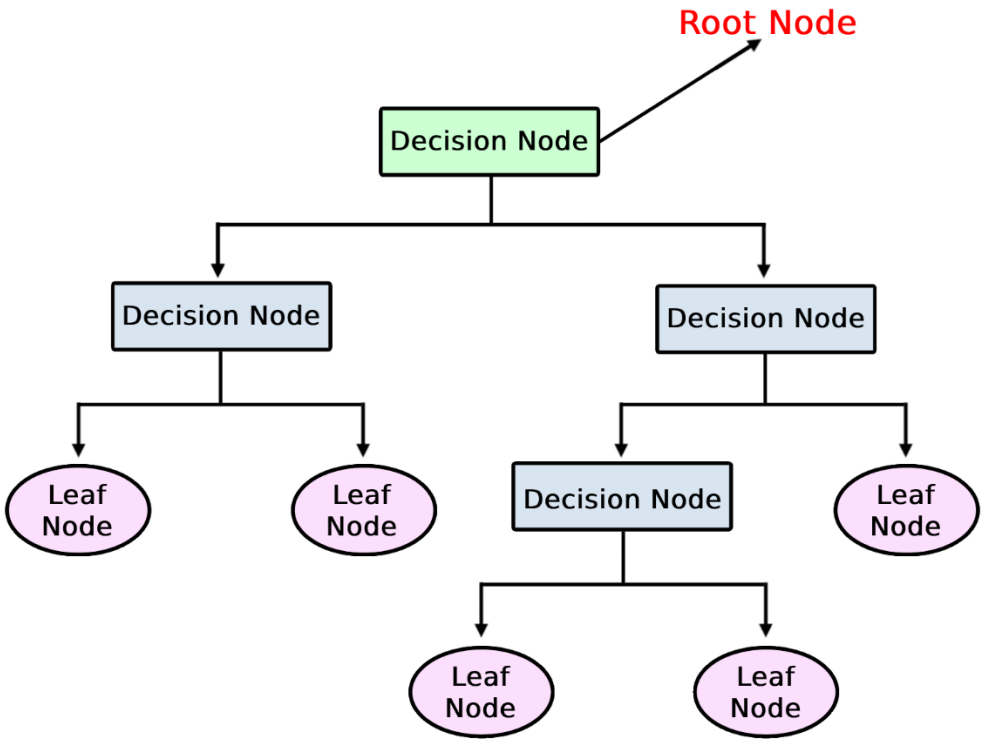

Figure 2. Decision Tree Diagram

Decision tree is a popular classification tool in identifying a chain of events that are most likely to lead to an outcome (Bing & Yuliang, 2016). Decision trees are quick to train and work well when there is non-linearity between the target and predictor variables. Additionally, this method can generate calculations using categorical and numerical predictor variables. However, longer training times are among some of the drawbacks of implementing decision trees depending on the size of the data.

Moreover, decision trees are high variance models and minor changes in the data such as feature selection can have a significant impact on the results of the model. Furthermore, this method is prone to overfitting nonetheless, multiple pruning methods are available to combat this problem (Yale et al., 2017). The decision tree produces a visual representation of decisions leading to the target variable that are easy to interpret, making it a fundamental model to implement.

Random Forest

Random forest is another supervised learning algorithm that can be used for classification tasks. This algorithm is like a decision tree however, this method creates multiple correlated decision trees. The final output is based on combining the results from multiple related decision trees based on majority and each tree neutralises the errors of other trees (Brieman, 1999; Ho, 1995). Multiple empirical studies have found Random forests to be the most accurate classification methods. Additionally, random forest works well with a large dataset consisting of multiple variables (Yale et al., 2017). Furthermore, random forest provides feature importance to explain the impact of features on the target variable. This method also works well with noisy data and outliers.

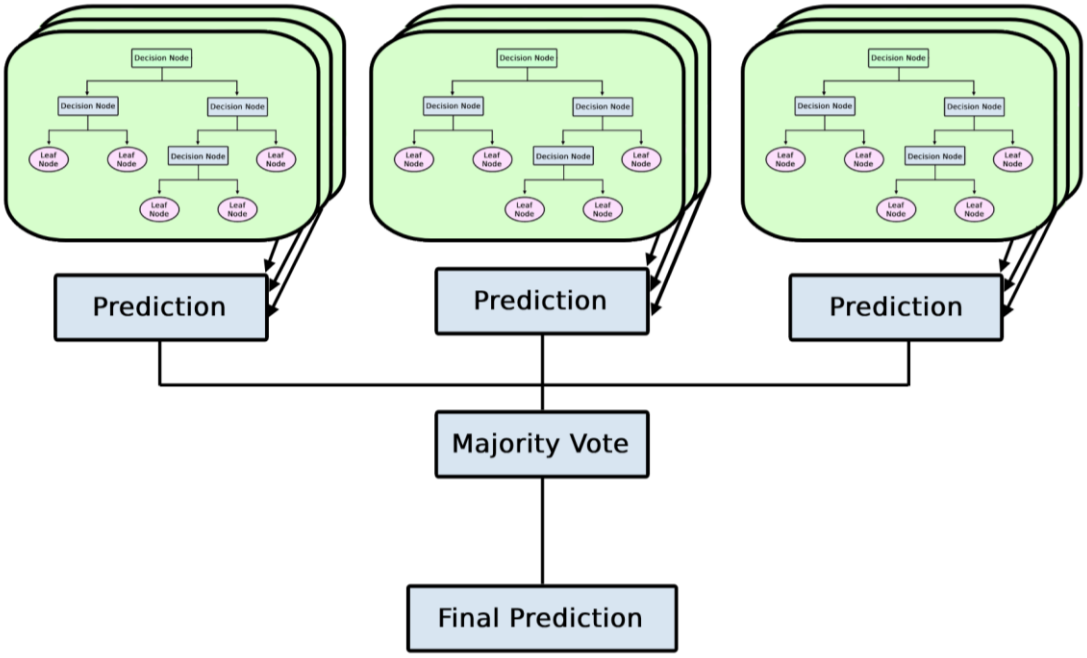

Figure 3. Random Forest Architecture

Despite the method's superior classification abilities, the computation of multiple decision trees leads to a slower training time, which can be detrimental when working with a large dataset in a time-limited environment. Moreover, this method is also prone to overfitting if the number of trees is too large. In addition, the black box problem exists with this method as multiple decision trees are created and there is difficulty understanding exactly how the model arrived at a specific decision. Nevertheless, the random forest ensemble method allows for the number of created decision trees to be specified for model training and testing. This value will be set to an optimal number of trees to reduce overfitting and any bias in the data that could occur if too many trees are used.

Naïve Bayes

The naïve bayes method of classification is a commonly used algorithm which uses probability to model the relationship between attributes and variables (Bing & Yuliang, 2016). This method uses bayes theorem (2) and estimates the conditional probability of the class the data belongs to.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (2)$$

$\mathrm{P(A|B)} = \mathrm{The\ probability\ of\ A\ being\ true\ given\ that\ B\ is\ true\ (posterior\ probability)}$

$\mathrm{P(A)} = \mathrm{The\ probability\ of\ A\ being\ true\ (prior\ probability)}$

$\mathrm{P(B)} = \mathrm{The\ probability\ of\ B\ occurring\ being\ true}$

$\mathrm{P(B|A)} = \mathrm{The\ probability\ of\ B\ being\ true\ given\ that\ A\ is\ true\ (likelihood)}$

$$P(C_k|x) = \frac{P(x|C_k) * P(C_k)}{P(x)} \quad (3)$$

$$P(C_1|x_1 \cap x_2 \cap x_3 \cap x_4) = \frac{P(x_1 \cap x_2 \cap x_3 \cap x_4|C_1) * P(C_1)}{P(x_1 \cap x_2 \cap x_3 \cap x_4)}$$

$C_k = \mathrm{Class\ label}$

$x = \mathrm{Feature}$

In this case of classification (3), the conditional probability of each class label of purchaser C_1 and non-purchaser C_0 can be calculated based on the intersection of feature(s) x. The intersection is used as naïve bayes calculates the conditional

probability with the assumption of all features being present simultaneously. Additionally, this model is fast to train and works well with categorical and numerical input variables. Naïve bayes only requires a small training dataset to identify the parameters needed for classification (Kohavi, 1996). Gaussian naïve bayes works well with variables of different distributions in classification tasks (Parihar & Yadav, 2022). Although naïve bayes is very effective even with highly dimensional data, it assumes independence for all variables which limits its ability to learn interrelationships between variables (Yale et al., 2017). Independence between input variables is also an assumption that is rarely accurate in real-world applications. Nonetheless, naïve bayes is capable of quickly producing optimal results with data of different sizes, making it a fundamental algorithm to be implemented.

Support Vector Machines

Support vector machines are a kernel-based algorithm commonly used for classification tasks (Cervantes et. al, 2020). Support vector machine perform classification by creating hyperplanes/decision boundaries to determine the best separation between two classes, with the closest samples to the hyperplanes being the support vectors (Vapnik, 2000). The principle of risk minimisation is utilised when determining the boundary distance between the two hyperplanes (Ozdemir & Turanli, 2021). Support vector machines can handle non-linear data using kernels, making this method an optimal choice for classification tasks (Parihar & Yadav, 2022). This method less like to overfit and is highly effective with highly dimensional data that has a clear margin of separation between classes.

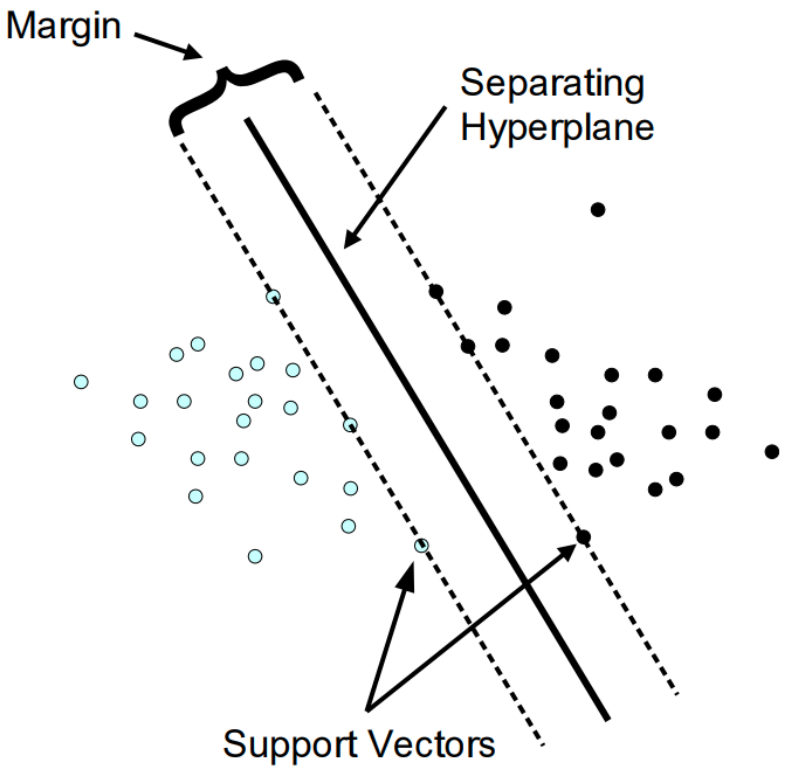

Figure 4. Support Vector Machine Architecture (Meyer & Wein, 2015)

Empirical studies have shown the support vector machine method of classification to be highly accurate with fewer calculations needed (Ozdemir & Turanli, 2021). However, Support vector machines have shown to not perform well with large datasets as they can be computationally expensive with longer training times when scaled to larger data. A polynomial and radial kernel support vector machine will be used to assess which of the two support vector machines would give the most accurate results. Radial and polynomials kernels were chosen over a linear SVM due to the data having non-linear relationships to the target variable.

Neural Network

Neural networks are a classification algorithm that is influenced by the natural structures of human biological neural networks (Noviantoro & Huang, 2021). Neural networks adopt the brain's abilities and as a result, neural network algorithms excel at recognising patterns in data and can create complex relationships between inputs and outputs (Noviantoro & Huang, 2021; Wilamowski, 2011). A Multilayer perceptron is a feedforward artificial neural network and consists of an input layer, hidden layer, and an output layer. These layers are connected and are linked by different weights. The nodes in the hidden layer able to model non-linear relationships between the input and hidden nodes, and between the hidden layer and output nodes (Yale et al., 2017). Each node in the hidden layer performs a weighted sum calculation from the previous layer and applies an activation function. Linear and non-linear activation functions exist however, in the case of classification, a sigmoid non-linear function is optimal as this introduces non-linearity and allows for complex relationships to be understood in the data. Once the activation function has been applied in the hidden layer, the sum is passed to the output layer and the final classification result is produced. Different methods exist to train neural networks and back-propagation is a commonly used algorithm to train feed-forward neural networks, and this involves adjusting the weights of inputs in the model by comparing them with the actual values in the training set to reduce output errors and optimise model performance (Suchacka & Stemplewski, 2017).

