Black Friday Sales Prediction

Anusha Chatra Anilkumar – 6720072

Surrey Business School

University of Surrey

Thesis submitted in the partial fulfilment of the requirements for the Degree of Master of Science in Business Analytics

January 06, 2023

Word Count: 15500

Executive Summary

As technology advances, business analysis has become increasingly important for companies. Many organizations choose to use high-tech methods to manage their operations, rather than traditional approaches Mao (2022). Sales, which are a crucial part of a company's operations and can impact its profit and management decisions, are an area where business analytics can be particularly useful. As a result, the use of machine learning for sales prediction has become popular among companies as a way to drive progress and improve performance. Sales prediction is primarily used to set sales performance goals for a business and manage inventory. This project provides accurate results and helps to estimate the quantity of raw materials needed for future production. The sales of products that are often purchased together can be predicted using this project. By understanding which inventory items are most popular, retailers and shopkeepers can improve their profit margins. The prediction is based on the company's past sales data and can be used to forecast future sales for the business or specific products.

The Friday after Thanksgiving is known as the busiest shopping day of the year in the United States and marks the start of the busiest shopping season. From a computer science perspective, one of the most interesting applications of machine learning in the retail industry is the ability to accurately predict how much a customer is likely to spend at a store based on their past purchasing patterns and behaviours. This can be helpful for retailers in planning and managing their inventory and sales during the busy holiday season. Black Friday marks the beginning of winter shopping season and this is the time where most of the products are on sale. As a result of this, customers tend to rush to buy things before it gets sold out. The stores need to prepare well in advance to deal with crowd and inventory management. Performing analysis on the previous data helps to predict the future sales which will the help the store to be well prepared. The aim of this study is to predict sales during Black Friday.

In the literature survey section, I have summarised the previous work done on sales prediction using research papers between the years 1987 to 2022. It consists of various machine learning algorithms used to forecast sales, tools used for visualisation and methods used for evaluating prediction accuracy.

For this study, I have used the dataset form Kaggle by Sammari (2020) to perform the analysis. The analysis has been performed using 50000 observations and 12 features with Purchase as the dependent variable. The independent variables are User_ID, Product_ID, Age, Gender, Marital_Status, Occupation, City_Category, Stay_In_Current_City_Years, Product_Category_1, Product_Category_2 and Product_Category_3. The dataset consists of both numerical and categorical variables. CRISP-DM approach was used to perform the analysis which consists of phases such as Data understanding, Data pre-processing, Modelling and Evaluation of results. I have used both R programming language on Rstudio and Tableau to perform data visualisation. All the graphs have been plotted using these two tools. All the interesting patterns and relationships found have been described in the data visualisation section. Basic information about the dataset such as datatype of the variables and summary of all the variables have been described. Categorical variables have been converted to factors before plotting the graphs for a better description.

In the data pre-processing phase, I have used the correlation plot to find the significance of the variables. The data was scaled since the values were on different scales. Based on the significance I have done feature selection that is, eliminated insignificant variables with p value greater than 0.05. I have also checked for missing values in the dataset. Product_Category_3 had almost 70% of missing values and I have dropped that variable.

Product_Category_2 had around 30% missing values which was treated by imputing mean value of the variable. Since variables User_ID and Product_ID does not impact the sales prediction in any way, those variables have been dropped.

After data pre-processing, the dataset was split into training and testing partition in the ratio 70:30. Machine learning algorithms such as Multiple linear regression, Support vector machine and Random forest were applied using different variations on the training partition of the dataset. The trained model was then applied to testing partition of the dataset to predict the sales. 4 models were developed using multiple linear regression approach with elimination of insignificant variable at every model improvement. Model 4 with variables Age, Gender, City_Category, Product_Category_1 and Product_Category_2 was observed to have the best result in comparison with other models. 2 models were developed with SVM algorithm using radial and polynomial kernel. Model with radial kernel performed better than the polynomial kernel. 3 models were developed using random forest using 100, 500 and 1000 trees respectively. Results for all the 3 models were quite similar with very little or negligible difference. Hence model with 100 trees was chosen to be the best model since it provided good result along with being cost and time efficient.

Root mean squared error (RMSE) and Mean absolute error (MAE) were used to check for model prediction accuracy. Random forest model for training dataset had the least RMSE and MAE value of 2345.05 and 1749.38 respectively. But RMSE and MAE values for testing dataset was 3037.156 and 2290.677. Since there is a huge difference between training and testing accuracy, this model can be unreliable to predict sales. SVM model with radial kernel had a RMSE and MAE value of 4176.846 and 3001.031 which was better than the residuals of multiple linear regression model. Multiple linear regression model had the highest RMSE and MAE values. The testing dataset of SVM radial kernel had a RMSE and MAE value of 4189.219 and 3032.828. The difference between training testing data was comparatively very minimal and thus making it the most effective model to predict sales with good accuracy.

Declaration of Originality

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

Anusha Anilkumar

…….. ……

MSc Business Analytics Student Signature of the Student

Anusha Chatra Anilkumar (6720072)

Surrey Business School

University of Surrey Date: 6th January 2023

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Philip Murray for his invaluable guidance and support throughout my research. I am also grateful for his continuous feedback and encouragement which helped me immensely to deliver on time. His profound knowledge on analytics helped me to choose the research subject and work towards the completion of it. I would also like to thank my family and friends for the support provided throughout the completion of my master's degree.

Table of Contents

List of Abbreviations

Consumer Goods	(CG)
Business-to-Business	(B2B)
Business-to-Consumer	(B2C)
Point of Sale	(POS)
Standard Deviation	(SD)
Root Mean Squared Error	(RMSE)
Mean Squared Error	(MSE)
Mean Absolute Error	(MAE)
K-Nearest Neighbor	(KNN)
Random Forest	(RF)
Support Vector Machine	(SVM)
Gradient Boosting	(GB)
Decision Tree	(DT)
Geographic Information System	(GIS)
eXtreme Gradient Boosting	(XGBoost)
Iterative Dichotomiser 3	(ID3)
Convolutional Neural Network	(CNN)
Artificial Neural Network	(ANN)
Back Propagation Network	(BPN)
Multivariate Intelligent Decision-Making	(MID)
Multi-Input-Single-Output	(MISO)
Data preparation and pre-processing	(DPP)
Cross Industry Standard Process for Data Mining	(CRISP-DM)
Not Available	(NA)

1 Introduction

Sales are essential to the success of any business, and sales forecasting plays a crucial role in managing a company. By looking at past data and conditions, good forecasting helps businesses develop and improve their strategies, increase their understanding of the market, and prepare for the future. A standard sales forecast involves analysing past resources and using that information to make predictions about customer acquisition, identify strengths and weaknesses, and set budgets and marketing strategies for the coming year. In summary, sales forecasting is a way of predicting future sales based on past data and resources. Businesses that prioritise sales forecasting tend to perform better than those that don't Punam, et al., (2018).

Forecasting sales is the process of determining how much money a company, group of people, or person will make during a specific time period. Research in retail management and retail geography has paid a great deal of attention to the science of sales forecasting. A more analytical approach to decision-making is made possible by the capacity to anticipate revenues accurately. Given the significant upfront expenses associated with building supermarkets and hypermarkets, the top merchants have become highly knowledgeable in this field and established specialised sales forecasting teams. By doing this, the rate of sales prediction accuracy has grown, enabling these big retailers to make educated site purchase decisions that have strengthened their dominance Wood (2006).

Sales forecast is mostly used to monitor inventory levels and set targets for the company's sales performance. Based on years of historical company data, it is simpler for established organisations to forecast future sales. In order to predict future business, newly established businesses must rely on less reliable information, such as market research. It is one of the foundations of sound financial management. Forecasting sales can help a business decide how to allocate its employees, cash flow, and resources. Accurate sales predictions help businesses plan their operations and forecast both short- and long-term performance Praveen (2022).

The consumer goods (CG) industry is highly competitive, with various manufacturers and retailers competing for market share for similar products. Marketing strategy plays a crucial role in the success of a product in this industry, and is influenced by various factors, some of which are observable while others are not. The main goal of any marketing strategy in the CG industry is to create a sustainable competitive advantage. The CG industry involves both business-to-business (B2B) and business-to-consumer (B2C) interactions, with manufacturers and retailers interacting on one hand, and retailers and consumers interacting on the other. Point of sale (POS) data, which includes information about products sold in stores, is important for the development of marketing strategy in the CG industry, as not all interactions in this space are recorded Sundararaman (2012).

The aim of this project is to predict the sales during Black Friday based on the previous sales transactions captured at a retail store.

In the United States, the Friday following Thanksgiving is regarded as the busiest shopping day of the year since it ushers in the busiest shopping season. One of the most intriguing uses of machine learning in the retail sector, from a computer science standpoint, is the capability to precisely forecast how much a customer would likely spend at a store based on their previous purchase patterns. During the busy Christmas season, this might be useful for merchants in planning and controlling their inventory and sales. The purpose of this project is to predict the sales during Black Friday. Enterprises must have an accurate sales estimate because missing a forecast might be harmful. They cannot afford to miss their forecasts. For instance, a business

may underperform and overestimate, which could have a detrimental impact on the value of the company. However, if a corporation overpromises and under promises, even while this is a positive thing, it won't provide organizations enough time to plan marketing campaigns, introduce new products, or hire new employees without losing money Kothandaraman (2021).

Generally the supermarkets are crowded on Black Friday amid the offers. Many products have limited-time discounts, which encourages customers to buy more of the products. Even with a solid arrangement, it is challenging for customers to buy the products. In any case, managing the group with few workers and concentrating on approaching customers provide the shop owners with far greater challenges. A few approaches have been used to address this problem, and they are not implausibly successful. A system with potential for resolving this problem is a model for prediction Ramasubbareddy (2021).

Decision-makers are frequently forced to base their conclusions on arbitrary mental models that represent their past experiences due to the complexity of business dynamics. However, studies have shown that when data-driven decision-making is used, businesses perform better Bohanec (2017). Companies in the top third of their industry are, on average, 5% more productive and 6% more lucrative than their rivals when using data-driven decision-making Brynjolfsson, et al., (2011).

Research has shown that users are more likely to accept and follow recommendations when an explanation of the decision is provided Gönül (2006). However, in terms of prediction accuracy, more complex and less transparent machine learning models, such as random forests, boosting, support vector machines, and neural networks, tend to outperform simpler and more interpretable models like decision trees, naive Bayes, and decision rules Caruana (2006). In summary, while providing explanations for decisions can improve acceptability, it is important to balance this with the need for accurate predictions, which may require the use of more complex machine learning models.

The machine learning algorithm involves four steps when applied to real-world business situations: demand, exploration, development, and evaluation Mao (2022). The first step is to define the prediction object and understand the need or demand for it. For example, sales prediction requires setting a target or expectation for the desired outcome, which determines the timeline for the prediction. The remaining steps involve exploring and preparing the data, developing the machine learning model, and evaluating its performance. In general, a company that wants to predict its sales for one month will need data from the past two years. Specifically, short-term predictions can impact a company's storage or production, while long-term predictions can affect its management operations. The importance of long-term and short-term predictions varies depending on the industry. It is important to determine the appropriate time frame for a prediction based on the needs of the company and the industry it operates in. When predicting sales, companies should consider variables that can affect changes in sales. However, the main goal of evaluating a prediction model is to increase profit, so in addition to considering sales, companies should also take into account other aspects of their operations. This will help them make more informed and comprehensive predictions that can benefit the overall performance of the company Eberly (2007).