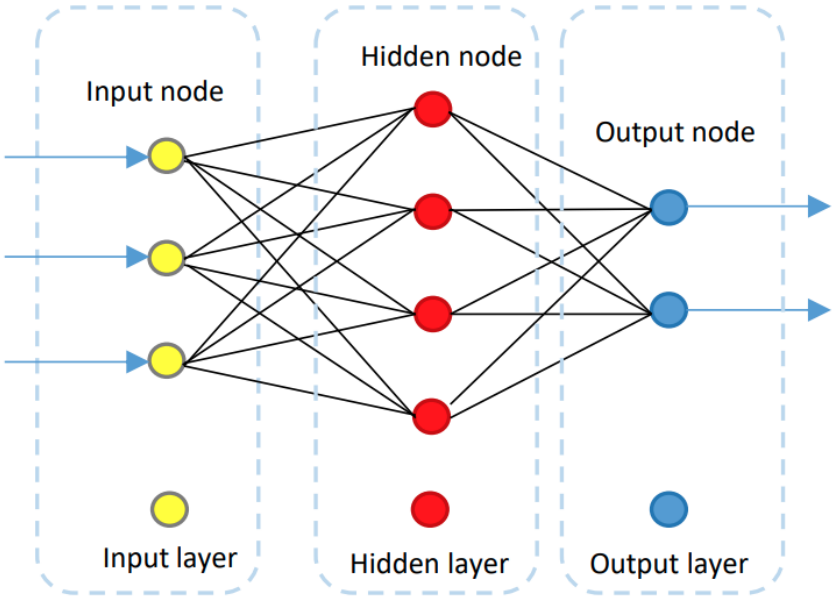

Figure 5. Neural Network Architecture (Noviantoro & Huang, 2021)

On the other hand, neural networks are sensitive to noise in the dataset and are prone to overfitting. They are computationally expensive and can suffer from the black box problem, leading to a lack of understanding of how the model arrived at specific decisions or which features it used for making predictions. Alternatively, explainable machine learning methods such as SHAP and a decision tree algorithm can be trained to identify the predictions of the neural network to eliminate the black box problem (Yale et al., 2017). In this research, a multilayer perceptron with a softmax activation function will be utilised on the dataset.

All computations will be performed on an Intel Core i7 CPU (3.3GHz) with a 16 GB memory.

3.3 Step 3. Evaluation

Once the models are tuned to give the highest accuracy on the cross-validation data, the models will fit to the testing data to give a final prediction of how accurate the models are. The last_fit function fits the best model to the training set and the evaluates its performance on new observations in the test set. The last_fit function takes the original data split as the function needs to understand the training and testing data. The model is then fit and trained on the training set once and then tested to give the final metrics for each model.

The *yardstick* package will be implemented to gain summary statistics of the models. A random forest variable importance plot will be utilised on the test set to visualise which attributes are the most significant when predicting whether people will purchase the Apple M1 laptop. To measure the different model's performances, multiple evaluation metrics will be utilised and compared to gain a deeper understanding of how well the models can predict purchase intentions. To measure performance a confusion matrix will be utilised to calculate model accuracy, sensitivity, specificity, Cohen's kappa and F1 score. Additionally, the AUC from the ROC will also be measured to understand model performance.

AUC and ROC curves are visual representations showcasing the rate which the model correctly distinguishes between the two classes of individuals who did and did not make a purchase. ROC curves examine the true positive rate against the false positive rate allowing us to evaluate models at different thresholds. Ideally, a ROC curve should be near the top left corner indicating a high level of true positives to false positives.

Table 1. Confusion Matrix

		Actual Values: Positive	Actual Values: Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
Predicted Values	Negative	False Negative (FN)	True Negative (TN)

A confusion matrix is commonly used to visualise the performance of machine learning models in classification tasks (Han et al., 2012). In this case, a positive prediction represents the participants who made a purchase and vice versa.

- True Positive (TP) represents the values belonging to the positive class that were correctly predicted by the classifier.
- True Negative (TN) represents the values belonging to the negative class that were correctly predicted by the classifier.
- False Positive (FP) represents the values belonging to the negative class that were incorrectly predicted as positive by the classifier.
- False Negative (FN) represents the values belonging to the negative class that were incorrectly predicted as positive by the classifier.

From the confusion matrix, the following equations can be calculated:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Accuracy measures the percentage of correctly predicted cases across all predictions.

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN} \quad (5)$$

Sensitivity (Recall or True Positive Rate) measures the proportion of correctly predicted positive cases out of all actual positive cases. This measure reflects the model's ability to identify positive cases.

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

Specificity (True Negative Rate) measures the proportion of correctly predicted negative cases out of all actual negative cases. This measure represents the model's ability to identify negative cases.

$$F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP+FN)} \quad (7)$$

F1 score provides a balanced measure of the model's performance.

$$(Cohen's\ Kappa)\ k = \frac{2*(TP*TN-FN*FP)}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)} \quad (8)$$

Cohens kappa statistic is considered a more robust method of model evaluation compared to AUC as it measures the model's agreement between the actual and predicted classes on a scale of -1 to +1 and considers the possibility of the agreement occurring by chance (Trivedi et al., 2022).

4 Results

The data was imported into R and the data was explored. The data was split into different data types including categorial variables for gender, income group, age, status etc. Different numeric variables were also included in the data such as variables measuring the individual's importance of different laptop traits (rated from one to five). The data can be categorised in three main sections: demographic, product specific, and general laptop questions.

Table 2. Transformed Data Summary

Variable	Data Type	Min	Max	Mean	Standard Dev.
User_pcmac	Categorial	0	1	-	-
Trust_apple	Categorial	0	1	-	-
Status	Categorial	0	1	-	-
Gender	Categorial	0	1	-	-
Domain	Categorial	0	2	-	-
Familiarity_m1	Categorial	0	1	-	-
Apple_count	Numerical	0	8	2.60	1.89
Age_computer	Numerical	0	9	2.82	2.44
Age	Numerical	18	60	27.78	9.24
Income	Numerical	0	75000	23063.91	26638.75
Interest_computers	Numerical	2	5	3.81	0.96
f_battery	Numerical	1	5	4.52	0.72
f_multi	Numerical	2	5	4.12	0.79
f_performance	Numerical	2	5	4.39	0.76
f_perloss	Numerical	1	5	3.37	1.12
f_size	Numerical	1	5	3.15	1.16
f_noise	Numerical	1	5	3.72	1.12
f_synergy	Numerical	1	5	3.46	1.27
f_price	Numerical	1	5	3.87	0.99
f_neural	Numerical	1	5	3.16	1.14
M1_consideration	Numerical	1	5	3.60	1.24
M1_purchase	Categorial	0	1	-	-

4.1 Data cleaning and transformation

Data cleaning and transformation were implemented to identify any errors/outliers in the data and to convert datatypes. No NULL values were found however, a spelling error in the status column was identified and corrected. Boxplots were created to visualise the distributions of the variables in the data. As a result, some outliers were identified in the data, and they were removed through aggregation by downsizing variables with multiple levels. For example, the variable domain showcased outliers as it contained 22 different levels describing the participant's occupational background, the levels were condensed down to three levels: Science & IT, Business & Legal and Humanities. Character variables were replaced with numeric data and label encoding was implemented. All categorial variables with multiple levels were aggregated to improve model performance. m1_purchase was converted to a factor as it is the target variable. The predictors were then normalised to ensure all variable were on comparable scale.

4.2 Exploratory analysis

Once data cleaning and transformation was completed, visualisations were created to further understand the relationship between the different variables and how they affect purchase intentions. From the visualisations created in RStudio, various interpretations could be made from data presented.

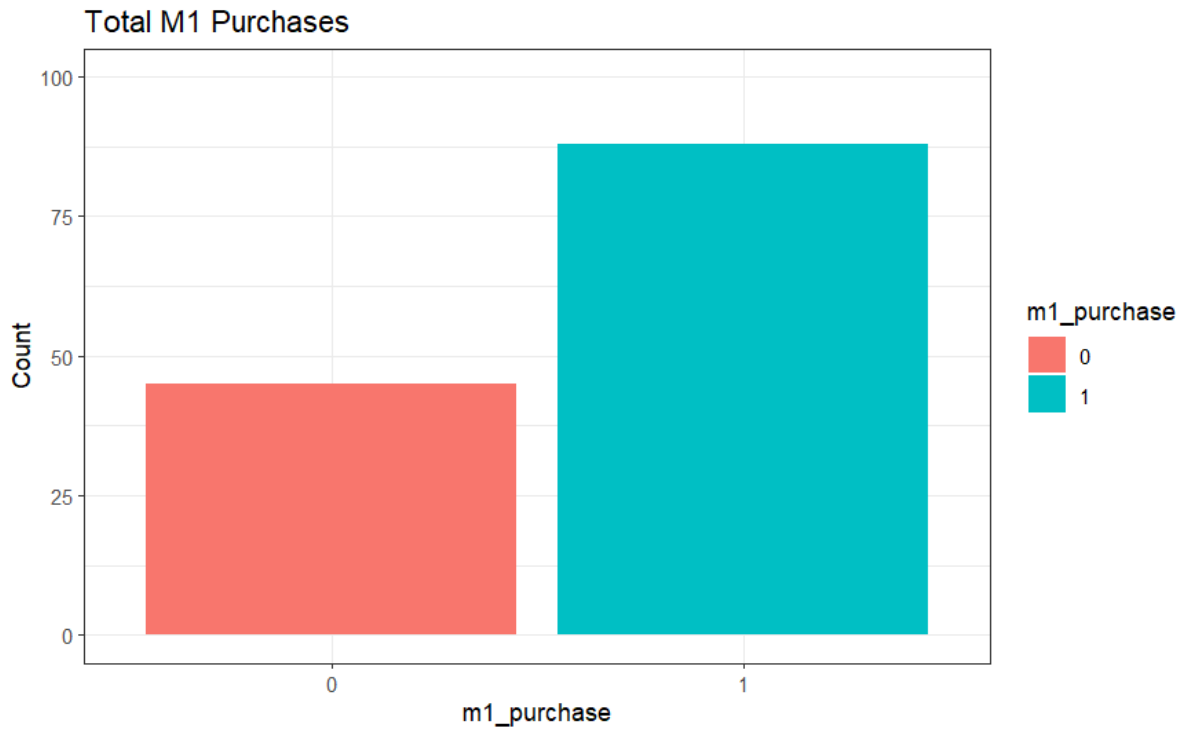

Figure 6. Total Purchasers

Firstly, a visualisation of the target variable m1_purchase was created, and the results showed that 88 of the 133 participants chose to purchase the M1 laptop. This also showcased a class imbalance of the target variable which will need to be addressed before models are fit to the data.

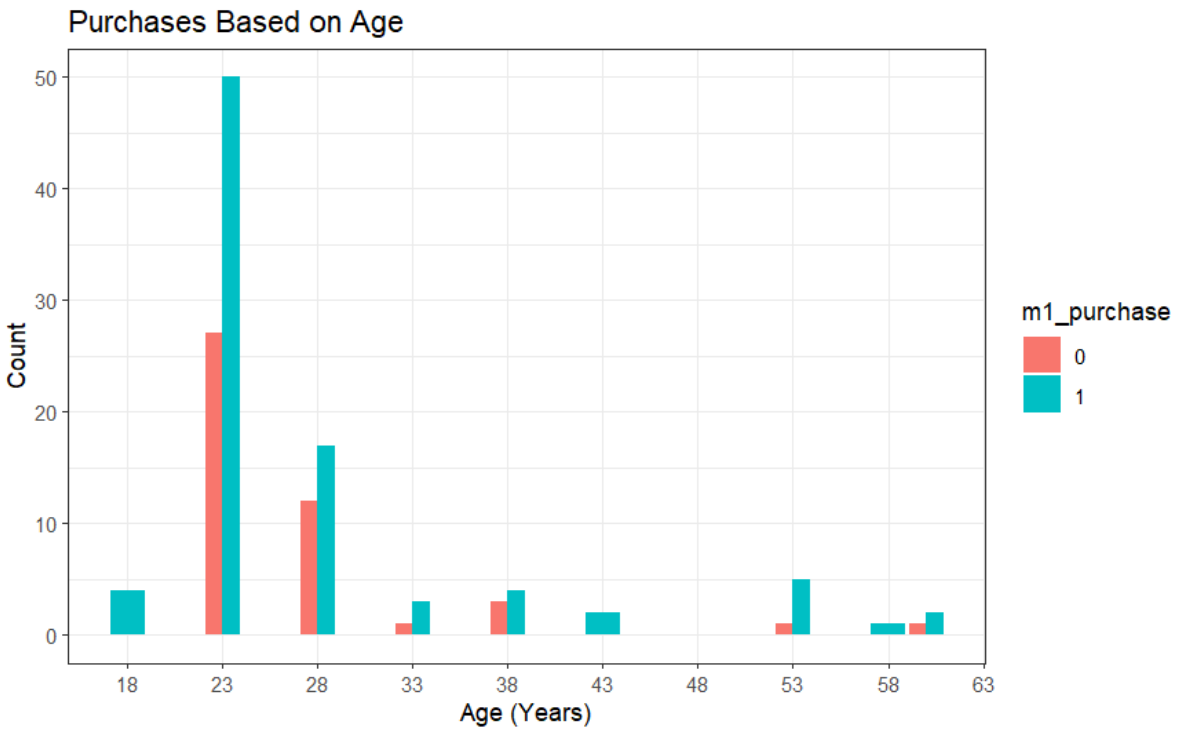

Figure 7. Purchasers by Age

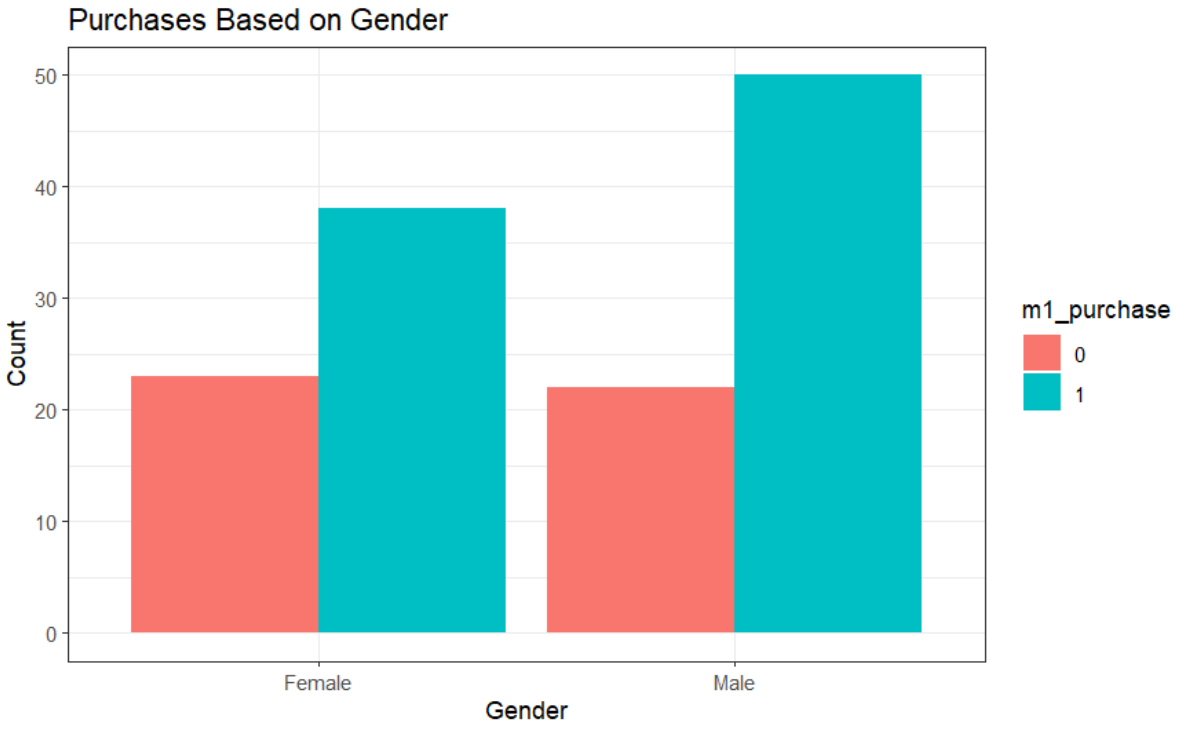

Figure 8. Purchasers by Gender

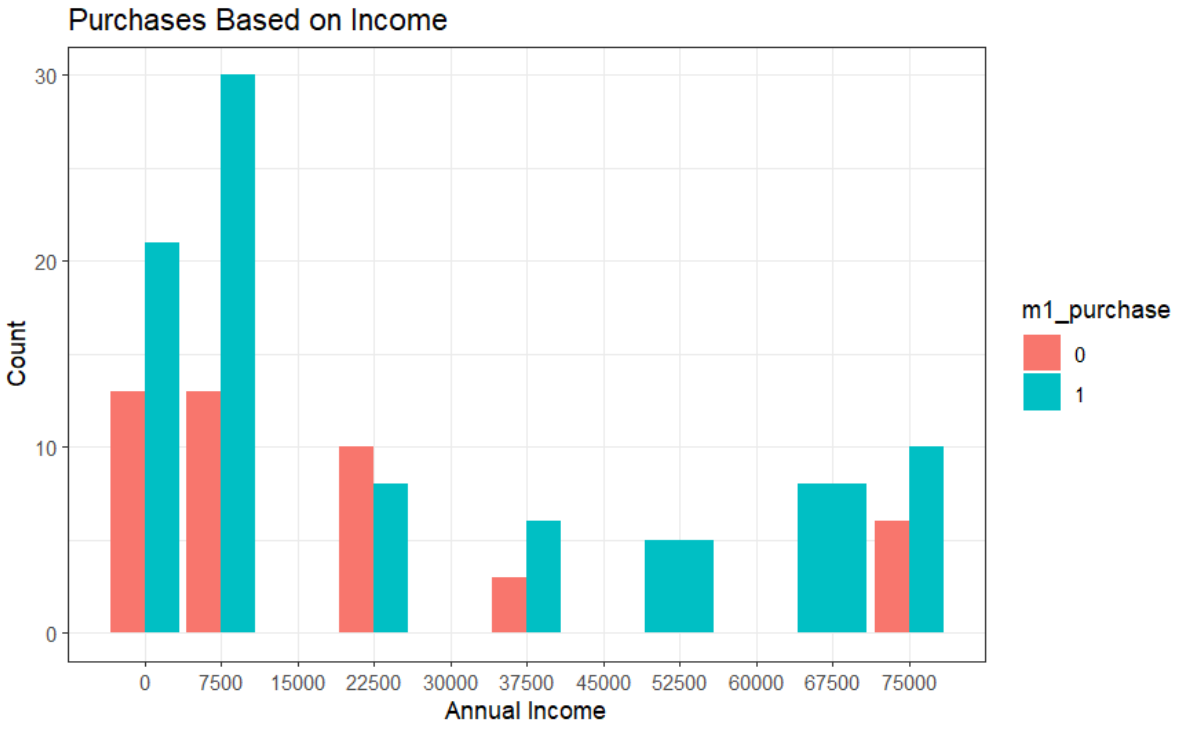

Figure 9. Purchasers by Income

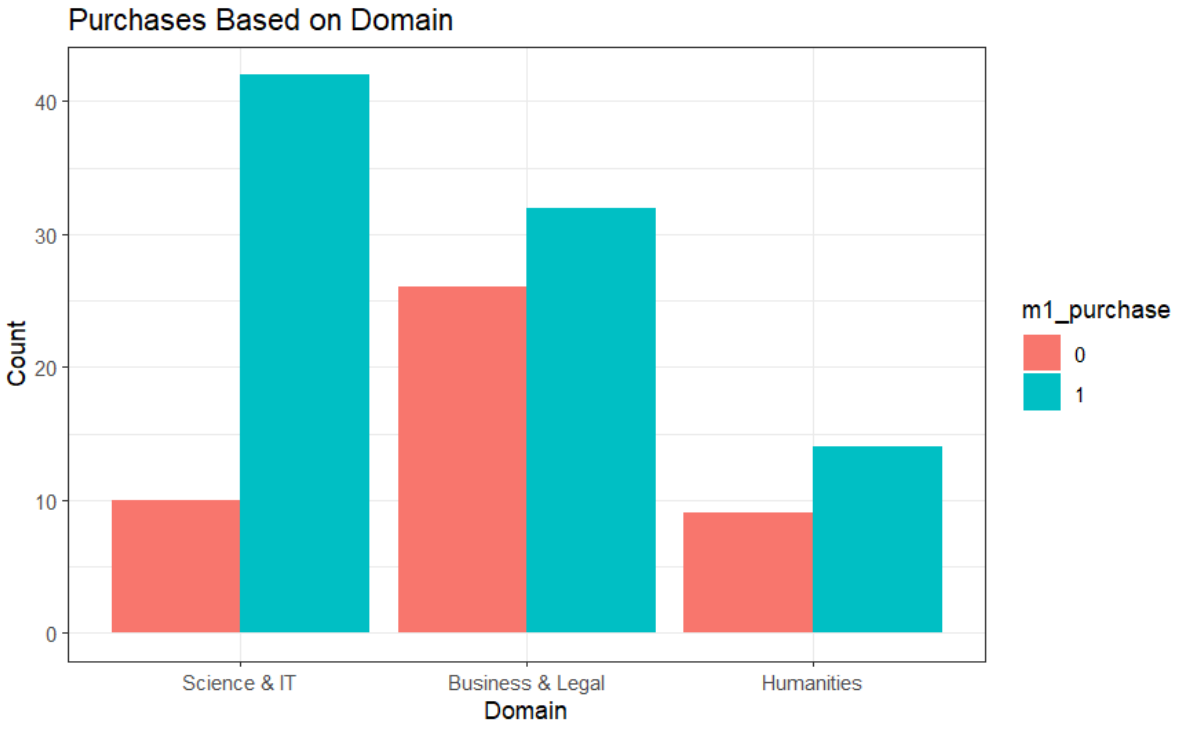

Figure 10. Purchasers by Domain

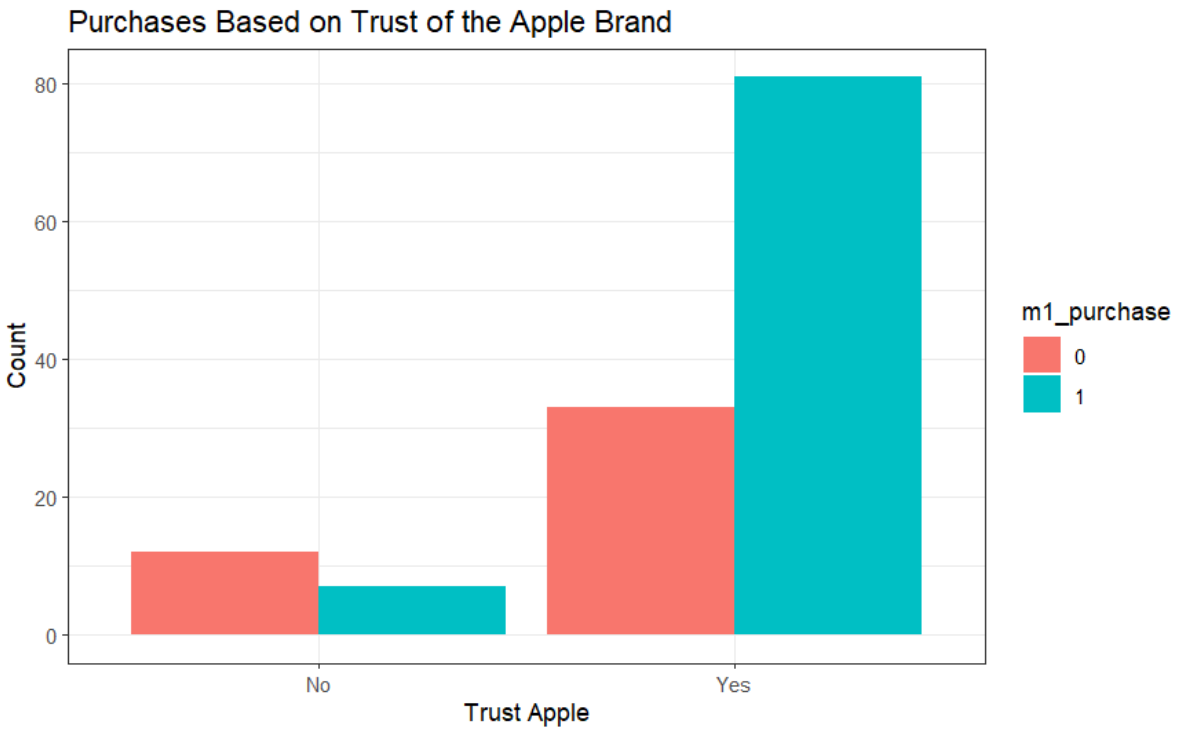

Figure 11. Participant's Trust in the Apple Brand

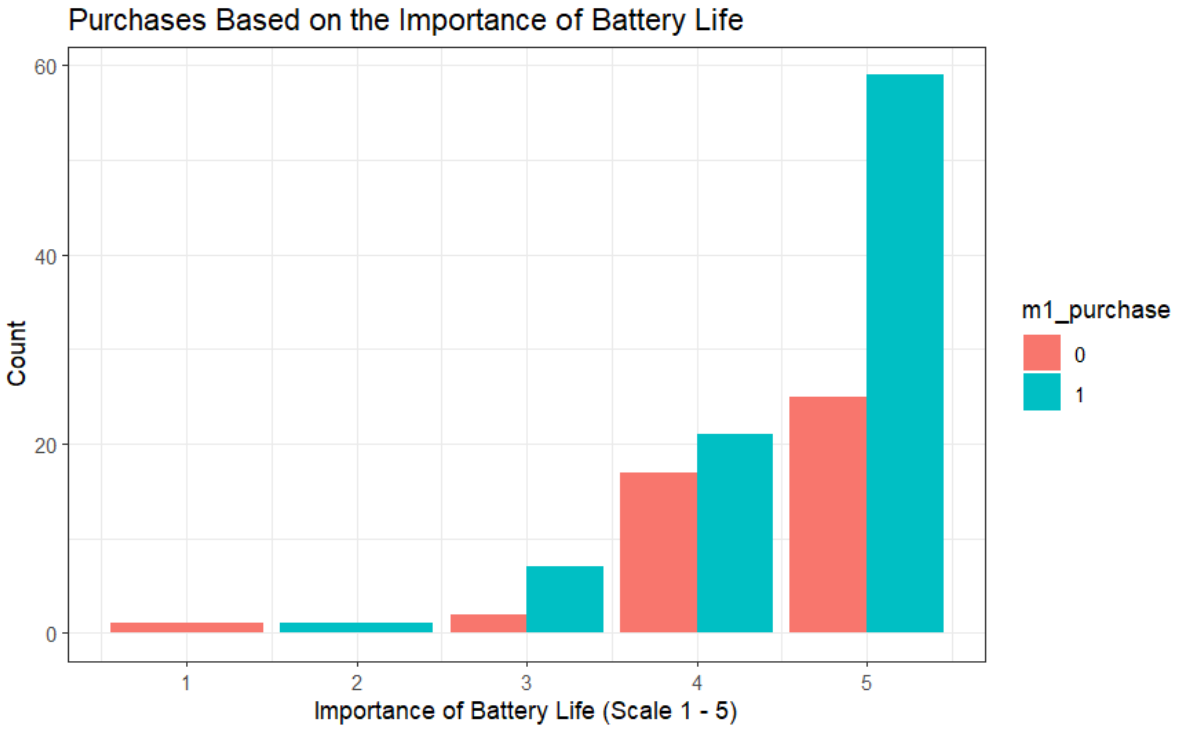

Figure 12. Participant's Importance of Battery Life

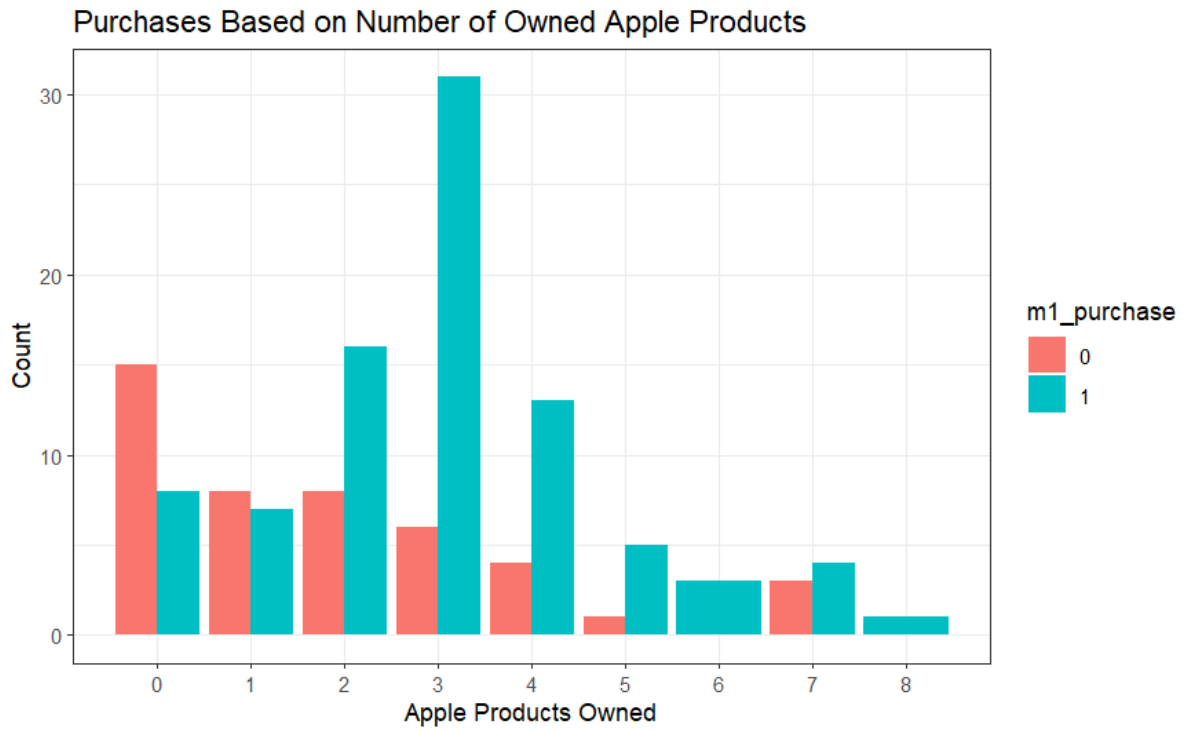

Figure 13. No. of Apple Products Owned by Participants

The visualisation of age showed 23-year-olds to be a significant proportion of the demographic choosing to purchase the M1 laptop with a decrease in purchases as age increases. Of the individuals who did choose to purchase the laptop, males were 57% of the total purchasers while only 43% were female. Participants earning £7,500 annually were shown to have the highest purchase intentions of the M1 laptop. The participants that showed to have the highest purchase intentions were part of the science & IT and business & legal domains. Trust in the Apple brand showed to have a significant influence on purchase intentions and showed to be on the main driving factors influencing purchase intentions. It should be noted that nearly 10% of individuals who did not trust the Apple brand still chose to purchase the laptop. The importance of different laptop features showed the laptop's battery life had a significant impact on influencing purchase intentions. Consequently, most individuals

rating this feature as very important had the highest number of purchases for the laptop. On the other hand, features including the neural engine, performance loss and size of the laptop were rated as moderately important by participants. The number of Apple products owned showed to have an impact on purchases. Most purchasers owned two to four Apple products with the highest number of non-purchasing participants owning no Apple products. The group owning two to four products were mainly students and individuals in employment. In addition, the participants currently owning an Apple computer showed to have a significant impact on purchase intentions as 77% of total purchasers own an Apple computer.