1.1 Deliverables

The aim of this project is to forecast sales during black Friday. Following are the deliverables,

- Initial data visualisation using Tableau and R programming on Rstudio
- Pre-processing the data that is, data cleaning, scaling and feature selection
- Applying various machine learning algorithms
- Comparing results of different models and selecting the best model based on the prediction accuracy

2 Literature review

This section gives an overview of the existing work that has been done on sales prediction. It showcases the academic analysis of the methods followed to forecast sales of a company. It will include studies conducted between 2000 and 2022 to forecast sales using machine learning algorithms. It will also include studies describing the tools and techniques used for data visualisation, modelling and performance metrics. Overall, this section outlines the previous work done on sales prediction using different machine learning techniques.

2.1 Applied Machine Learning for Supermarket Sales Prediction

There are a variety of methods for predicting supermarket sales, however historically, many supermarkets have primarily used the conventional statistical models Dinane (2015). However, machine learning has developed into a significant field of data science that has gained popularity because of its strong predictive and forecasting abilities, and as a result, it has replaced other equally crucial methods for extremely accurate sales forecasting.

Odegua (2020) used data examined using the machine learning methods K-Nearest Neighbor, Gradient Boosting, and Random forest to predict sales for the supermarket chain "Chukwudi Supermarkets". The information included a variety of supermarket factors, including the opening year, product pricing, supermarket location, etc. The data set included a sample of 4990 occurrences with 13 features.

Gradient Boosting Model - The number of boosted trees (n estimators) for gradient boosting was set to 200, the maximum depths to 6, the maximum features to be square roots, and the minimum sample split was set to 4. Other settings were set at their default values.

For the random forest model, the max depth was set to 5, and the n estimators value was set to 100. Other settings were left at their default values.

Mean Absolute Error (MAE) was used to evaluate the model. Lower the MAE, better the model. A tried-and-true metric, MAE provides a reliable indicator of model performance.

A measure of the difference between two continuous variables is called mean absolute error (MAE). The Mean Absolute Error (MAE) is the mean upright length between each observed data point and the forecasted position when X and Y are features from an event, say X is the given value and Y is the predicted value from the model obtained by machine learning algorithms.

The mean absolute error is given by:

$$\text{MAE} = \sum_{i=1}^{n} \frac{|y_i - x_i|}{n}$$

The Random Forest approach worked better than the other two algorithms, with a MAE of 0.409178, when the mean forecast of the 10-fold cross validation was used. The MAE of the Gradient Boosting model is quite similar to that of the KNN, although it has a substantially lower standard deviation. Figure 1 shows the learning curves of the 3 models.

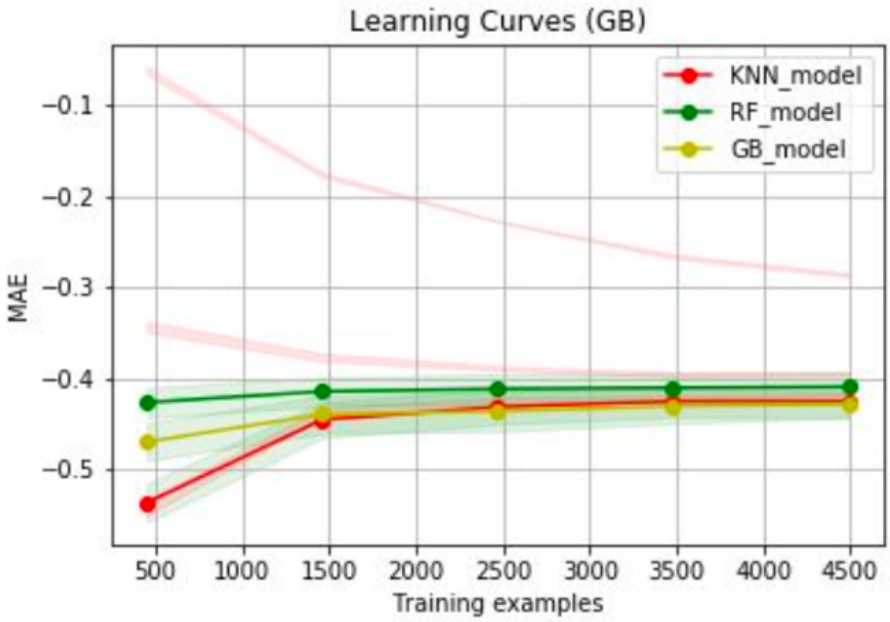

Figure 1: Learning curves of the 3 models. Odegua (2020)

Models	MAE	SD
KNN	0.425103	0.018008
RF	0.409178	0.014420
GB	0.428260	0.014420

Figure 2: MAE and SD values for the 3 models. Odegua (2020)

From Figure 2 we can see that random forest algorithm had the least mear residual than the KNN and GB, although for all three models, the training value and cross-validation value are initially poor but improve as the training size grows. Additionally, we see that the MAE for KNN is rising, indicating that adding more data sets will improve performance. Also, the

models used in this study were trained on a little amount of data, which may have contributed to the usually high error rates. This might be one of the limitations of our project as well.

We will apply SVM and linear regression algorithm along with SVM in this project to compare the results. We also increase the number to records to obtain better accuracy. For evaluation, RMSE value is also considered.

The purpose of the research by Kaneko and Yada (2016) was to forecast the sales of a retail store and evaluate the effectiveness of the system. They collected the data between 2002 to 2004 from supermarkets in Japan. They used 3 categories of data; Category 1 had 62 broadcast attributes, Category 2 had 569 further divided attributes and Category 3 ha 3312 of the finest divided attributes. From Category 1 to Category 3, the traits are more specialised, and the product characteristics are more precise. In summary, the authors used POS data from supermarkets to develop a system for predicting retail store sales using logistic regression model. They found that the predictive accuracy of the system varied from 75% to 86% depending on the number of product attributes. The authors also used daily data to create categories and make decisions about which products to prioritize based on their sales performance. These categories helped in giving better insights by targeting particular segments. As opposed to the previous model by Odegua (2020) which had 13 variables, this study only used 3 categories to build the model. Since that dataset used for our project also uses categories in City, Age, Products and Occupation, it will give targeted results.

Demographics are statistical characteristics of a population, often including socio-economic features such as age, education level, occupation, income, marital status, and average family size Zhang, et al., (2016). In the context of a website, these demographics can be collected from visitors and analysed to understand their activities on the site. Data for an offline store is collected using the bills generated when a customer makes a purchase. Apriori algorithm is a type of association rule that is used to identify frequent patterns in data. It works by iteratively generating item-sets of increasing size and identifying those that meet a minimum support threshold. This process continues until no more item-sets can be generated. The use of association rules in this system allows the store owner to understand which products are frequently bought together and can help inform decisions on which products to stock in the store. By using data mining techniques such as association rule mining, retailers can better understand customer behaviour and make informed decisions about their product offerings. In this project, we will be using data collected from an offline supermarket.

The system for an online stationery store presented in Setiawan, et al. (2017) paper intends to enhance the current manner of operation by introducing association rules and offering advice to the owner on which items to include with the current ones. By grouping things based on support and confidence, association rules are employed to make suggestions for the store owner that are pertinent and related. Items that do not satisfy the needed support are removed, and candidate item sets are generated using the Apriori algorithm from larger sets in earlier stages. This enables the retailer to provide a wide selection of goods that satisfy client demand and increase the store's profitability. We will perform the same process but use different algorithms and offline supermarket data.

The paper by Ezhilarasan & Ramani (2017) discusses the use of data mining techniques to predict the performance of a hosted website. It allows for flexible logical reasoning and is effective for predicting outcomes in specific clusters. In summary, the authors used data mining techniques to predict the performance of a hosted website, using website traffic and conversion rate as attributes and fuzzy logic as the technique for data mining and prediction. The goal was to improve the website's performance and achieve the desired outcomes for the online business.

Wu (2018) conducted the research with 550k sales transaction records with 12 features. The data set also contained customer demographics (age, gender, marital status, city type, stay_in_current_city), product details (product_id and product category) and total purchase _ amount from last month. Our study also uses the same 12 features but with a reduced data size of 50k records.

Different machine learning methods were constructed to estimate the purchase amount using multiple linear regression, and their accuracy and performance were compared to choose the best model. Because it's a regression issue, the Root Mean Squared Error (RMSE) loss function is applied.

The Root Mean Squared Error (RMSE) is given by,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Linear regression, MLK classifier, Deep learning model using Keras, Decision tree, Decision tree with bagging, XGBoost algorithms were applied to the dataset and RMSE was calculated to check which model performed better. Below Figure 3 depicts the RMSE comparison of different algorithms.

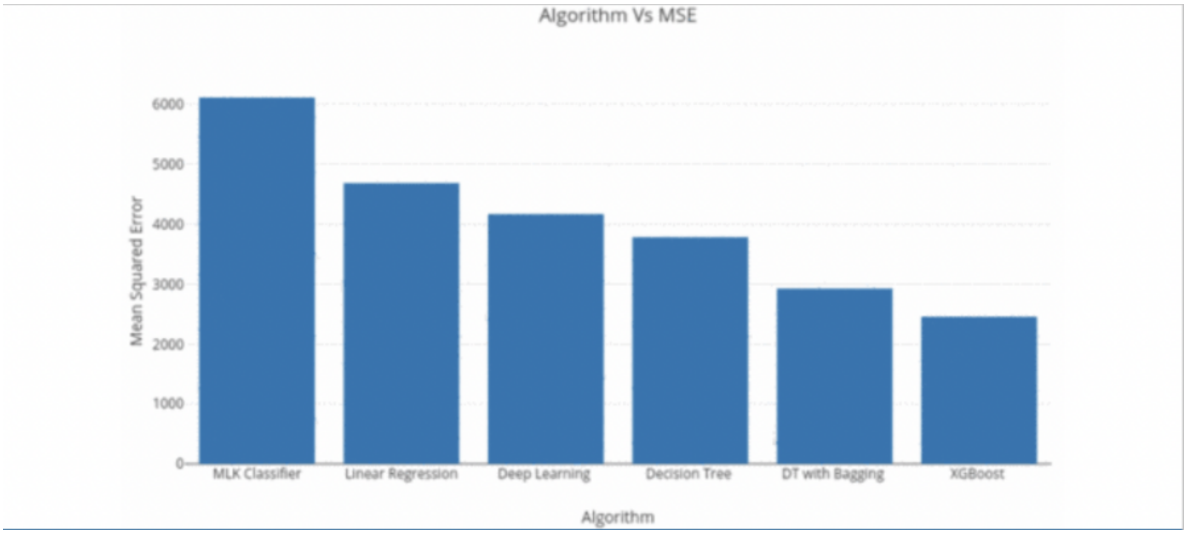

Figure 3: Algorithm vs RMSE. Wu (2018)

The study also finds that pre-processing the data is necessary to provide a useful dataset for creating the prediction model. This study covered a number of methods for getting the optimal model. But there is still no conclusive answer as to what the right method is to produce a model with great accuracy. A dataset with enough features and an increase in size must be obtained in order to improve the findings. It is necessary to perform additional research to improve the current machine learning approaches so they can operate in real time and produce an effective model. In order to assess the effectiveness and scalability of the constructed models, they must also be tested on data of various volumes.