Overall, 23-year-olds with an income of £7500 had the highest purchase intentions for the M1 laptops and most of them were students in the science & IT and business & legal domains. Exploratory analysis of the data showed that most participants not making purchases of the laptop came from participants who did not own any Apple products and did not trust the Apple brand. Furthermore, participants outside of the science & IT, business & legal domains showed lower purchase intentions for the laptop.

4.3 Feature selection

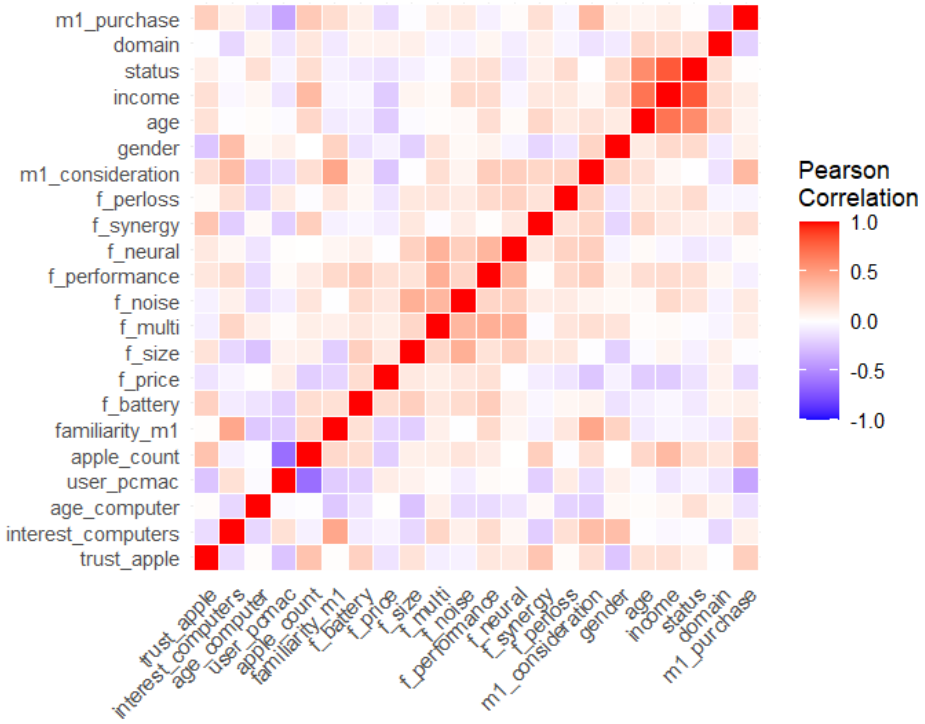

Figure 14. Correlation Matrix

Before fitting models to the data, feature selection was conducted through Pearson correlation and random forest. This data has many features, and the correlation matrix will assist in feature selection by identifying the closely related variables that provide no further benefit in this research. From the correlation matrix we can see that age has a strong positive correlation on status and income. The number of Apple products the participants own is shown to have a strong negative correlation with whether the participants own a PC or Mac. As a result, age, income, and status were removed from the data due to their high correlation. Despite their high negative correlation, the count of participant's Apple products and the brand of their computer was kept in the study to further understand the impact of the Apple brand on purchase intentions.

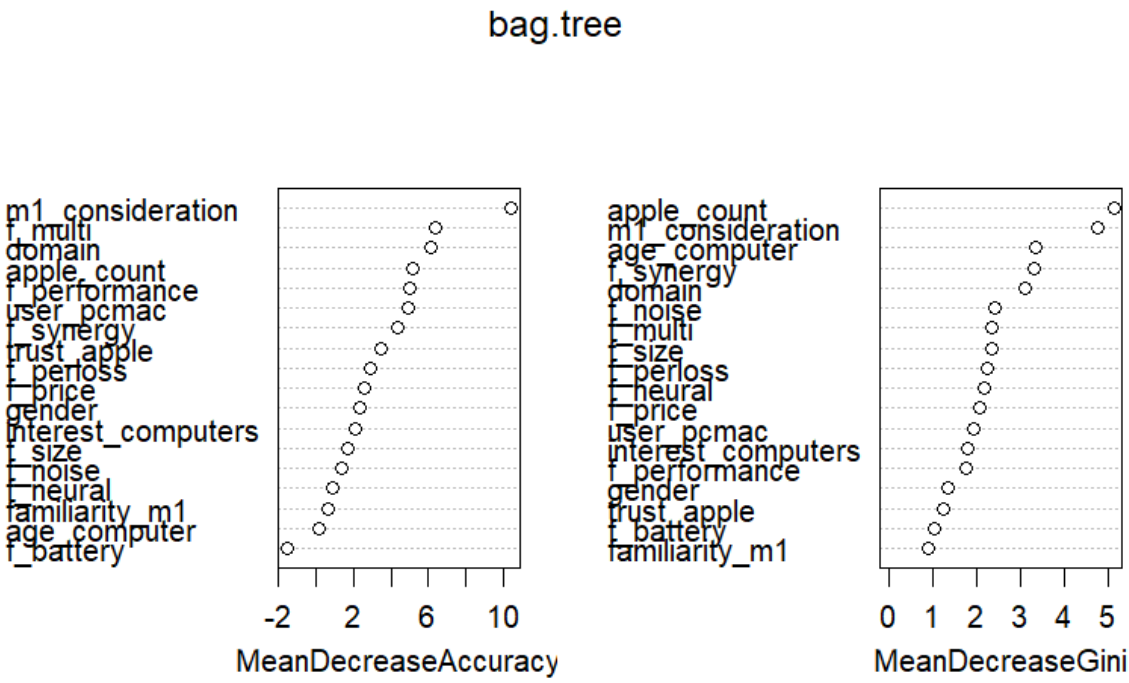

Figure 15. Accuracy and Impurity Variable Importance Plot

The random forest variable importance was used to visualise the variable impact on the target class. As a result, the 8 features: domain, user_pcmac, trust_apple, m1_consideration, apple_count, age_computer, f_synergy and f_multi showed to have high importance and were chosen as the main predictor variables for the main data set.

4.4 Data preparation

A training and test set were created using the rsample package. The initial_split function was used on the dataset of 133 instances to divide the data using a 3/4 split (75% training, 25% test). The split was stratified on the m1_purchase column to ensure that the training and test set had the same proportion of individuals who did and did not make a purchase of the M1 laptop. The training set contained 99 instances while test set contained 34 instances.

The recipes package was used to create a new recipe where m1_purchase was identified as the target class and the training set was used as the main data. Additionally, up-sampling on the m1_purchase target class was added to the recipe through SMOTE to deal with the class imbalance between purchasers and non-purchasers in the data. Previous studies working with imbalanced datasets found SMOTE to be the most effective technique when compared to the other methods (Wong & Marikannan, 2020). SMOTE was able to balance the classes by generating synthetic observations by extrapolating the minority class to improve representation of the non-purchaser's class. Once the recipe was completed, it was added to a workflow ready to be fitted with a machine learning model.

4.5 Training

As the training set is small for comparing and tuning models, resampled datasets were created to validate the performance on the models. Cross validation was implemented using 10-fold cross validation to create 10 different iterations of the training set. Consequently, the different created models can be fit to the cross-validation data sets to compare evaluation metrics and tune the models before implementing them on the final test set. The parsnip package within tidymodels was used to create of multiple machine learning models to be fit to the data. This step was used to define the model specification and set any model parameters before the models were fit to the data. All models were set to the "classification" mode, in line with the research objective. The models and engines used are as follows.

Table 3. Implemented Models and Engines

Model	Engine
Logistic Regression	glm
Decision Tree	rpart
Random Forest	ranger
Naïve Bayes	naivebayes
SVM (Polynomial)	kernlab
SVM (Radial)	kernlab
Neural Network	nnet

The models were trained on the cross-validation dataset and tuned accordingly to improve model performance. The tune function was used to find the optimal parameters for different models. To understand the benefits of the tuning the hyperparameters, a normal decision tree and a tuned decision trees' performance on the training data were compared and the results showed the tuned decision tree to be more accurate and had a higher AUC score. This was also evident when comparing the performance of a normal random forest to a tuned random forest.

Once the models were tuned for optimal performance, their ability to make accurate predictions were evaluated on the withheld data in the testing set.

Table 4. Model Performances

Evaluation Metric	LR	DT	RF	NB	SVM (R)	SVM (P)	NN
Accuracy	0.71	0.74	0.76	0.74	0.74	0.72	0.76
AUC	0.78	0.81	0.82	0.81	0.81	0.77	0.82
Sensitivity	0.71	0.51	0.57	0.57	0.57	0.68	0.75
Specificity	0.73	0.71	0.87	0.82	0.74	0.72	0.74
F1	0.57	0.60	0.64	0.56	0.52	0.58	0.64
Kappa	0.40	0.44	0.47	0.41	0.42	0.40	0.45

Random forest and neural network performed optimally when exposed to new data.

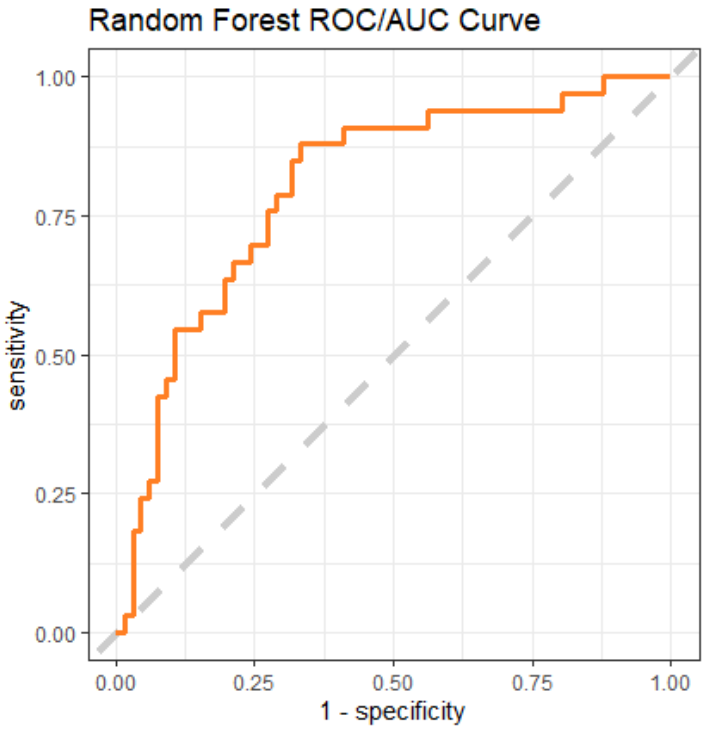

Figure 16. Random Forest ROC Curve

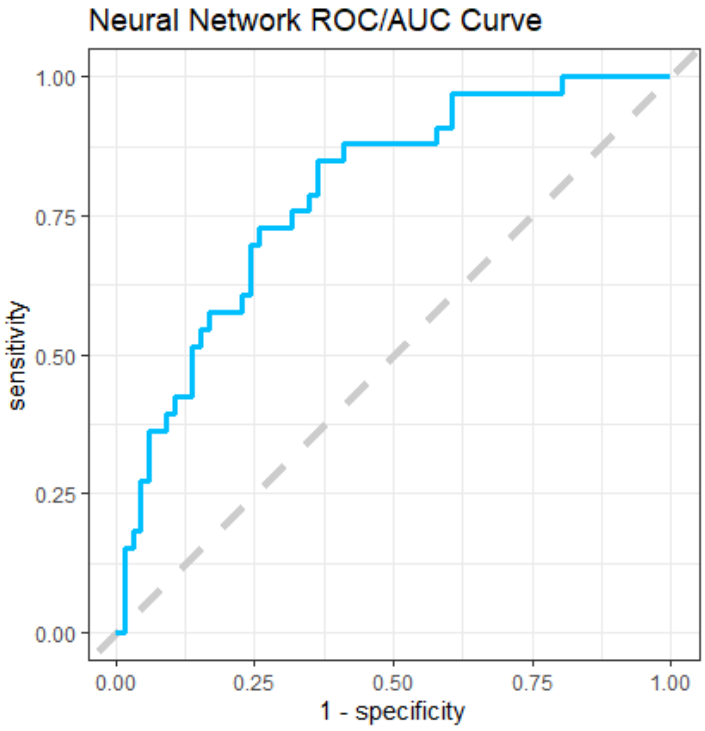

Figure 17. Neural Network ROC Curve

A variable importance plot was created on the final testing set to identify which features have the highest impact on influencing purchase intentions.

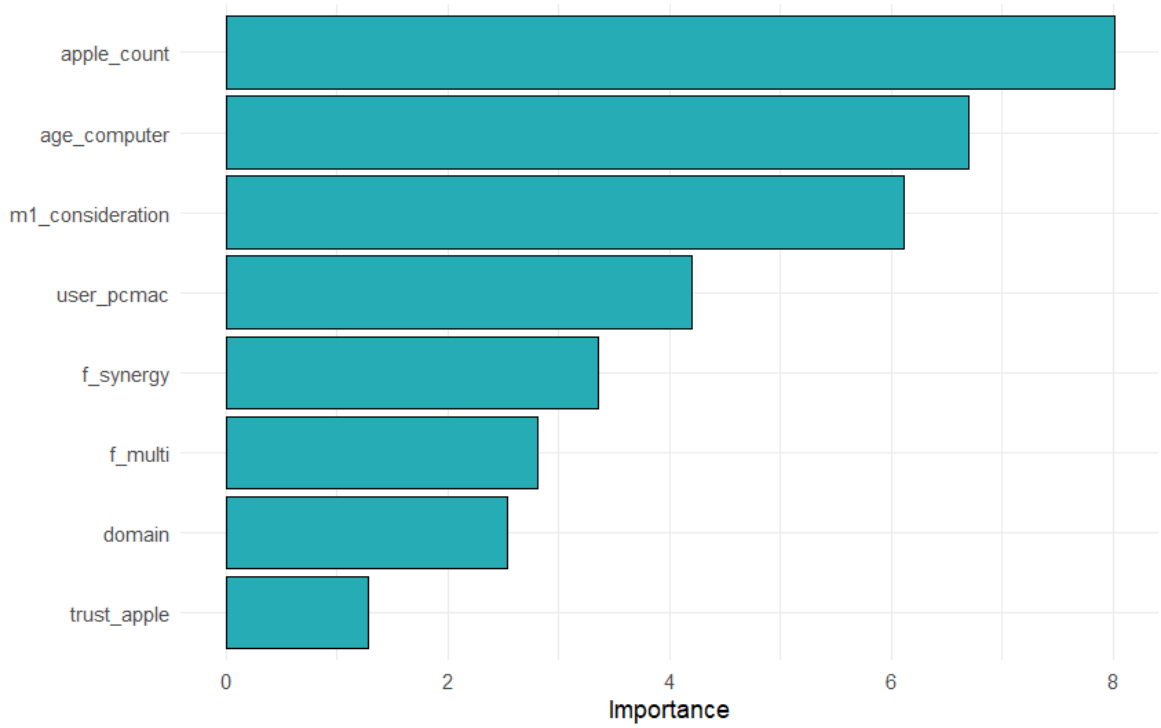

Figure 18. Variable Importance Plot

The variable importance plot found the number of Apple products the participants own, the age of their current computer and the importance of the M1 laptop chip to be the top three most important variables influencing purchase intentions.

4.6 Bayesian model

As the dataset contains a small number of instances, the evaluation metrics of different models may not produce as accurate results as possible. Alleviating this issue is possible through implementing a Bayesian statistical model in R.

Using the glm function in R and utilising a logit link, the predictor variable coefficients can be calculated. The participant's current brand of computer, domain, and perception of the M1 chips' importance showed to have a significant influence on purchase intentions with all parameters having a p-value < 0.05. Keeping the other variables constant, the results showed that participants who considered Apple's M1 computer chip to be very important were 1.7 times more likely to purchase the laptop than participants who scored the M1 chip as not important. For every participant who did not own an Apple computer, the odds of them purchasing the M1 laptop decreased by a factor of 0.20 (at a credible interval of 0.63). For participants not working in the science & IT domains, the odds of them purchasing the M1 laptop decreased by a factor of 0.43 (at a credible interval of 0.80).

Table 5. Bayesian Logistic Model Coefficients and Significance

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.33868    1.51475  -0.884  0.37682
trust_apple       0.93501    0.63952   1.462  0.14373
age_computer     -0.10035    0.09747  -1.030  0.30320
user_pcmac       -1.62333    0.60640  -2.677  0.00743 **
apple_count       0.04534    0.16239   0.279  0.78007
f_multi           0.16434    0.27380   0.600  0.54836
f_synergy         0.05182    0.18516   0.280  0.77957
m1_consideration  0.55077    0.19782   2.784  0.00537 **
domain           -0.85256    0.33224  -2.566  0.01029 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6. Odds Ratios of Predictors

		2.5 %	97.5 %
(Intercept)	0.262	0.013	5.089
trust_apple	2.547	0.727	9.189
age_computer	0.905	0.743	1.092
user_pcmac	0.197	0.058	0.632
apple_count	1.046	0.766	1.459
f_multi	1.179	0.684	2.020
f_synergy	1.053	0.729	1.515
m1_consideration	1.735	1.192	2.606
domain	0.426	0.216	0.803

The glm produces insightful information, however it is a simple variation of binary logistic regression. By utilising Bayesian multiple binary logistic regression, prior information about the model parameters can be incorporated. In this case, the unknown parameters from beta zero to beta seven were specified to have a normal distribution with low precision. Using JAGS and BUGS code within R allows for inferences about the model to be performed.

Table 7. Bugs Model Output

```
3 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
n.sims = 3000 iterations saved
         mu.vect sd.vect    2.5%     50%   97.5%  Rhat n.eff
beta_0    -1.527   1.584  -4.661  -1.507   1.454 1.001  3000
beta_1     1.046   0.668  -0.259   1.050   2.352 1.002  1200
beta_2    -0.113   0.104  -0.320  -0.111   0.088 1.001  3000
beta_3    -1.769   0.652  -3.073  -1.770  -0.478 1.001  3000
beta_4     0.049   0.171  -0.266   0.043   0.395 1.002  1600
beta_5     0.184   0.283  -0.374   0.189   0.756 1.001  3000
beta_6     0.057   0.196  -0.331   0.055   0.441 1.001  2500
beta_7     0.607   0.210   0.209   0.602   1.032 1.001  3000
beta_8    -0.924   0.350  -1.633  -0.917  -0.273 1.001  3000
deviance 136.379   4.474 129.912 135.678 147.395 1.001  3000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1)

DIC info (using the rule, pD = var(deviance)/2)
pD = 10.0 and DIC = 146.4
```

The trace plot (Figure 19) shows a consistent convergence after 4996 iterations. Additionally, the density plot (Figure 20) shows the chains of the different betas have converged around the same area, indicating that the Markov Chain Monte Carlo (MCMC) reached a stable state and is consistently sampling from the posterior distribution. The density distribution for the betas exhibits a normal distribution as opposed to a bimodal distribution, further reinforcing the stable convergence of the model. The Rhat values obtained are close to 1, demonstrating that the model is credible and provides reliable estimates.