To summarise, Odegua (2020) has performed analysis on very limited number of records. In this project number of records is increased to improve the accuracy. Research by Kaneko and Yada (2016) describes the importance of categories. In this project, we will determine if the categories provide better results. Zhang, et al., (2016), Setiawan, et al. (2017) and Ezhilarasan

& Ramani (2017) have done the analysis on online data. In this project, we will use the same process but for an offline retail store. Wu (2018) used 550k records to perform the analysis. XGBoost algorithm gave the least residual. We will compare our model's results with this.

2.2 Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques

Rajeshwari & Sushma (2021) conducted the research using Black Friday Sales dataset Ostwal (2019) with 550k observations and 12 features. Machine learning algorithms such as Linear regression, Ridge regression, Lasso regression, Decision tree regressor and Random forest regressor are applied. Performance is evaluated using Mean Square Error (MSE). In our study, we will be using the same 12 features but 50k records. Different algorithms such as Support Vector Machine algorithm with radial and polynomial kernel are applied to improve the mean square error.

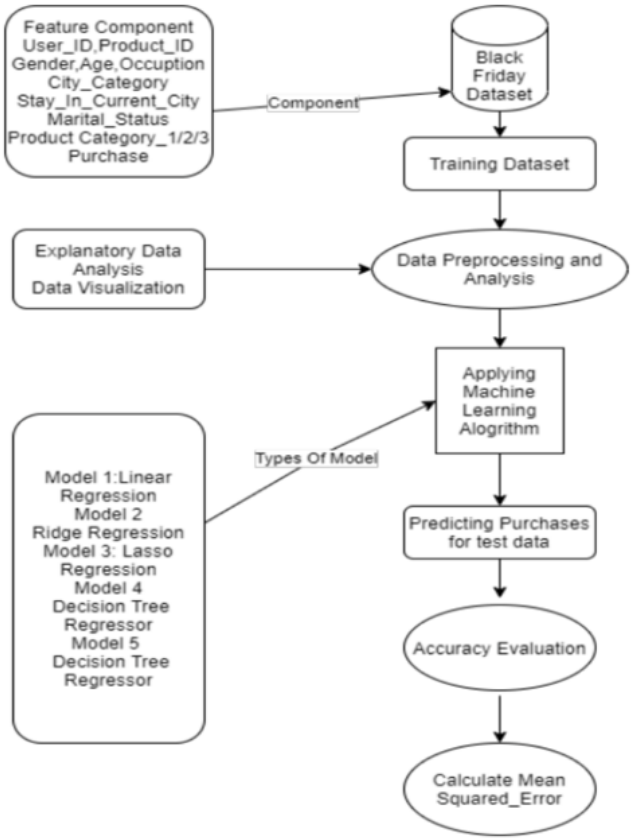

Figure 4: Flowchart of proposed system. Rajeshwari & Sushma (2021)

Exploratory data analysis and data visualisation is done to find any interesting relationship between the features. Data is then pre-processed to remove any outliers, missing values etc and then scaled to apply machine learning algorithms. Various algorithms are applied on the training data. The models are then applied to test data to predict purchase amount. Accuracy is measured using Mean Square Error(MSE).

Model	MSE
Linear Regression	4617.99
Ridge Regression	4687.75
Lasso Regression	4694.14
Decision Tree Regressor	3363.87
Random Forest Regressor	**3062.72**

Figure 5: MSE values for different algorithms. Rajeshwari & Sushma (2021)

Figure 5 depicts the performance of various algorithms on the Black Friday sales data. Their study concluded that Random Forest algorithm works the best as it has the lowest MSE value. Hyperparameter tuning is suggested as an improvement to the model.

In this project, we will perform better data cleaning for 50k records and compare the results.

2.3 Data Analysis and Visualisation of Sales Data

Wajgi (2016) has shown a system that is necessary in today's environment for the analysis and visualisation of sales data so that the organization's owners and investors may make wise decisions and make money. Visualisation's primary objective is to use graphics to connect information in a clear and effective way. To find patterns for potential future forecasts, data mining techniques were used. For analysis and visualisation, a data set from one of the USA's stores was used. Numerous attributes were included in the data collection, including order ID, order date, order priority, sales, customer name, region, product name, product category, and others. The data set provided to the system had a number of attributes, some of which were not pertinent to the end user. As a result, the dataset was cleaned and only the pertinent attributes were exported before saving in the database. The research explains the importance of data cleaning and visualisation to predict sales. In this project, Tableau is used for find any interesting patters in the data and to plot the results.

2.4 Data Visualisation using Tableau

In the field of data analytics, Tableau is a popular free data visualisation software. Data visualization is a tool of data literacy: “Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data” Tableau.com (2019). Data visualisation enriches a presentation in the same manner that a thousand words would while using less space and assisting the researcher in finding relationships in the data Hennessey (2014). The comparative advantage of Tableau is in the usage of numerous large data sets to produce charts in interactive dashboards from which the

user may drill down to find patterns in the data Shaptunova (2017). Additionally, Tableau provides an intuitive drag-and-drop user interface which makes it easy to use Batt, et al., (2020).

Tableau is a multipurpose data visualisation application that is increasingly popular in the data analytics sector than other tools such as GIS, FRED Louis (2019), Infographics or Excel. While GIS focuses on thematic geographic mapping, FRED plots time series data, Excel cleans data, and Infographics draws attention to a particular statistic Rouse (2012), Tableau also offers all of these functions by itself without requiring any other tool.

In this project we will plot the graphs using both Tableau and R programming language and discuss the findings.

2.5 Sales Prediction using machine learning algorithms

The sales of several Big Mart shops have been forecasted using machine learning methods like Linear Regression, K-Nearest Neighbors, XGBoost, and Random Forest. For each of the four methods, many metrics that affect the accuracy of results were tabulated, including Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracy, and more. With an accuracy of 93.53%, the Random Forest Algorithm is discovered to be the best suitable of all Bajaj & Purvika (2020).

In one the studies, the use of deep learning techniques is to foresee their sales plan for electronic components. To increase the system's effectiveness, other optimization techniques are also applied: such as a genetic algorithm Baba & Suto (2000).

In a research conducted by Maharajan (2019), data is cleansed, pre-processed, and defined in terms of the attributes of the customers before applying the algorithm. The Apriori algorithm is applied with the least amount of confidence and support. The Apriori algorithm's created rule is used for evaluation. The paper suggests that accessibility, simplicity of use, and perceived rewards are the main variables that influence consumers' online purchasing decisions. We will check if this prediction holds true for our dataset.

Machine learning is now the primary strategy for resolving the AI problem due to the existence of artificial intelligence Fradkov (2020). One of the business analytics backed by machine learning approaches that improves an organization's efficiency is sales prediction. A master of machine learning is essential to the business area because sales prediction is an interdisciplinary field. One of the key components of business planning is now sales forecasting Yin (2020).

One machine learning strategy is the decision tree. The assessment uses the algorithms ID3, C4.5, and C5.0. A Decision Tree will divide one node into two or more sub-nodes in a tree-like manner. The results of each sub-node are represented by a previous sub-node. The end nodes show the outcome of a classification Abdulkareem (2021).

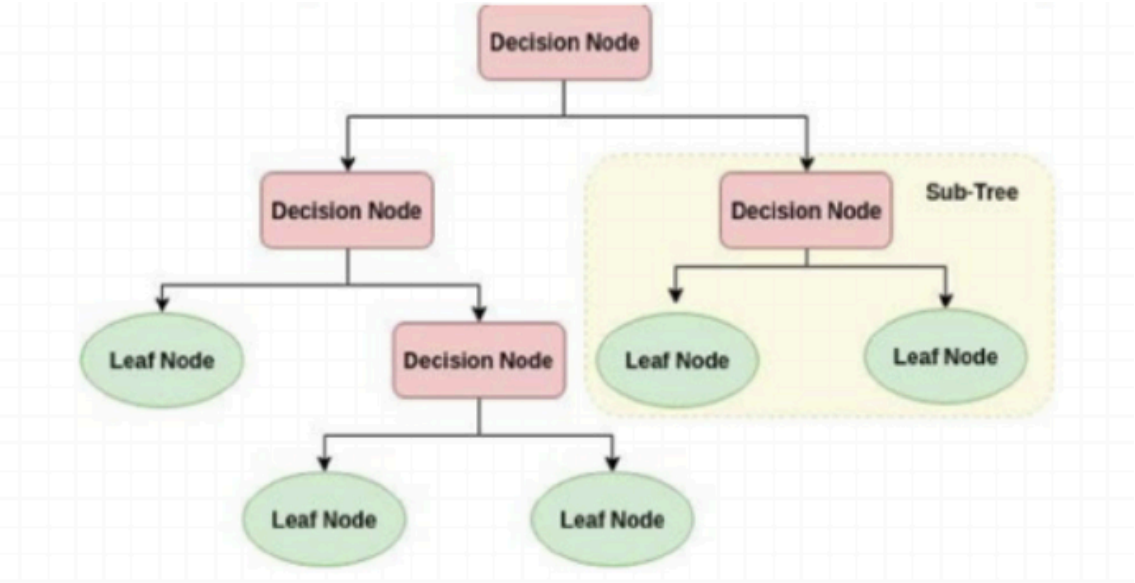

Figure 6: Structure of Decision Tree. Mao (2022)

Classification and Regression Tree is a non-traditional approach to analysing data that relies on binary splits to partition the dataset and predict the value of y based on the value of x, using a greedy algorithm to grow a tree and prune it to prevent overfitting. This method involves sequentially growing the tree based on split criteria in order to fit the standard node as closely as possible. An alternative method for analysing data is the Bayesian approach, which involves using probability theory to make predictions about the likelihood of certain outcomes based on past events. The Bayesian rule in statistics refers to a method of adjusting one's subjective judgment about the probability of an event occurring based on the probability distribution given an observation Chipman, et al., (1998). As the number of analytical samples approaches the size of the entire population, the probability of an incident occurring among the samples tends to align with the probability of it occurring in the overall population.

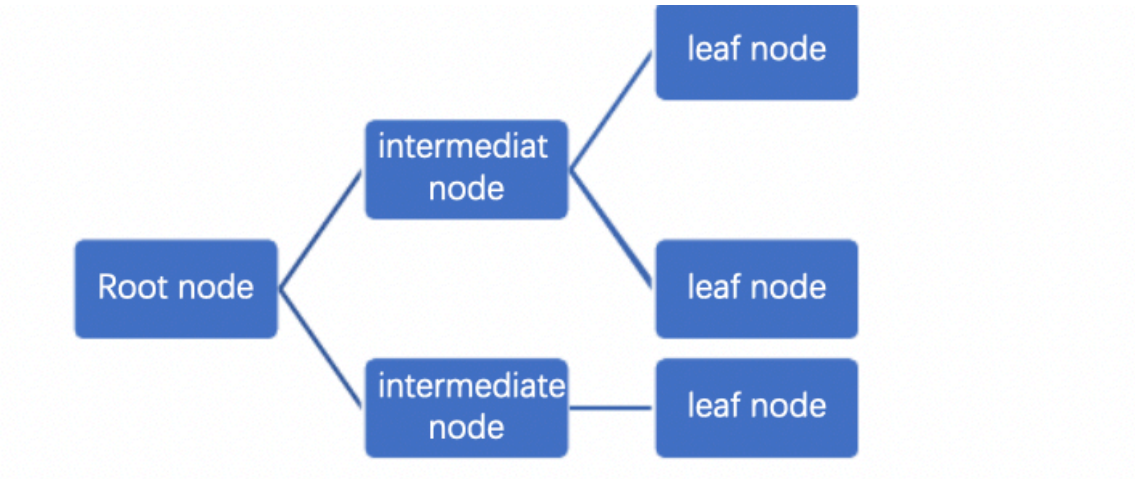

Figure 7: A sketch of CART. Mao (2022)

Gradient boosting is a technique that involves combining multiple weak learning models, also known as an ensemble, in order to build a model that can make predictions. The idea behind boosting is that if there is a polynomial learning algorithm that can accurately learn a concept, it is considered strongly learnable. On the other hand, if there is a polynomial learning

algorithm that can learn a concept, but the accuracy is low, it is considered weakly learnable. The XGBoosted tree model consists of three phases: data processing, feature selection, and training and prediction Xia, et al., (2020). During the data processing phase, the original dataset is analysed and cleaned by removing any defaulted data and performing preliminary processing. In the feature selection phase, the technique of feature engineering is used to train and predict the features based on one-hot coding. During the training and prediction phase, the XGBoosted model uses the selected features and adjusts its parameters in order to improve the accuracy of its predictions. Finally, the XGBoosted model itself makes the final prediction using the most accurate model it has developed.