Figure 19. Trace Plot for Beta Groups

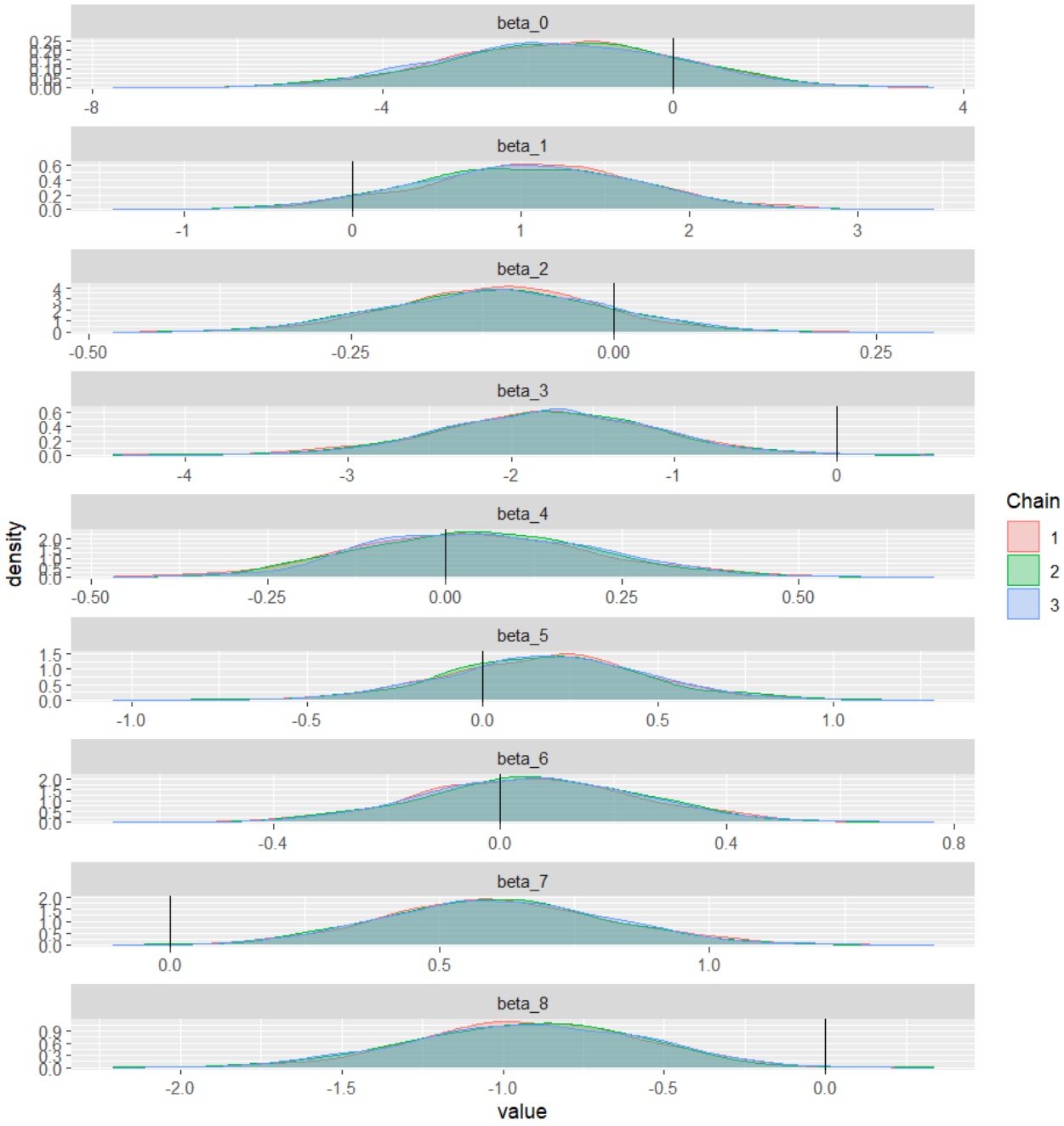

Figure 20. Density Plot for Beta Groups

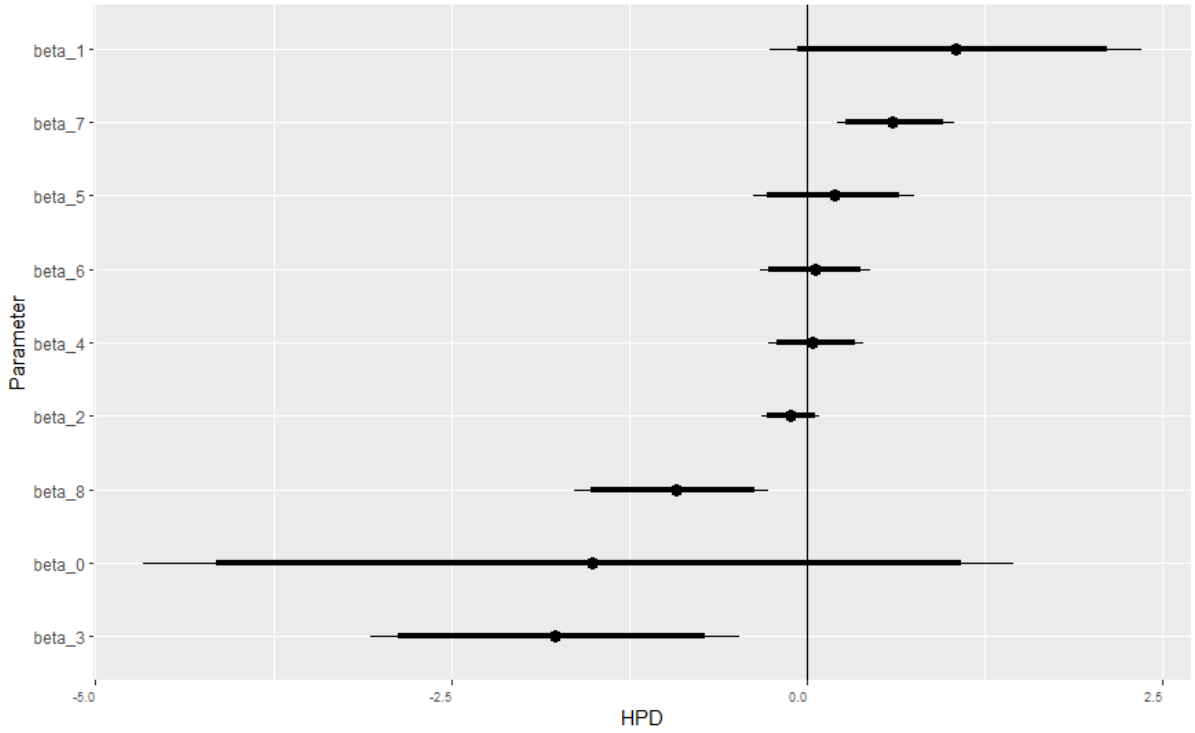

Figure 21. Caterpillar Plot for Beta Groups

Figure 21 visualises the 95% confidence intervals for the unknown parameters and three parameters show no posterior support to zero indicating an influence on the target variable. Beta six corresponding to m1_consideration shows to have a positive influence on purchase intentions. As a result, the higher the participants scored the importance of the M1 chip, the higher the probability that they would purchase the laptop. On the other hand, we can observe a negative influence on purchase intentions for beta eight and beta three corresponding to the participant's domain and the brand of the current computer, respectively. The dummy variables 0 for the user_pcmac variable represented all the participants who owned an Apple computer. Consequently, the consumers who currently own an Apple computer had a higher probability to purchase the M1 laptop as opposed to the consumers who own a PC. Additionally, the participants who were in the science & IT domains had a higher

probability of purchasing the M1 laptop as opposed to those in the business & legal and humanities domains. The confidence intervals for beta one, two, four, five and six all intersect 0 on the highest probability density suggesting that there is uncertainty about these parameters when influencing purchase intentions of the M1 laptop.

5 Discussion

The visualisations show us that trust had a significant influence on purchase intentions. However, multiple participants who did not trust Apple still chose to purchase the M1 laptop. This can be attributed to the positive reputation of the brand and their superior product superseding any trust concerns a consumer may have about the brand itself. Participants aged 23 showed to have the highest purchase intentions of the M1 laptop and this can be attributed to most participants being students who mainly rely on the use of computers for their studies. Additionally, Clemes et al. (2014) found younger consumers to have more internet experience which explains why the younger participants may showcase more purchasing interest than their older counterparts. Product related features also showed to have produce high purchase intentions particularly the battery life of the laptop as this is an important feature consumers will consider when purchasing a laptop. This aligns with the study conducted by Gatautis et al. (2014) who found product information and quality to have a significant impact on purchase intentions. Exploratory analysis of the data showed that most participants not making purchases of the laptop came from people who did not own any Apple products and did not trust the Apple brand. Furthermore, participants outside of the science & IT, business & legal domains showed lower purchase intentions for the laptop. This is likely due to the influence of trust affecting purchasing behaviour and repeat purchasing of similar products from the same brand. Additionally, participants working in technological domains such as IT are more likely to require a computationally powerful computer like the M1 laptop as opposed to participants working in the humanities domain.

The model evaluation metrics shows us that random forests and neural networks performed best across all metrics when tested on unseen data. Both models had the highest AUC score indicating both models are efficient at distinguishing between the two classes. Neural networks are gaining more popularity in business machine learning applications, and this can be attributed to their ability generate numerous accurate predictions, making them highly beneficial in predicting human consumer behaviour (Borres et al., 2023). Random forests producing good results is congruent with previous studies as this ensemble method has shown to consistently produce high evaluation metrics. Random forests also produced the highest specificity score of 87% showcasing the model's superior ability to identify the participants who would not make a purchase. However, random forest's sensitivity, and ability to correctly identify the participants who made a purchase was significantly lower at 57%. This poses a problem as the ability to correctly predict purchasers is crucial in the context of this study and for businesses seeking to understand their consumers. On the other hand, neural network showed to have a more balanced sensitivity and specificity at 75% and 74% respectively. As a result, neural network was considered the optimal model in this study due to its balanced ability to correctly identify the participants who did and did not make a purchase.

Table 8. Previous Work Comparison

	Previous Work		This Study	
	LR	SVM (P)	LR	SVM (P)
Accuracy	81%	66%	71%	71%

Data scientists working in Python employed logistic regression on the same dataset and achieved 81% accuracy on the final test set. Additionally, SVM with a polynomial kernel was also employed and achieved 66% accuracy. Both experiments stated that hyperparameter tuning should be applied to increase accuracy however, optimal parameters may not increase performance due to the small size of the dataset. Due to the small dataset, the model performance was not optimised as the model's ability to identify patterns leading to purchases may not be as accurate with a small number of data points to make observations. Despite the small data set, the different models from the tidymodels framework used in this study showed to have good performance on new data with an average accuracy of 73% across all models. Furthermore, this study fell short on LR accuracy but improved on the previously implemented polynomial SVM and revealed that the radial kernel SVM performs optimally for tasks like this study. Comparison with

The variable importance plot shows us that the number of Apple products owned by the participants had the most significance on influencing their purchase intentions. Studies have shown that a consumer's familiarity with a product can influence their purchase behaviour (Simon, 1955). As a results, participants owning multiple Apple products are more likely to be aware of the M1 laptop and the synergy between products is likely to influence their purchase intentions. This can also be linked to loyalty towards the brand and previous purchase behaviour. Despite trust in Apple having lowest importance in the variable importance plot, the visualisations showed that trust played a significant role in influencing purchase intentions. Extant literature has shown that a consumer's trust in a brand will directly influence purchase intentions with a lack of trust leading to a negative impact on purchase behaviours as consumers will avoid shopping with vendors considered untrustworthy (El Ansary &

Roushdy, 2013; Kamtarin, 2012). Additionally, consumers owning an Apple computer and two to four Apple products also showed to have the highest purchase intentions for the laptop. This showcases the importance of brand reputation and synergy between products as it can lead to increased loyalty for future purchases. This was closely followed by the age of the participant's current computer and the importance of the laptop's M1 chip. The age of participant's current computer may lead to increased purchase intentions for the laptop as a computer's performance declines overtime, eventually needing to be replaced. The laptop offers a wide variety of high-end unique features that are not available in their other products such as the M1 chip and the neural engine allowing apps on the laptop to use machine learning technologies (Apple, 2020). The top three most important variables found in this study were also found to be the most important in the previous work on the same dataset.

The Bayesian model found the participants domain, the type of computer they use and the importance of the M1 chip to be the variables influencing purchase intentions. The outcomes of the Bayesian model differed slightly to the random forest variable importance but both models showed to have the M1 chip consideration and the brand of the user's current computer to be the variables influencing purchase intentions. The participant's domain was the variable also found to influence purchase intentions by the Bayesian method. Exploratory analysis showed that the highest purchase intentions were by participants operating in the business & legal and science & IT domains. This variable may be significant as these domains' practices utilise computation heavy methods that are facilitated by powerful computers such as the M1 laptop and it's M1 chip. Overall, the Bayesian model can understand which variables influence purchase intention and allows for prior

assumptions about the data to be flexibly changed leading to more accurate and representative results. Additionally, conducting model inference on data of many sizes, allows for easy comparison with other models to assess performance and predictive ability. The results from this method show that Bayesian logistic regression serves as a reliable tool for gaining a deeper insight into consumer purchase intentions.

6 Conclusion

Based on the results of the variable importance plot and the Bayesian model, it is evident that the number of products a consumer own from the same brand has a significant impact on influencing purchase intention. Unique product features and the domain of consumers also showed to influence purchase intentions. Consequently, businesses can learn from machine learning analyses such as the ones presented in this research and can tailor their business strategies to capitalise on the factors that influence purchase intentions. For example, a business can optimise their target acquisition methods to reach the demographics showing the least interest in their products. Furthermore, businesses can provide discounts to consumers who purchase multiple products from the same brand to increase sales and consumer loyalty. In addition, this information can also be used to help improve the customer life cycle of the customers identified as their main demographics by offering after-sales services such as customer service support and further discounts.

Tidymodels offers a comprehensive framework for modelling and machine learning. It provides a streamlined service allowing for multiple models to be created, tuned, and fit to data promptly. In addition, the simple model creation process allows the performance of the different models to be quickly compared and evaluated. There are multiple methods for rectifying any issues with the data before it is trained and tested on machine learning models to ensure the best performances. Although some coding knowledge is required, this framework is easy to follow and is supported by a plethora of online tutorials and guides. This method produced good results that can be improved by utilising more comprehensive datasets. As a result, the framework

serves as a viable insight tool for non-experts seeking to gain a deeper understanding of their data and how different variables affect a target variable.

The main limitation to this study would be the size of the dataset as there are only 133 instances, of which 99 were used for training and 34 were used for testing. The small dataset may not show a clear representation of decisions leading to the target classes and can directly impact the models' abilities to understand the underlying patterns in the data. Furthermore, hyperparameter tuning can struggle to improve model performance due to the small data set size. Future studies can improve on this research by training and testing models on a larger data set as this will allow for the models to learn generalisable patterns and reduce overfitting. Additionally, a more representative group of participants should be utilised in future studies as the dataset mostly contained responses from students aged 23 years old. This demographic is not representative of the total population and may lead to biased results that produce skewed interpretations.

Furthermore, model training occasionally led to low accuracy scores ranging from 50 – 71%. However, when the models were fit to the testing data, the accuracy was significantly higher ranging from 79 – 85% accuracy and AUC scores up to 91%. Despite the testing scores being favourable, the score imbalances indicate the models learned to fit the data too closely as opposed to learning underlying patterns in the data and this generalisation led to overfitting. This can be attributed to the small dataset as the models were only trained on 99 instances and tested on 33. In addition, the data exhibits class imbalance in the target variable with more participants in the data showing an interest to purchase the laptop compared to non-purchasers. SMOTE was introduced in the preprocessing step to alleviate class

imbalance issues in the dataset. Although implementing SMOTE can fix the class imbalance issue, this method of oversampling the minority class can introduce issues with noisy instances and overfitting (Meng & Li, 2022). The utilisation of other oversampling/under sampling methods should be explored in future research, as well as implementing different machine learning models to find the most accurate classifier.

List of References

Abhishek, V., Hosanagar, K., and Fader, P. S. (2015) ‘Aggregation Bias in Sponsored Search Data: The Curse and the Cure’, *Marketing Science*, 34(1), pp. 59–77.

Adnan, H. (2014) ‘An Analysis of the Factors Affecting Online Purchasing Behavior of Pakistani Consumers’, *International Journal of Marketing Studies*, 6(5), pp. 133-148. Doi: 10.5539/ijms.v6n5p13

Aghdaie, S., Piraman, A., and Fathi, S. (2011) ‘An Analysis of Factors Affecting the Consumer’s Attitude of Trust and their Impact on Internet Purchasing Behavior’, *International Journal of Business and Society*, 2(23), pp. 147-158.

Akar, E. and Nasir, V.A. (2015) ‘A review of literature on consumers' online purchase intentions’, *Journal of Customer Behaviour*, 14(3), pp. 215-233.

Alin, A. (2010) ‘Multicollinearity’, *Wiley interdisciplinary reviews: computational statistics*, 2(3), pp. 370-374.

Amin, A., Shah, B., Khattak, A.M., Moreira, F.J.L., Ali, G., Rocha, A. and Anwar, S. (2019) ‘Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods’, *International Journal of Information Management*, 46, pp. 304-319.

Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B. and Zhang, J.D. (2020) ‘An introduction to machine learning’, *Clinical pharmacology & therapeutics*, 107(4), pp. 871-885.

Berman, R. (2018) ‘Beyond the last touch: Attribution in online advertising’, *Marketing Science*, *37*(5), pp. 771-792. Doi: 10.1287/mksc.2018.1104.

Blier-Wong, C., Cossette, H., Lamontagne, L. and Marceau, E. (2020) ‘Machine learning in P&C insurance: A review for pricing and reserving’, *Risks*, 9(1), 4.

Borres, R.D., Ong, A.K.S., Arceno, T.W.O., Padagdag, A.R., Sarsagat, W.R.L.B., Zuñiga, H.R.M.S. and German, J.D. (2023) ‘Analysis of Factors Affecting Purchase of Self-Defense Tools among Women: A Machine Learning Ensemble Approach’, *Applied Sciences*, 13(5), pp. 3003.

Bucklin, R. E., and Sismeiro, C. (2009) ‘Click here for Internet insight: Advances in clickstream data analysis in marketing’, *Journal of Interactive marketing*, 23(1), pp. 35-48.

Chaitanya C. and Gupta D. (2017) ‘Factors influencing customer satisfaction with usage of shopping apps in India’, *2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 1483-1486. Doi: https://doi.org/10.1109/RTEICT.2017.8256844.

Chaturvedi, S., and Gupta, S. (2014) ‘Effect of Social Media on Online Shopping Behaviour of Apparels in Jaipur City: An Analytical Review’, *Journal of Business Management, Commerce & Research*, 2(7), pp. 1-8.

Chen H., Yan Q., Xie M., Zhang D., and Chen Y. (2019) 'The Sequence Effect of Supplementary Online Comments in Book Sales,' *IEEE Access, vol. 7*, pp. 155650-155658. Doi: 10.1109/ACCESS.2019.2948190.

Chen, Y., Liu, H., Wen, Z., and Lin, W. (2023) 'How Explainable Machine Learning Enhances Intelligence in Explaining Consumer Purchase Behaviour: A Random Forest Model with Anchoring Effects', *Systems*, 11(6), pp. 312. Doi: https://doi.org/10.3390/systems11060312.

Cheng, F., Zhang, X., Zhang, C., Qiu, J., and Zhang L. (2018) 'An Adaptive Mini-Batch Stochastic Gradient Method for AUC Maximization', *Neurocomputing*, pp. 2–25.

Clemes, M.D., Gan, C., and Zhang, J. (2014) 'An empirical analysis of online shopping adoption in Beijing', *Journal of Retailing and Consumer Services*, 21(3), pp. 364-371. Doi: 10.1016/j.jretconser.2013.08.003

Dangeti, P. (2017) *Statistics for Machine Learning: Build Machine Learning Models with a Sound Statistical Understanding*. Birmingham: Packt Publishing Limited.

Doolin, B., Dillon, S., Thompson, F., and Corner, J. (2005) 'Perceived Risk, the Internet Shopping Experience an Online Purchasing Behavior: A New Zealand Perspective', *Journal of Global Information Management*, 13(2), pp. 66-88. Doi: 10.4018/jgim.2005040104

Downing, C. E. (2010) 'Is web-based supply chain integration right for your company?', *Communications of the Association for Computing Machinery (ACM),* 53(5), pp. 134–137.

Duan, L. and Xiong, Y. (2015) 'Big data analytics and business analytics', *Journal of Management Analytics*, 2(1), pp.1-21.

El Ansary, O., and Roushdy, A. (2013) 'Factors Affecting Egyptian Consumers' Intentions for Accepting Online Shopping', *The Journal of American Academy of Business, Cambridge*, 19(1), pp. 191-201.

Fan, J., Han, F. and Liu, H. (2014) 'Challenges of big data analysis', *National science review*, 1(2), pp. 293-314.

Forbes (2020) *PC Sales Surge During Coronavirus Crisis - And HP's The Big Winner*. Available at: https://www.forbes.com/sites/barrycollins/2020/07/10/pc-sales-surge-during-coronavirus-crisisand-hps-the-big-winner/?sh=7c03fac75227 (Accessed 8 June 2023).

Gatautis, R., Kazakeviciute, A., and Tarutis, M. (2014) 'Controllable Factors Impact on Consumer Online Behaviour', *Economics and Management*, 19(1), pp. 63-71. Doi: 10.5755/j01.em.19.1.5692

Gong, W., and Maddox, L. (2011) 'Online Buying Decisions in China', *The Journal of American Academy of Business, Cambridge*, 17(1), pp. 43-50.

Ha, N.T., Nguyen, T.L.H., Van Pham, T., and Nguyen, T.H.T. (2021) 'Factors influencing online shopping intention: An empirical study in Vietnam', *The Journal of Asian Finance, Economics and Business*, 8(3), pp.1257-1266.

Hagger-Johnson, G. (2014) *Introduction to Research Methods and Data Analysis in the Health Sciences*. London: Taylor & Francis Group.

Hamami, F. and Muzakki, A. (2021) 'Machine learning pipeline for online shopper intention classification', *AIP Conference Proceedings*, 2329, 050014. Doi:10.1063/5.0043452.

Hamet, P. and Tremblay, J. (2017) 'Artificial intelligence in medicine', *Metabolism*, 69, pp. 36-40.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann Publishers.

Han, J., Kamber, M., and J. Pei. (2011) *Data Mining Concepts and Techniques*. Waltham: Morgan Kaufmann Publishers.

Hanafy M, Ming R. (2021) 'Machine Learning Approaches for Auto Insurance Big Data', *Risks*, 9(2), pp. 42. https://doi.org/10.3390/risks9020042

Hardesty, D.M. and Suter, T.A. (2005) 'E-tail and retail reference price effects', *Journal of Product & Brand Management*, 14(2), pp. 129-136. Doi: 10.1108/10610420510592626

Jannach, D., Ludewig, M. and Lerche, L. (2017) 'Session-based item recommendation in e-commerce: on short-term intents, reminders, trends, and discounts', *User Model User-Adap Inter* 27, pp. 351–392. Doi: 10.1007/s11257-017-9194-1

Jia, J. (2019) 'Analysis of alternative fuel vehicle (AFV) adoption utilizing different machine learning methods: a case study of 2017 NHTS', *IEEE Access*, 7, pp.112726-112735.

Kabir, M.R., Ashraf, F.B. and Ajwad, R. (2019) 'December. Analysis of different predicting model for online shoppers' purchase intention from empirical data', *International Conference on Computer and Information Technology (ICCIT)*, pp. 1-6.

Kamalul Ariffin, S., Mohan, T., and Goh, Y. N. (2018) 'Influence of consumers' perceived risk on consumers' online purchase intention', *Journal of Research in Interactive Marketing*, 12(3), pp. 309–327. Doi: 10.1108/JRIM-11-2017-0100

Kamtarin, M. (2012) 'The Effect of Electronic word of Mouth, Trust, and Perceived Values on Behavioral Intention from the Perspective of Consumers', *International Journal of Academic Research in Economics and Management Sciences*, 1(4), pp. 56-66.

Kaplancan, G.V. (2017) *A Case Study on the Development and Problems of E-Commerce, Virtual Business and Virtual Merchandising in Turkey and the World.* Istanbul: Nisantasi University.

Karegowda, A. G., Manjunath, A., S. and Jayaram, M. A. (2010) 'Feature subset selection problem using wrapper approach in supervised learning', *International Journal of Computer Applications*, 1(7), pp. 13–17.

Kiang, M.Y. (2003) 'A comparative assessment of classification methods', *Decision support systems*, 35(4), pp. 441-454.

Kiri L. Wagstaff. (2012) 'Machine learning that matters', *Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML'12)*, pp. 1851–1856.