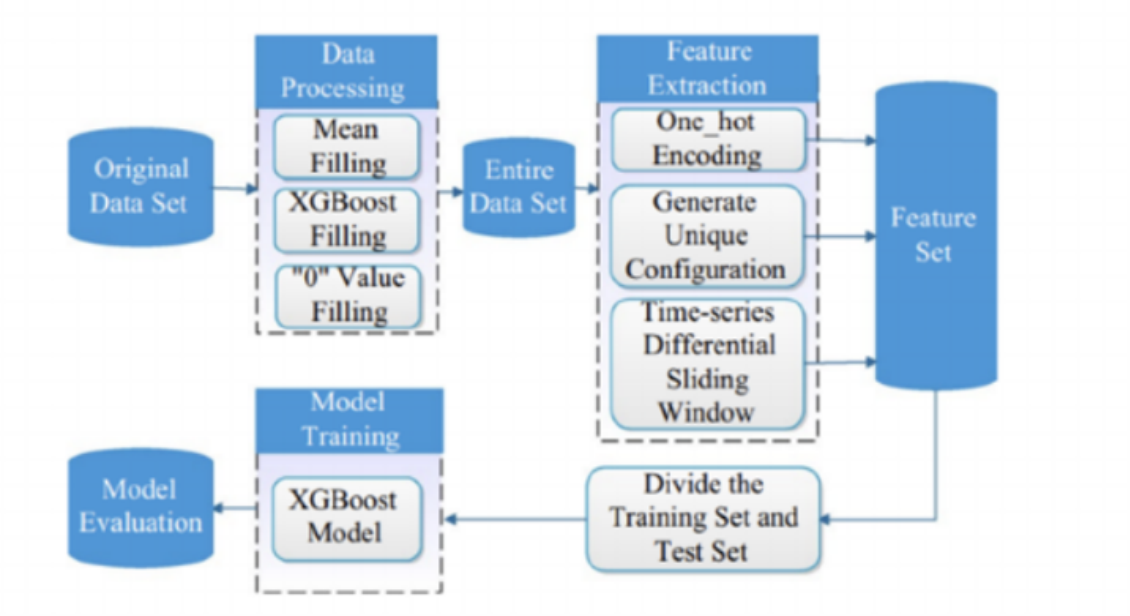

Figure 8: System architecture of XGBoosted tree. Sharkawy (2020)

A neural network is a mathematical model that is inspired by the structure and function of the neural network found in animals. It consists of a hierarchical directed graph with multiple nodes in each layer. With advances in computer processing technology, neural network models are able to handle a large number of features in machine learning models. When using a neural network, users input a series of values for various features and the network calculates the process to determine which category the new feature should be classified in Sharkawy (2020).

Additionally, neural networks have the ability to learn from representative datasets and make predictions about new data in the future. These networks are able to solve problems involving basic images relatively quickly and more complex images over a longer period of time. Generally, neural networks have three layers: the input layer, the hidden layer, and the output layer. The input layer consists of all the independent variables, the output layer typically consists of a single neuron that represents the dependent variable y, and the hidden layer is a set of nodes that receive weighted signals from the previous layer and transform them using an activation function Miller (1992).

Convolutional Neural Networks (CNNs) are a major advancement in the field of computer vision O'Shea (2015). They use a mathematical operation called convolution in their structure,

rather than the matrix multiplication used in traditional neural networks. Because CNNs operate on two-dimensional data structures, they are better able to understand and recognize patterns in images and speech than other machine learning structures. CNNs are made up of convolutional layers, pooling layers, and dense layers. The convolutional layer applies a convolution operation to the image data input to the network, using grids to scan the data and convert it into a three-dimensional structure containing information about the red, green, and blue channels of the image. There are hundreds of neurons in a CNN that are able to recognize patterns in the data and produce an activation map, which highlights the areas of the image that contain features of interest. Specific neurons in the network are then activated based on the presence of these features Mao (2022).

The study conducted by Wang, et al., (2019) focuses on the use of Artificial neural network for sales prediction. The first ever neuron like perceptron system was developed in 1969, and the artificial neural networks were created in the 1950s Minsky & Papert (1969). The original hypothesis, however, was not considered credible until Hopfield introduced the contemporary ANN version in 1982 because it was too basic. Following this, a wide range of ANNs have been developed that incorporate new architectures and theories, and the tremendous growth in computer power has made it possible for modern, potent ANNs to be extensively employed with a large range of useful ANN applications in a variety of industries. On the basis of back-propagation ANNs, a number of market sales prediction models have been suggested. These models include better ways to account for recognised back-propagation shortcomings, particularly for issues with agnostic rules.

For major wholesalers in Victoria, Kong and Martin Kong & Martin (1995) employed back-propagation networks (BNPs) to forecast upcoming food sales. They demonstrated that, when compared to conventional methods and market analysis, such as a basic linear regression algorithms, the novel BPN method produced improved prediction results. Consequently, the BPN model might be a helpful system for sales forecasting. Thiesing et al., (1997) employed an ANN trained with the help of BPN to forecast time series predictions, such as weekly consumption for grocery store items, while taking into account costs, promotional campaigns, and seasonal influences.

Retail outlets play a crucial role in the retail sector by transferring products from suppliers to customers. While other retail outlets offered an abundance of the same products, some were out of stock of the items customers wanted. Nevertheless, regardless of how outstanding the service was, if wanted goods were out of stock, customer satisfaction was always adversely harmed. As a result, managing orders and inventory had become a top concern for convenience shop management. In order to monitor orders and manage inventories in retail outlets based on operational parameters of the economic circle and sales projection, Chen et al., (2010) developed an ANN-based system. The suggested approach increased order and discard rates to better guarantee that the correct products and the appropriate volumes were ordered. To forecast sales of oral care items and error rates for a variety of products, Vhatkar & Dias (2016) took into account a number of sales prediction methods and created an opposite transfer ANN system. To increase the accuracy of daily average rice sales projection in supermarkets, Mo et al. suggested an enhanced BPN technique. Weron (2014) suggested projecting the price of electricity. This article tries to describe the problems of the solutions that are currently accessible, their advantages and disadvantages, and the risks and possibilities that forecasting tools present or might experience. Cincotti et al., (2014) examined Italian Power Exchange electricity spot pricing using an ANN.

Even though the Neural Network algorithms provide good results, it is complex to implement and takes a lot of time for large datasets. The study by Bajaj & Purvika (2020) proved Random Forest algorithm to have better result than other algorithms. In this project, we will check if RF algorithm works better than others.

2.6 A multivariate intelligent decision-making model for retail sales forecasting

The MID model combines the strengths of different intelligent methods, including the flexibility of artificial neural networks (ANNs) and the interpretability of decision tree (DT) models. The authors applied the MID model to a real-world sales forecasting problem in a retail company and compared its performance with that of ANN and DT models (Guo, et al., 2013). The results showed that the MID model worked better than the other two models with respect to prediction accuracy, which demonstrates the effectiveness of the hybrid intelligence approach in sales forecasting Wong (2010).

This study implies that there are m input features, n pairs of multi-input-single-output (MISO) samples are provided, and that the faster sales of a consumer goods and other affecting factors are candidate input features. (X_i, y_i) represents the ith input/output data pair $(1 \leq i \leq n)$. The research finds the relationship between m input features $(x_{i1}, x_{i2}, \ldots, x_{im})$ of X_i and the total sales volume y_i and forming a subset of input variables from X_i and then forms an constructive system to approximate the associations with respect to n given datasets. Finally, based on early sales of other consumer products and associated input variables, the created model is used to anticipate the volume of sales for those products. To put the above processes into practise, a multivariate intelligent decision-making (MID) system is created.

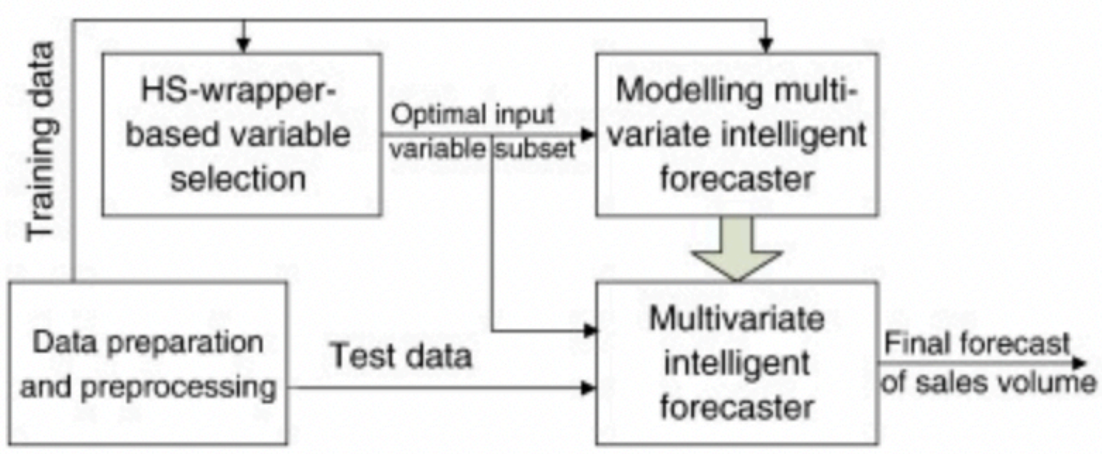

Figure 9: Structure of the MID model. Guo, et al., (2013)

A data preparation and pre-processing (DPP), an HS-wrapper-based variable selection (HWVS), and a multivariate intelligent forecaster (MIF) module are the three modules that make up the proposed MID model's architecture, which is depicted in Figure 9. In order to create and validate the MID model, the DPP module pre-processes sales information for previous goods from retailer point-of-sale databases. The HWVS module chooses the best input feature subset from a list of possible input features in order to eliminate unnecessary and irrelevant ones and to better grasp the fundamental properties of sales data. The MIF is then created using the subset of the chosen input features to simulate the link between the chosen

input features and the retail product sales volumes. Last but not least, the developed MIF is used to project the total retail product sales volumes.

This study examines the issue of early sales-based retail forecasting in the retail sector, which is beneficial for associated retail firms to provide accurate replenishment forecasts and enhance the effectiveness of their retail supply chains. To assess the suggested MID model against a wide range of typical datasets from real-world retail data, numerous experiments were carried out. The experimental results showed that the suggested model could perform significantly better than the generalised linear model in addressing the researched multivariate sales forecasting problem. The study also suggests that performing feature selection would improve the accuracy of the model. In this project, we will perform feature selection to improve the results.

2.7 A Big Data Approach to Black Friday Sales

Awan, et al., (2021) have applied machine learning algorithms such as linear regression and random forest using big data technology. For managing a vast volume and variety of data as well as for doing real-time analysis, there are numerous big data frameworks available. Because it stores massive datasets in memory and performs distributed computation on them, the Apache Spark clustered computing framework is significantly faster and more dependable than Apache Hadoop.

All classification, regression, and clustering model used in conventional machine learning are included in spark machine learning Meng, et al., (2016). An open-source, distributed, scalable, and platform-independent machine learning library is Apache Spark MLlib 2.0 Assefi, et al., (2017). On clusters, MLlib operates in parallel. Additionally, it has access to all programming languages and all machine learning models Jumaa & Omar (2019).

Evaluation	Linear Regression	Random Forest
Machine Learning	68% (R2)	74% (Precision)
Spark ML	72% (R2)	81% (Precision)

Figure 10: Evaluation results of machine learning vs. Spark MLlib for linear regression and random forest. Awan, et al., (2021)

From Figure 10, we can see that on Spark, the random forest model delivered 81% accuracy whereas single node machine learning only delivered 74%. The accuracy for linear regression on Spark is 72% which is greater than single node machine learning which delivered 68% accuracy. In this project, we will use R programming language instead of Python and check if the model performs better.

2.8 Two-Level Statistical Model for Sales Prediction

In the research done by Punam, et al., (2018), the ensemble technique known as stacking, or stacked generalization, is a two-level approach to using data mining and predictive techniques. It is called two-level because it involves stacking multiple learning algorithms in two layers: a bottom layer with one or more algorithms and a top layer with a single algorithm Liu (2009). The top layer algorithm is trained to combine the predictions made by the algorithms in the bottom layer. To use stacking, the bottom layer algorithms are first trained using a dataset, and then the top layer combiner algorithm is trained using the predictions made by the bottom layer algorithms to produce a final prediction. Stacking tends to perform better than a single model because it incorporates the predictions of multiple algorithms and considers multiple aspects of the dataset Enko (2005). Mean Absolute Error (MAE) is used to evaluate the accuracy of predictive models. It is specifically used to measure the accuracy of continuous variables by calculating the average absolute error of a set of predictions, regardless of their direction. Absolute error is the difference between the predicted value and the actual value. MAE is a useful metric for assessing the accuracy of a model and determining how well it is able to make predictions.

To obtain experimental results, each regression model made use of a cross validation model with ten folds to evaluate its predictive significance. In cross validation technique, the dataset is randomly split into 10 subsets, with approximately same sizes. 90% of the data that is, 9 subsets were used as training partition and the left over 10% data was used for testing partition. The regression model is trained using the training data and the test data is used to measure the model's predictive accuracy. This process is repeated until each subset has been used as test data once. This allows for a more thorough evaluation of the model's performance. In this project we will split the data in 70:30 ratio and check the accuracy using both RMSE and MAE values.

2.9 Evaluating prediction accuracy using RMSE and MAE

Root Mean Square Error (RMSE) depicts the standard deviation of residual predictions. The presence or absence of the regression line data points is gauged using these residuals. RMSE is a measurement of the distribution of these residuals Kvalheim (2018). They are commonly used in forecasting problems to check for accuracy.

$$RMSE_{f0} = \left[\sum_{i=1}^{N} \left(Z_{fi} - Z_{0i}\right)^2 \ /N\right]^{1/2}$$

Where Σ = Summation,

$\left(Z_{fi} - Z_{0i}\right)$ = differences squared,

and N = Sample size. Ramachandra, et al., (2021)

Mean Absolute Error (MAE) is another function used to calculate average error or residual. MAE is said to be unambiguous when compared to RMSE Chai & Draxler (2014). MAE is given by,

$$MAE_{f0} = \sum_{i=1}^{N} |Z_{fi} - Z_{0i}| / N$$

Where Σ = Summation,

$(Z_{fi} - Z_{0i})$ = absolute differences,

and N = Sample size. Ramachandra, et al., (2021)

In conclusion, we will be using Random forest, Multiple linear regression algorithms as mentioned in sections 2.1, 2.2 and 2.5. SVM algorithm is applied in addition to see if the model performs better. We will also use Tableau and R as mentioned in sections 2.3 and 2.4. In contrast to section 2.8, this project will have a single level model with dataset split in the ratio 70:30. RMSE and MAE values are calculated as shown in section 2.9 to evaluate the models. All the findings in section 2 will be compared with actual results obtained from this project on section 5.

3 Research Methodologies

3.1 Introduction

This section aims to describe the data used to build the model, relationship between the variables and methodology applied to achieve the results. Initial data exploration is performed to find any strong relationships between the variables and to find any interesting pattern in the purchase behaviour of a customer. Data is then cleaned by checking for redundant or missing values, outliers etc. Once the data is cleaned and scaled, machine learning algorithms are applied. Finally, model outcomes are compared to check which model gives the best accuracy. CRISP-DM (CRoss Industry Standard Process for Data Mining) approach is followed to perform this research.

3.2 Research Framework

A business model that creates the basic structure of a data science system is the CRoss Industry Standard Process for Data Mining (CRISP-DM) Hotz (2022). CRISP-DM methodology comprises of 6 iterative phases from business analysis to deployment. Figure 11 depicts the 6 stages of CRISP-DM methodology.

Figure 11: CRISP-DM process. Brian (2021)

In order to help the business plan and carry out a data mining project, CRISP-DM provides a roadmap, best practises, and frameworks for better and faster results while employing data mining techniques.

Business understanding is the first step in the data mining process and involves understanding the business needs and translating them into a clear problem definition. The data understanding phase involves familiarizing oneself with the data, identifying any issues with data quality, gaining preliminary insights into the data, and identifying interesting subsets to generate hypotheses about hidden information. The data preparation phase involves selecting relevant tables, records, and attributes, as well as transforming and cleaning the data to make it suitable for use in modelling tools. In summary, the first three phases of the data mining process focus on understanding the business requirements, getting to know the data, and preparing the data for modelling (Nadali, 2011).

The creation of the model and choosing of the modelling approach include the data modelling step. Any data mining technique can be applied. The decision generally depends on the data and the business situation. Specific parameters must be set in order to build the model. It is appropriate to compare the model to the evaluation criteria and choose the best ones for assessment. In the project's evaluation phase, a model that seems to be of high quality from the standpoint of data analysis is built. It is vital to do a more complete evaluation of the model and assess the procedures used to develop the model before moving forward with its final deployment to ensure that it correctly accomplishes the business objectives. The user guide provides an overview of the deployment phase. A software component or a final report could be involved. According to the user manual, the deployment phase entails planning, monitoring, and maintenance (Schröer, 2021).

3.2.1 Business Understanding

Black Friday marks the beginning of winter shopping season. Many products will be on sale and hence customers tend to shop more which results in a crowded store. But managing the crowd with a little staff and attracting potential clients is equally harder for the store owners. Many approaches have been used to address this issue, but they haven't been very effective. A method that has shown promise in resolving the issue is a prediction model (Wu, 2018). This research focuses on the area of prediction models to create a precise and effective algorithm to assess customer spending in the past and produce the customer's future spending with the same variables. Initially I will perform data visualisation and data pre-processing to make the data ready for modelling. I will be using Tableau and R programming to plot the graphs. After visualisation, I will perform data pre-processing to make the data ready for modelling. I will then apply machine learning techniques such as multiple linear regression, random forest and SVM. The results are then compared based on the accuracy of the prediction.

3.2.2 Data Understanding

For this study, I will be using the dataset from Kaggle provided by Sammari (2020). The dataset consists of sales transactions that were recorded at a retail store. The dataset is a CSV file with 550,069 observations and 12 features. I will be using 50,000 records with 12 variables for my research. The dataset consists of both numerical and categorical data. I will be splitting the data into training and testing in the ratio 70:30. Purchase is the dependent variable which will be taken out from the testing partition. Product_ID, User_ID, Gender, Age, Occupation, City_Catrgory, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3 are the independent variables. Occupation is encoded with numbers 0 to 20 representing respective occupations. Figure 12 depicts the data type of each variable.

```
'data.frame':   50000 obs. of  12 variables:
 $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 10
00004 1000005 ...
 $ Product_ID                : chr  "P00069042" "P00248942" "P00087842" "P00085442" ...
 $ Gender                    : chr  "F" "F" "F" "F" ...
 $ Age                       : chr  "0-17" "0-17" "0-17" "0-17" ...
 $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
 $ City_Category             : chr  "A" "A" "A" "A" ...
 $ Stay_In_Current_City_Years: chr  "2" "2" "2" "2" ...
 $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
 $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
 $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
 $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
 $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
```

Figure 12: Data type of all variables in the data frame

Each of the variables in the dataset are explained below,

User_ID is the unique identification number given to each customer. The dataset includes a total of 5891 users.

Product_ID is the unique alphanumeric id given to each product. The dataset consists a total of 3623 products.

Gender indicates the gender of the user making a transaction. Male is represented by M and Female is represented by F.

Age indicates the age group of the user making a transaction. A user can fall into one of the following categories; 0-17, 18-25, 26-35, 36-45, 46-50, 51-55 and 55+.

Occupation indicates the job of the user pre-defined with numbers 0 to 20. The values of these encoded numbers are masked.

City_Category gives information about where the user is currently living. It is categorised into 3 parts namely A, B and C.

Stay_In_Current_City_Years indicates for how many years the user has lived in the current city. The value ranges are 0, 1, 2, 3, and 4+.

Marital_Status indicated whether a user is married or not. If the user is married, it is indicated with 1 else 0.

Product_Category_1 to _3 describes the category of the product. Each product purchased falls under one of the three categories. The product category values are masked.

Purchase is the dependent variable which indicates the amount spent by the customer on a product.

3.2.3 Data pre-processing

Raw data may contain many duplicate pieces of information, as well as missing or inconsistent values. The three categories of most frequent issues with raw data are as follows:

Missing information can also be viewed as erroneous information because it leaves gaps that could affect the outcome of the study. Missing data frequently shows up when there is an issue during the data collection phase, such as a bug that caused a system outage, errors in data entry, or problems with the usage of biometrics, among others.

Erroneous data and outliers that are present in the data collection but have no practical application are included in the category of noisy data. Human error, unusual exceptions, incorrect labelling, and other problems during data collection can all contribute to noise.

Data inconsistencies occur when files containing the same information are kept in several formats and files. Data inconsistency introduced by duplicates in various formats, typos in name codes, or a lack of data constraints frequently results in inconsistent data that must be corrected before analysis (Miller, 2019). Data cleaning, data transformation and data reduction are some of the techniques involved in data pre-processing which makes the data organised and cleaned before training the model.

3.2.3.1 Handling missing values

I will first plot a graph to check for missing values in the dataset. Figure 21 depicts the graph of NA values in all the variables in the dataset.