Kuhn, M. and Wickham, H. (2020) *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Available at: https://www.tidymodels.org (Accessed 22 June 2023).

Kumar, V. (2014) 'Feature selection: A literature review', *The Smart Computing Review*, 4(3), pp. 211-229.

Kurniawan, I., Abdussomad, Akbar, M.F., Saepudin, D., Azis, M.S., & Tabrani, M. (2020) 'Improving the Effectiveness of Classification Using the Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction', *Journal of Physics: Conference Series*, 1641.

Lee, J., Jung, O., Lee, Y., Kim, O. and Park, C. (2021) 'A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion', *Journal of Theoretical and Applied Electronic Commerce Research*, 16(5), pp. 1472-1491.

Li H. and Peng T. (2020) 'How Does Heterogeneous Consumer Behavior Affect Pricing Strategies of Retailers?', *IEEE Access*, vol. 8, pp. 165018-165033.

Li, H., and Kannan, P. K. (2014) 'Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment', *Journal of Marketing Research*, 51(1), pp. 40–56. Doi: 10.1509/jmr.13.0050

Li, R., Kim, J., and Park, J. (2007) 'The Effects of Internet Shoppers' Trust on Their Purchasing Intention in China', *Journal of Information Systems and Technology Management*, 4(3), pp. 269-286. Doi: 10.1590/S1807-17752007000300001

Lim J., Grover V., and Purvis R. L. (2012) 'The Consumer Choice of E-Channels as a Purchasing Avenue: An Empirical Investigation of the Communicative Aspects of Information Quality', *IEEE Transactions on Engineering Management*, 59(3), pp. 348-363. Doi: 10.1109/TEM.2011.2164802.

Liu, J. (2022) 'Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data', *Soft Computing*, 26(3), pp.1141-1163.

Lu, C.W., Lin, G.H., Wu, T.J., Hu, I.H. and Chang, Y.C. (2021) 'Influencing factors of cross-border e-commerce consumer purchase intention based on wireless network and machine learning', *Security and Communication Networks*, 2021, pp.1-9.

Lundberg, S. M., and Lee, S. I. (2017) ‘A unified approach to interpreting model predictions’, *Advances in neural information processing systems*, 30.

Mahesh, B. (2020) ‘Machine learning algorithms-a review’, *International Journal of Science and Research (IJSR),* 9(1), pp. 381-386.

Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C. and Haltmeier, M. (2020) ‘A machine learning framework for customer purchase prediction in the non-contractual setting’, *European Journal of Operational Research*, 281(3), pp. 588-596.

Meng, D. and Li, Y. (2022) ‘An imbalanced learning method by combining SMOTE with Center Offset Factor’, *Applied Soft Computing*, 120, pp. 108618.

Meyer, D. and Wien, F.T. (2015) ‘Support vector machines’, *The Interface to libsvm in package e1071*, 28(20), pp. 597.

Moe, W. W., and Fader, P. S. (2004) ‘Capturing evolving visit behavior in clickstream data’, *Journal of Interactive Marketing*, 18(1), pp. 5–19

Mokryn, O, Bogina, V and Kuflik, T (2019) 'Will this session end with a purchase? Inferring current purchase intent of anonymous visitors', *Electronic Commerce Research and Applications*, 34, 100836. Doi: 10.1016/j.elerap.2019.100836.

Momtaz, H., Islam, A., Ariffin, K., and Karim, A. (2011) ‘Customers Satisfaction on Online Shopping Malaysia’, *International Journal of Business and Management*, 6(10), pp. 162-169. Doi: 10.5539/ijbm.v6n10p162

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011) ‘The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature’, *Decision Support Systems*, 50(3), pp. 559–569.

Noviantoro, T., and Huang, J.-P. (2021) ‘Applying data mining techniques to investigate online shopper purchase intention based on clickstream data’, *Review of Business, Accounting, & Finance*, 1(2), pp. 130-159. Available at: https://fortunepublishing.org/index.php/rbaf/article/view/15 (Accessed: 24 February 2023).

Oncioiu, I. (2014) ‘The Impact of Tourist Feedback in the Virtual Community on the Purchase Intention’, *International Business Research*, 7(3), pp. 28-33. Doi: 10.5539/ibr.v7n3p28

Özdemir, R. and Turanli, M. (2021) ‘Comparison of machine learning classification algorithms for purchasing forecast’, *Journal of Life Economics*, 8(1), pp. 59-68.

Parihar, V. and Yadav, S. (2022) ‘Comparative Analysis of Different Machine Learning Algorithms to Predict Online Shoppers' Behaviour’, *International Journal of Advanced Networking and Applications*, 13(6), pp. 5169-5182.

Policarpo, L.M., Silveira, D.E., Righi, R.D., Antunes, R.S., Costa, C.A., Barbosa, J.L., Scorsatto, R., and Arcot, T. (2021) ‘Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review’, *Comput. Sci. Rev*., 41, 100414.

Qiu, J., Lin, Z. and Li, Y. (2015) ‘Predicting customer purchase behavior in the e-commerce context’, *Electronic commerce research*, 15, pp. 427-452.

Quan J., Wang X., and Quan Y. (2019) ‘Effects of Consumers’ Strategic Behavior and Psychological Satisfaction on the Retailer’s Pricing and Inventory Decisions’, *IEEE Access*, vol. 7, pp. 178779-178787. Doi: 10.1109/access.2019.2958685.

Rana, Z.A., Mian, M.A. and Shamail, S. (2015) ‘Improving Recall of software defect prediction models using association mining’, *Knowledge-Based Systems*, *90*, pp.1-13.

Rawat, S., Rawat, A., Kumar, D., and Sabitha, A.S. (2021) ‘Application of machine learning and data visualization techniques for decision support in the insurance sector’, *International Journal of Information Management Data Insights*, 1(2), pp. 100012.

Rubi, M.A., Bijoy, M.H.I., Chowdhury, S., and Islam, M.K. (2022) ‘Machine Learning Prediction of Consumer Travel Insurance Purchase Behavior’, *International Conference on Computing Communication and Networking Technologies (ICCCNT),* pp. 1-5.

Sakar, C.O., Polat, S.O., Katircioglu, M. and Kastro, Y. (2018) ‘Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks’, *Neural Computing and Applications*, 31, pp. 6893-6908.

Salchenberger, L.M., Cinar, E.M. and Lash, N.A. (1992) ‘Neural networks: A new tool for predicting thrift failures’, *Decision Sciences*, *23*(4), pp. 899-916.

Saprikis, V. (2013) ‘A Longitudinal Investigation on Greek University Students’ Perceptions towards Online Shopping’, *Journal of Electronic Commerce in Organizations*, 11(1), pp. 43-62. doi: 10.4018/jeco.2013010103

Seckler, M., Heinz, S., Forde, S., Tuch, A. N., and Opwis, K. (2015) ‘Trust and distrust on the web: User experiences and website characteristics’, *Computers in Human Behavior*, 45, pp. 39-50. Doi: 10.1016/j.chb.2014.11.064

Setyaningsih E. R and Listiowarni I. (2021) ‘Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing’, *East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pp. 330-333. Doi: 10.1109/EIConCIT50028.2021.9431862.

Severino, M.K. and Peng, Y. (2021) ‘Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata’, *Machine Learning with Applications*, 5, p.100074.

Shao, X., Li, L. (2011) ‘Data-driven multi-touch attribution models’, *17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 258-264.

Shapley, L.S. (1953) ‘Stochastic games’, *Proceedings of the national academy of sciences*, 39(10), pp. 1095-1100.

Simon, H.A. (1955) ‘Behavioral Model of Rational Choice’, *The Quarterly Journal of Economics*, Volume 69, Issue 1, pp. 99–118. Doi: 10.2307/1884852

Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G. and Bostrom, J. (2019) ‘Deep reinforcement learning for multiparameter optimization in de novo drug design’, *Journal of chemical information and modeling*, *59*(7), pp. 3166-3176. Doi: 10.26434/chemrxiv.7990910.v2

Statista (2023) *E-commerce worldwide - statistics & facts*. Available at: https://www.statista.com/topics/871/online-shopping/#topicOverview (Accessed 12 June 2023)

Statista (2023) *Laptops – Worldwide*. Available at: https://www.statista.com/outlook/cmo/consumer-electronics/computing/laptops/worldwide?currency=GBP (Accessed 7 June 2023)

Surjandy, Cassandra C., Meyliana, Eni Y., Marcela Y., and Clarissa S. (2021) ‘Analysis of Product Trust, Product Rating and Seller Trust in e-Commerce on Purchase Intention during the COVID-19 Pandemic’, *International Conference on Information Management and Technology (ICIMTech)*, pp. 522-525. Doi: 10.1109/ICIMTech53080.2021.9534964.

Thamizhvanan, A., and Xavier, M. (2013) ‘Determinants of customers’ online purchase intention: an empirical study in India’, *Journal of Indian Business Research*, 5(1), pp. 17-32. Doi: 10.1108/17554191311303367

Trivedi, S.K., Patra, P., Srivastava, P.R., Zhang, J.Z. and Zheng, L.J. (2022) ‘What prompts consumers to purchase online? A machine learning approach’, *Electronic Commerce Research,* pp.1-37. Doi: 10.1007/s10660-022-09624-x

Tsuboi, Y., Jatowt, A. and Tanaka, K. (2015) ‘Product purchase prediction based on time series data analysis in social media’ *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),* 1, pp. 219-224.

Tufail H., Ashraf M. U., Alsubhi K., and Aljahdali H. M. (2022) ‘The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection’, *IEEE Access*, 10, pp. 25555-25564. Doi: 10.1109/access.2022.3152806.

Udell, M. (2014) ‘From insight to action’, *Marketing Insights*, 26(6), pp. 38–43.

Vaidehi, P.U. (2014) ‘Factors Influencing Online Shopping Behavior of Students in Engineering Colleges at Ranga Reddy District’, *Sumedha Journal of Management*, 3(1), pp. 50-62.

Vamosi, S., Platzer, M. and Reutterer, T. (2022) ‘AI-based Re-identification of Behavioral Clickstream Data’, *Proceedings of the European Marketing Academy*, 51, 106830.

Vinerean, S., Certina, I., Dumitrescu, L., & Tichindelean, M. (2013) ‘The Effects of Social Media Marketing on Online Consumer Behavior’, *International Journal of Business and Management*, 8(14), pp. 66-79. Doi: 10.5539/ijbm.v8n14p66

Wang, N. (2021) ‘Research on the influence of the cross-border e-commerce development of small and medium-sized enterprises in Dongguan in the post-epidemic era’, *2nd International Conference on E-Commerce and Internet Technology (ECIT)*, pp. 176-180. Doi: 10.1109/ECIT52743.2021.00047.

Wang, P. and Xu, Z. (2020) ‘A novel consumer purchase behavior recognition method using ensemble learning algorithm’, *Mathematical Problems in Engineering*, pp.1-10.

Wen, Z., Lin, W. and Liu, H. (2023) ‘Machine-Learning-Based Approach for Anonymous Online Customer Purchase Intentions Using Clickstream Data’, *Systems, 11(5), pp. 255.*

Wong, A.N. and Marikannan, B.P. (2020) ‘Optimising e-commerce customer satisfaction with machine learning’, *Journal of physics: Conference series*, 1712(1), pp. 012044.

Yale, K, Nisbet, R, Miner, G.D. (2017) *Handbook of Statistical Analysis and Data Mining Applications*. San Diego: Elsevier Science & Technology.

Zeng, M., Cao, H., Chen, M., and Li, Y. (2018) ‘User behaviour modeling, recommendations, and purchase prediction during shopping festivals’, *Electronic Markets*, 29(2), pp. 263-274. Doi: 10.1007/s12525-018-0311-8

Zhai, X., Shi, P., Xu, L., Wang, Y. and Chen, X. (2020) ‘Prediction Model of User Purchase Behavior Based on Machine Learning’, *IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1483-1487. Doi: 10.1109/ICMA49215.2020.9233677.

Zhao H. H., Luo X. C., Ma R., and Lu X. (2021) ‘An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation with Correlated Variables’, *IEEE Access*, 9, pp. 48405-48412. Doi: 10.1109/access.2021.3067499.

Zhao T., Hu M., Rahimi R., and King I. (2017) ‘It's about time! Modeling customer behaviors as the secretary problem in daily deal websites’, *International Joint Conference on Neural Networks (IJCNN)*, pp. 3670-3679. Doi: 10.1109/IJCNN.2017.7966318.

Zhao, Y., Yao, L., and Zhang, Y. (2016) ‘Purchase prediction using Tmall-specific features’, *Concurrency Computat.: Pract. Exper.*, 28: pp. 3879– 3894. Doi: 10.1002/cpe.3720.

Zhou, Z. H. (2021). *Machine learning*. Springer Nature.