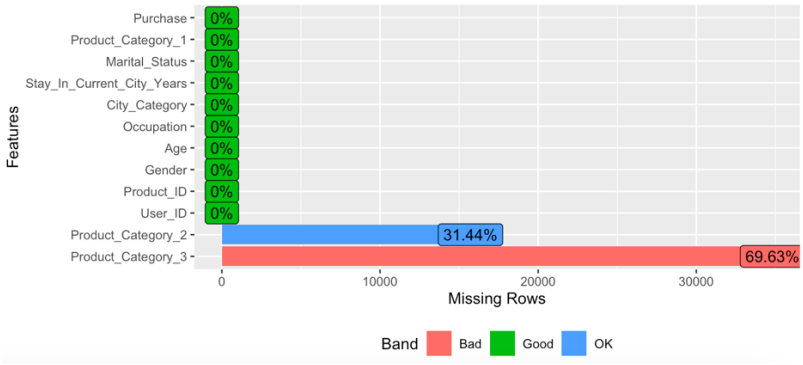

Figure 25: Missing values in the dataset

From figure 25, we can see that Product_Category_2 and Product_Category_3 has missing values. Product category has almost 70% of missing values which implies that only 30% of data is available in that category. Product_Category_2 has around 31% of missing values indicating that 69% of the data in this category is valuable. Missing values can be handled in many ways. Option one is to drop the variable entirely, second option is to calculate the mean value of the variable and assign it to the missing rows and the last option it to replace NA values with 0.

Since product category 3 has almost 70% missing data, I will drop this variable. Replacing NA values with 0 or with arithmetic mean does not give much insight in this case as most of the data will be fabricated which will make the model less accurate.

Product category 2 has around 31% missing values which is much better when compared to product category 3. To handle missing values in this variable, I will calculate arithmetic mean of the variable Product_Category_2 and assign it to all missing values. Dropping the variable will not make sense in this case as we will be losing 69% of good data that will help us in prediction.

3.2.3.2 Data Transformation

The dataset used for this study consists of both numerical and categorical data. The character variables needs to be converted to factors to give a meaningful insight. The below figure 22, shows the summary of all variables after factoring categorical variables.

```
> summary(newdata_train)
    User_ID           Product_ID    Gender       Age          Occupation     City_Category
 Min.   :1000001   P00265242:  159   F:12188   0-17 : 1409   Min.   : 0.000   A:14049
 1st Qu.:1001015   P00025442:  156   M:37812   18-25: 9615   1st Qu.: 2.000   B:20721
 Median :1002103   P00112142:  154             26-35:19731   Median : 7.000   C:15230
 Mean   :1002553   P00117442:  143             36-45: 9847   Mean   : 8.143
 3rd Qu.:1004072   P00110742:  139             46-50: 4006   3rd Qu.:14.000
 Max.   :1006040   P00059442:  136             51-55: 3490   Max.   :20.000
                   (Other)  :49113             55+  : 1902
 Stay_In_Current_City_Years Marital_Status   Product_Category_1 Product_Category_2
 0 : 6987                   Min.   :0.0000   Min.   : 1.000     Min.   : 2.000
 1 :17374                   1st Qu.:0.0000   1st Qu.: 1.000     1st Qu.: 5.000
 2 : 9102                   Median :0.0000   Median : 5.000     Median : 9.000
 3 : 8730                   Mean   :0.4098   Mean   : 5.305     Mean   : 9.869
 4+: 7807                   3rd Qu.:1.0000   3rd Qu.: 8.000     3rd Qu.:15.000
                            Max.   :1.0000   Max.   :18.000     Max.   :18.000
                                                                NA's   :15721
 Product_Category_3    Purchase
 Min.   : 3.00      Min.   :  185
 1st Qu.: 9.00      1st Qu.: 5852
 Median :14.00      Median : 8045
 Mean   :12.71      Mean   : 9279
 3rd Qu.:16.00      3rd Qu.:12033
 Max.   :18.00      Max.   :23958
 NA's   :34817
```

Figure 26: Summary of the data

From figure 26, we can see the count of each category of a particular categorical variable. It also describes minimum, maximum, mean, median, first quarter and third quarter values.

Variable User_ID just gives information about unique id assigned to the customer. It does not play any role in predicting sales and hence I will be dropping this variable.

Variable Product_ID gives information about unique ids assigned to the product. It does not help us in predicting sales. So, I will be dropping this variable as well.

We need to find the correlation between all the variables in order to determine their importance. Although we have converted the categorical data into factors, we cannot plot correlation plot since it's not numerical. To convert factors into numerical data I have done encoding as below.

In variable Gender, Female is encoded as 1 and male as 2.

In variable Age, age groups of 0-17, 18-25, 26-35, 36-45, 46-50, 51-55 and 55+ are encoded as 1, 2, 3, 4, 5, 6 and 7 respectively.

In variable City_Category, City A, City B and City C are encoded as 1, 2 and 3.

In variable Stay_in_Current_City_Years, years 0, 1, 2, 3 and 4+ are encoded as , 1, 2, 3, 4 and 5.

We use the glimpse function to check the data type of the variables. Figure 27 depicts the output of glimpse function.

```
Rows: 50,000
Columns: 9
$ Gender                    <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 2,…
$ Age                       <dbl> 1, 1, 1, 1, 7, 3, 5, 5, 5, 3, 3, 3, 3, 3, 6, 6, 6, 6, 4, 3,…
$ Occupation                <int> 10, 10, 10, 10, 16, 15, 7, 7, 7, 20, 20, 20, 20, 20, 9, 9, …
$ City_Category             <dbl> 1, 1, 1, 1, 3, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3,…
$ Stay_In_Current_City_Years <dbl> 3, 3, 3, 3, 5, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 5,…
$ Marital_Status            <int> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1,…
$ Product_Category_1        <int> 3, 1, 12, 12, 8, 1, 1, 1, 1, 8, 5, 8, 8, 1, 5, 4, 2, 5, 1, …
$ Product_Category_2        <dbl> 9.869308, 6.000000, 9.869308, 14.000000, 9.869308, 2.000000…
$ Purchase                  <int> 8370, 15200, 1422, 1057, 7969, 15227, 19215, 15854, 15686, …
```

Figure 27: Output of glimpse function after converting factors into numeric variables

We can plot the correlation graph as the dataset is now completely numeric.

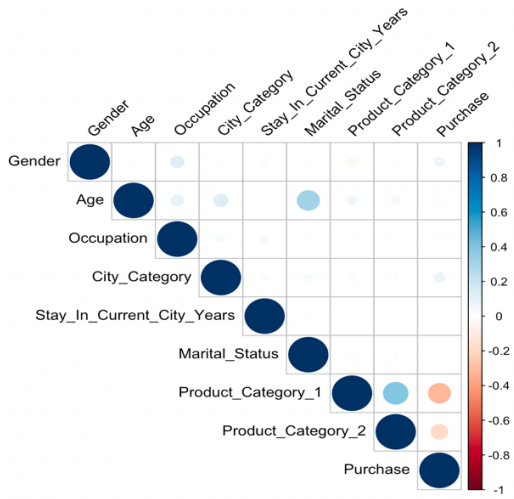

Figure 28: Correlation plot between all variables

From figure 28, we can see that age and marital status has a good positive relationship. Product_Category_1 and Product_Category_2 also positive relationship. We can also see significant negative relationship between Product_Category_1 and Purchase. There is a good negative association between Product_Category_2 and Purchase as well. Including these variables in the model gives us good results.

3.2.4 Modelling

After the data preparation phase, the selected modelling techniques are applied to the prepared data. The performance of the resulting models is then evaluated using evaluation metrics. These metrics can be chosen based on the specific characteristics of the data and the goals of the data mining project. For example, in a classification problem, the accuracy of the model could be used as an evaluation metric, while in a regression problem, the root mean squared error (RMSE) or mean absolute error (MAE) could be used. Modelling stage of this study in described in detail in section 4.

3.2.5 Evaluation

Once the model is created, it is important to evaluate its performance and effectiveness. This may involve comparing the model's results to actual outcomes, or using various evaluation metrics such as accuracy, precision, and recall to assess its performance. If the model is not performing as expected, it may be necessary to go back to earlier stages of the process, such as data preparation or model selection, to make improvements. Evaluation stage is described in detail in section 5.

3.2.6 Deployment

Once the model is performing satisfactorily, it can be deployed in the real world and used for its intended purpose. It is important to continuously monitor the model's performance and make any necessary updates or adjustments to ensure its continued effectiveness Yun, et al., (2014).

These are the 6 phases involved in Crisp-DM methodology. Modelling and Evaluation phases will be explained in further sections 4 and 5 respectively.

4 Modelling

The data modelling phase involves choosing a modelling technique, creating a test case, and building the model. Any data mining technique can be used, but the choice should be based on the business problem and the data available. It is also important to explain the reasoning behind

the choice of technique. When building the model, specific parameters must be set. To evaluate the model, it is important to assess it against evaluation criteria and select the best-performing model Schroer, et al., (2021).

During the evaluation phase, the results of the data modelling are compared to the defined business objectives to determine their effectiveness. The results must be interpreted and further actions must be defined based on the findings. Additionally, the overall process should be reviewed to identify any potential improvements or areas for optimization.

4.1 Multiple Linear Regression

Regression analysis is a statistical method used to identify the relationship between a dependent variable (the target) and one or more independent variables (the predictors). It is used for forecasting, time series modelling, and determining the causal relationship between variables. There are various types of regression, which are typically classified based on the shape of the regression line, the type of dependent variable, and the number of independent variables Wu, et al., (2018). Regression analysis is a powerful tool for making predictions and understanding the relationships between variables. Linear Regression is a statistical method that involves finding the best fit straight line (called the regression line) that represents the relationship between a dependent variable (Y) and one or more independent variables (X). The relationship is represented by the equation Y= a+b∗X+e, where a is the intercept, b is the slope of the line, and e is the error term. This equation can be used to predict the value of the dependent variable based on the value of the independent variable(s).

The extension of simple linear regression to incorporate many explanatory variables is known as multiple linear regression Tranmer (2008). Because we presume that the answer variable is directly related to a linear combination of the explanatory factors, we continue to use the term "linear" in both linear and multiple regression.

Even though the equation for multiple linear regression has more terms, it has the same basic structure as the equation for simple linear regression as shown below,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

where β_0 represents the intercept (the average of Y when all $X_k = 0$), and each β_k depicts a slope in terms with X_k (the effect of change in the average of Y when X_k is greater by one unit and all other predictor variables are kept unchanged). When all explanatory variables are 0, the predicted value of y in the simple situation is 0, which is represented by the constant 0. In a model with p explanatory variables, each explanatory variable has a β_coefficient Tranmer (2008). The analysis gives us the opportunity to look at the relationships between a set of explanatory variables and an interesting response variable, but it does not allow us to draw conclusions about causes.

The assumptions for multiple linear regression are as follows:

1. the y_i are exclusive of each other;
2. the y_i has a normal distribution;
3. the average of that distribution is a linear function of each x_{ik}; and

4. the variance of that distribution is equal for all y_i (constant variance, or *homoscedasticity*) Eberly (2007).

4.2 Applying Multiple Linear Regression

Linear regression algorithm is applied on the training dataset with all variables after data pre-processing. The dataset consists of 33334 observations with 9 variables. User_ID, Product_ID and Product_Category_3 variables have been removed in the previous stage. Below figure 29, depicts the output of the linear regression model with all 9 variables. We will call this model as model1. We can see that Gender, Age, City_Category, Product_Category_1 and Product_Category_2 have good significance value. It has a R-squared value of 0.1195 and Adjusted R-squared value of 0.1191 which represents approximately 12% variance. The standard residual error is 4638. We can also observe that Stay_In_Current_City_Years is very slightly significant and Occupation and Marital_Status do not show any significance. To improve the model, least significant variables are removed one by one and check the output.

```
Call:
lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
    Stay_In_Current_City_Years + Marital_Status + Product_Category_1 +
    Product_Category_2, data = train_dataset1)

Residuals:
     Min       1Q   Median       3Q      Max
-10292.7  -3086.7   -622.9   2260.0  17833.4

Coefficients:
                            Estimate Std. Error t value             Pr(>|t|)
(Intercept)                10994.280    185.098  59.397 < 0.0000000000000002 ***
GenderM                      500.467     59.873   8.359 < 0.0000000000000002 ***
Age18-25                     133.526    162.332   0.823              0.41077
Age26-35                     330.326    158.139   2.089              0.03673 *
Age36-45                     440.301    163.077   2.700              0.00694 **
Age46-50                     401.904    180.127   2.231              0.02567 *
Age51-55                     871.687    183.584   4.748           0.00000206 ***
Age55+                       556.337    202.376   2.749              0.00598 **
Occupation                     1.785      3.916   0.456              0.64843
City_CategoryB               274.875     62.647   4.388           0.00001149 ***
City_CategoryC               760.061     67.751  11.218 < 0.0000000000000002 ***
Stay_In_Current_City_Years1  144.315     81.053   1.781              0.07500 .
Stay_In_Current_City_Years2  152.033     90.517   1.680              0.09304 .
Stay_In_Current_City_Years3  115.329     91.538   1.260              0.20771
Stay_In_Current_City_Years4+ 161.124     94.230   1.710              0.08729 .
Marital_Status               -64.295     55.102  -1.167              0.24328
Product_Category_1          -401.992      7.508 -53.545 < 0.0000000000000002 ***
Product_Category_2           -80.209      6.587 -12.177 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4638 on 33316 degrees of freedom
Multiple R-squared:  0.1195,    Adjusted R-squared:  0.1191
F-statistic: 266.1 on 17 and 33316 DF,  p-value: < 0.00000000000000022
```

Figure 29: Linear regression model with all variables

In figure 30, linear regression model is applied by removing Stay_In_Current_City_Years variable. We can observe that Adjusted R-squared value is still the same as model1 which is 0.1191. Since the Adjusted R-squared value is not decreased, we can be sure that the model is not deteriorating. We can also see that Occupation and Marital_Status is still not significant. The significance of other variables have not changed either. We will call this model as model2 and to improve it, we will remove insignificant variables.

```
> model2 <- lm(Purchase ~ Gender + Age + Occupation + City_Category + Marital_Status + Product_Categor
y_1 + Product_Category_2, data = train_dataset1)
> summary(model2)

Call:
lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
    Marital_Status + Product_Category_1 + Product_Category_2,
    data = train_dataset1)

Residuals:
   Min     1Q Median     3Q    Max
-10422  -3085   -620   2264  17863

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)       11107.557    174.963  63.485 < 0.0000000000000002 ***
GenderM             499.572     59.762   8.359 < 0.0000000000000002 ***
Age18-25            134.443    162.247   0.829              0.40732
Age26-35            336.123    158.040   2.127              0.03344 *
Age36-45            442.710    163.057   2.715              0.00663 **
Age46-50            405.787    179.926   2.255              0.02412 *
Age51-55            878.720    183.355   4.792           0.00000165 ***
Age55+              558.264    202.321   2.759              0.00580 **
Occupation            2.040      3.911   0.522              0.60192
City_CategoryB      281.941     62.531   4.509           0.00000654 ***
City_CategoryC      768.281     67.624  11.361 < 0.0000000000000002 ***
Marital_Status      -63.641     55.098  -1.155              0.24808
Product_Category_1 -401.989      7.506 -53.553 < 0.0000000000000002 ***
Product_Category_2  -80.315      6.586 -12.194 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4638 on 33320 degrees of freedom
Multiple R-squared:  0.1194,    Adjusted R-squared:  0.1191
F-statistic: 347.6 on 13 and 33320 DF,  p-value: < 0.00000000000000022
```

Figure 30: Linear regression model without Stay_In_Current_City_Years variable

Figure 31 depicts the output of linear regression model without Occupation and Stay_In_Current_City_Years variable. We will name this model as model3. We can see that Adjusted R-squared value is still that same as model1 and model2. The residual error is also same that is, 4638. Marital_Status is still not significant and hence we will remove it in the next model. All other variables show good significance at 95% confidence interval.

```
> model3 <- lm(Purchase ~ Gender + Age + City_Category + Marital_Status + Product_Category_1 + Product
_Category_2, data = train_dataset1)
> summary(model3)

Call:
lm(formula = Purchase ~ Gender + Age + City_Category + Marital_Status +
    Product_Category_1 + Product_Category_2, data = train_dataset1)

Residuals:
     Min       1Q   Median       3Q      Max
-10427.8  -3082.0   -619.6   2264.5  17868.2

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)       11123.316    172.333  64.545 < 0.0000000000000002 ***
GenderM             503.214     59.352   8.478 < 0.0000000000000002 ***
Age18-25            129.916    162.013   0.802              0.42263
Age26-35            333.655    157.967   2.112              0.03468 *
Age36-45            442.317    163.054   2.713              0.00668 **
Age46-50            405.114    179.919   2.252              0.02435 *
Age51-55            879.000    183.352   4.794           0.00000164 ***
Age55+              559.347    202.308   2.765              0.00570 **
City_CategoryB      282.188     62.528   4.513           0.00000641 ***
City_CategoryC      769.038     67.608  11.375 < 0.0000000000000002 ***
Marital_Status      -64.128     55.090  -1.164              0.24441
Product_Category_1 -402.011      7.506 -53.557 < 0.0000000000000002 ***
Product_Category_2  -80.315      6.586 -12.194 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4638 on 33321 degrees of freedom
Multiple R-squared:  0.1194,    Adjusted R-squared:  0.1191
F-statistic: 376.6 on 12 and 33321 DF,  p-value: < 0.00000000000000022
```

Figure 31: Linear regression model without Occupation and Stay_In_Current_City_Years variables

Figure 32 shows the output of linear regression model without Marital_Status, Occupation and Stay_In_Current_City_Years variables. We can see that now all variables are significant. The age group 18-25 is not significant but other age groups show good significance and we will not remove it from the model. Adjusted R-squared value is still 0.1191 which means that our model is stable. Further, to check if the model improves, variable age is removed which resulted in decreased Adjusted R-squared value and hence it is not efficient. We will call this model as model4 which will be our final model with the linear regression algorithm.

```
> model4 <- lm(Purchase ~ Gender + Age + City_Category + Product_Category_1 + Product_Category_2, data
 = train_dataset1)
> summary(model4)

Call:
lm(formula = Purchase ~ Gender + Age + City_Category + Product_Category_1 +
    Product_Category_2, data = train_dataset1)

Residuals:
     Min       1Q   Median       3Q      Max
-10444.7  -3080.1   -618.1   2266.2  17867.8

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)      11123.240    172.334  64.545 < 0.0000000000000002 ***
GenderM            503.509     59.352   8.483 < 0.0000000000000002 ***
Age18-25           116.143    161.582   0.719              0.47228
Age26-35           308.724    156.509   1.973              0.04855 *
Age36-45           415.070    161.366   2.572              0.01011 *
Age46-50           358.979    175.501   2.045              0.04082 *
Age51-55           831.827    178.819   4.652           0.00000330 ***
Age55+             520.132    199.485   2.607              0.00913 **
City_CategoryB     281.679     62.527   4.505           0.00000666 ***
City_CategoryC     768.581     67.607  11.368 < 0.0000000000000002 ***
Product_Category_1 -401.988      7.506 -53.554 < 0.0000000000000002 ***
Product_Category_2  -80.300      6.586 -12.192 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4638 on 33322 degrees of freedom
Multiple R-squared:  0.1194,    Adjusted R-squared:  0.1191
F-statistic: 410.7 on 11 and 33322 DF,  p-value: < 0.00000000000000022
```

Figure 32: Linear regression model without Marital_Status, Occupation and Stay_In_Current_City_Years variables

Once we select the best model, we plot the graph to check for residuals and variance. Model4 is used to plot the below graphs as it gave the best result with good p value.

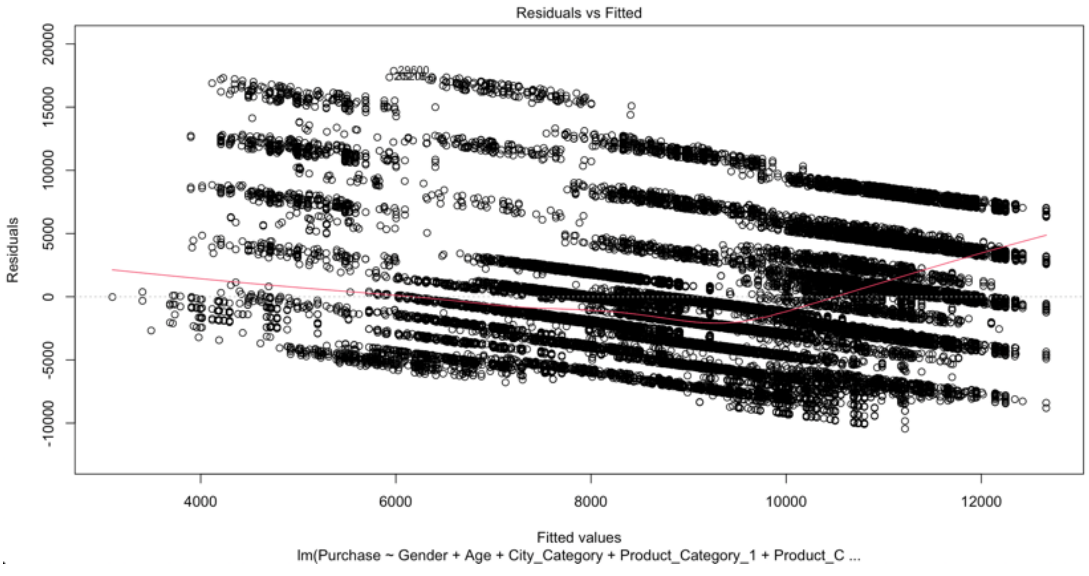

Figure 33: Residuals vs Fitted plot

From figure 33, we can see that the linearity assumptions are fairly met as we do not see any pattern in the data points. The red line is a little curved which indicates that the linearity assumption is not completely met. This is due to the homoscedasticity in the data. We can also see that there is no particular pattern in the distribution of points which tells that the variance is constant.

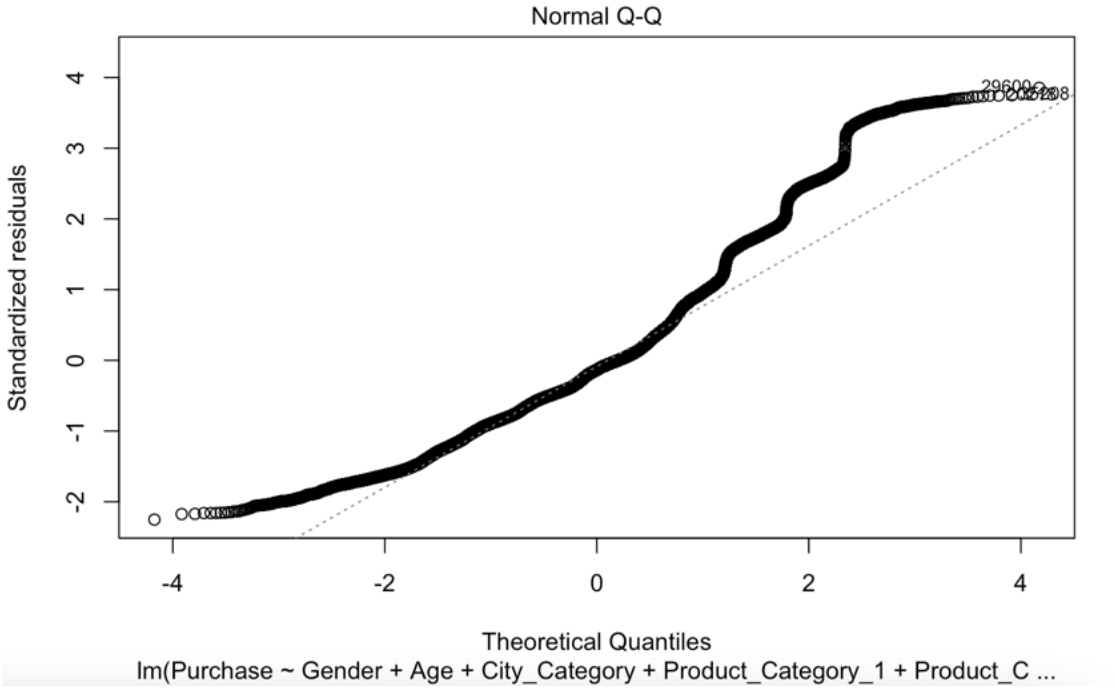

Figure 34: Normal Q-Q plot

In figure 34, Y-axis represents ordered, observed and standardised residuals and X-axis represents theoretical residuals. We can see that residuals or errors are normally distributed as most of the data points fall on the diagonal line.

After finalising the model, we calculate the Root Mean Square Error(RMSE) and Mean Absolute Error (MAE) value to check for accuracy of the models. We will calculate these values for both model3 and model4 for comparison.

Model	**RMSE**		**MAE**	
	Training	Test	Training	Test
Model3	4663.596	4671.083	3577.313	3577.884
Model4	4663.757	4670.785	3577.372	3577.564

Table 1: RMSE and MAE values for multiple linear regression mode3 and model4

From table 1, we can see that the RMSE and MAE values are quite similar for model3 and model4. Model4 provides slightly better results than model3 for testing dataset. Since model4 uses less variables when compared to model3, it makes the model efficient by saving time and cost and providing good results. We will perform the regression with other algorithms and compare the results of the models.

4.3 Support Vector Machine

For many classification and regression issues, the SVM (Support Vector Machine) technique is a well-liked choice. It is one of the methods for supervised learning that uses hyperplanes as decision boundaries. The hyperplane is a line in two-dimensional space. SVM assigns each sample to a certain membership class based on the training samples. As it builds the model, the SVM training algorithm divides up fresh instances into different classes and assigns a hyperplane to the output. It is a non-probabilistic binary linear classifier as a result. An SVM model is a representation of instances as points in space that is mapped so that points in various categories are as far apart from the decision border as possible Madge & Bhatt (2015). Instead of using a direct method, SVM estimates probability estimates using the five-fold cross-validation methodology. It becomes more costly as a result. They work well in higher dimensions, particularly when there are more dimensions than samples. This is because the SVM classifier has hyperparameters, such as gamma, regularisation parameter(C), and a selection of kernels. C dictates how much data misclassification is permitted, gamma indicates the range of influence a training sample has, and kernel, which may be linear, polygonal, or rbf, controls how the hyperplane is learned Hearst, et al., (1998).

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. SVMs are based on the idea of finding a hyperplane in a high-dimensional space that maximally separates the different classes. The points that are closest to the hyperplane are called support vectors, and the distance between the hyperplane and the nearest support vector is called the margin. The goal of the SVM is to find the hyperplane with the greatest margin, as this maximally separates the different classes and results in the best generalization performance. In the case of binary classification, the SVM finds the hyperplane that maximally separates the two classes, with all the points of one class on one side of the hyperplane and all the points of the other class on the other side. SVMs are particularly useful when the data is not linearly separable and a non-linear decision boundary is needed. In these cases, SVMs can use the kernel trick to transform the data into a higher-dimensional space, where it may be possible to find a linear decision boundary.

Consider an n-dimensional vector $x = (X_1,...,X_n)$. Linear boundary (hyperplane) is given by,

$$\beta_0 + \beta_1 X_1 + ... + \beta_n X_n = \beta_0 + \sum_{i=1}^{n} \beta_i X_i = 0$$

The sum of the components in one category will then be more than zero, while the sum of the elements in the other category will be less than zero. With labelled examples, $\beta_0 + \sum_{i=1}^{n} \beta_i X_i = y$, where y is the label. In our classification, $y \in \{-1, 1\}$.

We can rewrite the hyperplane equation using inner products.

$$y = \beta_0 + \sum \alpha_i y_i \, x(i) * x$$

where $*$ represents the inner product operator. The ideal hyperplane maximises the separation between the plane and any given point. The margin is that area. A maximum margin hyperplane (MMH) divides the data most effectively. However, since the differentiation may not be exact, we can add error variables 1 through n and keep the sum of those variables below a certain budget B. Only the points closest to the boundary are important for hyperplane selection; all

other points are irrelevant. Since the hyperplane assigns each of these points to one of two classes, it is known as a Support Vector Classifier (SVC). These points are referred to as the support vectors Moore (2014).

4.4 Applying Support Vector Machine algorithm

Training dataset with 33334 observations and 9 variables is used in this model. User_ID, Product_ID and Product_Category_3 variables are removed as it does not give us insight with respect to purchase. Since the study is about regression, we specify the SVM type as eps-regression. We also use the radial kernel which gives the output in exponential basis. Cost is by default 1 which represents cost of constraints violation. Epsilon is also by default 0.1 which represents insensitive loss function. From figure 35, we can see that gamma value approximately is 0.056 which explains around 56% variance. From this we can tell that SVM model works better than linear regression for this dataset. I will calculate RMSE and MAE values to check the accuracy of the model.

```
Call:
svm(formula = Purchase ~ ., data = train_dataset1, type = "eps-regression", kernel = "radial")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.05555556
    epsilon:  0.1


Number of Support Vectors:  29579
```

Figure 35: Output of Support vector machine algorithm with radial kernel

The kernel type is changed to polynomial and the algorithm is applied to check if the accuracy improves

```
Call:
svm(formula = Purchase ~ ., data = train_dataset1, type = "eps-regression", kernel = "polynomial")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.05555556
     coef.0:  0
    epsilon:  0.1


Number of Support Vectors:  29376
```

Figure 36: Output of Support vector machine algorithm with polynomial kernel

From figure 36, we can see that gamma value is still the same so there is no change in the variance. The RMSE and MAE values are calculated for this model and compared with the previous model to check which model performs better.

Kernel	**RMSE**		**MAE**	
	Training	Test	Training	test
Radial	4176.846	4189.219	3001.031	3032.828
Polynomial	4211.201	4215.808	3033.626	3058.767

Table 2: Comparison between RMSE and MAE values with respect to Radial and Polynomial kernel

From table 2, we can see that RMSE and MAE values of Radial kernel values are lesser than Polynomial kernel for both training and test data. Lower the value of RMSE and MAE, better the model and hence we can say that model with radial kernel provides better accuracy as the error rate is lesser than that of the polynomial model.

4.5 Random Forest

The random forest algorithm, which Leo Breiman proposed in 2001 as a general-purpose classification and regression technique, has proven to be incredibly effective. The method has demonstrated good performance in situations where the number of variables is significantly greater than the number of observations Biau & Scornet (2016). It combines a number of randomised decision trees and averages their predictions. It may also be applied to complex issues, is simple to customise for different ad hoc learning assignments, and provides metrics of varying importance.

A Random Forest, as its name suggests, is an ensemble of trees where each tree depends on a set of random variables Cutler, et al., (2012). Formally, we assume an unknown joint distribution $P_{XY}(X, Y)$ for a p-dimensional random vector X =$(X_1, \ldots, X_p)^T$ representing the real-valued input or predictor variables and a random variable *Y* representing the real-valued response. The aim is to discover a *f(X)* prediction function that can predict *Y*. A loss function *L(Y, f(X))* defines the prediction function to minimise the expected value of the loss.

$$E_{XY}(L(Y, f(X)))$$

where the subscripts represent expectations in relation to the combined distribution of X and Y.

L(Y, f(X)) is also used to measure the proximity between *f(X)* and *Y*. *L* is defined by squared error loss *L(Y, f(X)) = (Y – f(X))^2* for regression problems and zero-one loss for classification:

$$L(Y, f(X)) = I(Y \mathrel{!}= f(X)) = \begin{cases} 0 \; if \; Y = f(X) \\ 1 \; otherwise \end{cases}$$

Minimizing $E_{XY}(L(Y, f(X)))$ for squared error loss gives conditional expectation,

$$f(X) = E(Y|X = x)$$

which is mostly known as regression method.

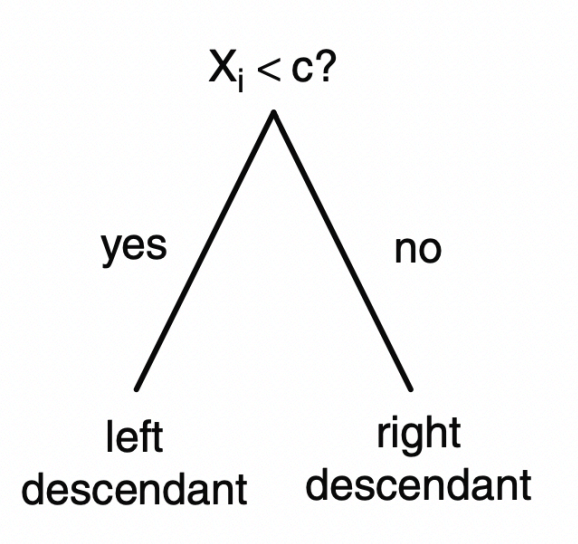

Figure 37: Dividing on a predictor variable X_i with split point c. Cutler, et al., (2012)

Ensembles form *f* in terms of a series of base learners $h_1(x), \ldots, h_j(x)$ and these base learners are joined together to give the ensemble predictor *f(x)*. In regression, the base learners are averaged.

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x)$$

In Random Forest models, the *j*th base learner is a tree represented by hj(X, Θ_j), where Θj is a group of random variables and the Θj 's are unconstrained for *j = 1,… , J* Cutler, et al., (2012).

4.6 Applying Random Forest algorithm

Training dataset with 33333 observations and testing dataset with 16667 observations and 9 variables are used for this model. Random forest algorithm is applied with different number of trees and mtry values. The ntree argument represents the number of trees to be used in the model. The ntree parameter should have a value such that it stabilises the error rate without having too many unnecessary trees. Different ntree values of 100, 500 and 1000 are applied and results are compared. The parameter mtry represents the number of variables to randomly sample as candidates at each split. The optimal value of mtry for a regression problem is (number of features)/3. Since we have chosen the dataset with 9 features, the ideal value of mtry is 9/3 which is 3.

```
Call:
 randomForest(formula = Purchase ~ ., data = train_dataset1, ntree = 100,      mtry = 3)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 3

          Mean of squared residuals: 9411645
                    % Var explained: 61.64
```

Figure 38: Output of random forest algorithm with ntree = 100

From figure 38, we can see that around 61.64% of variance is explained in the dataset which is much better than the previous algorithms used.

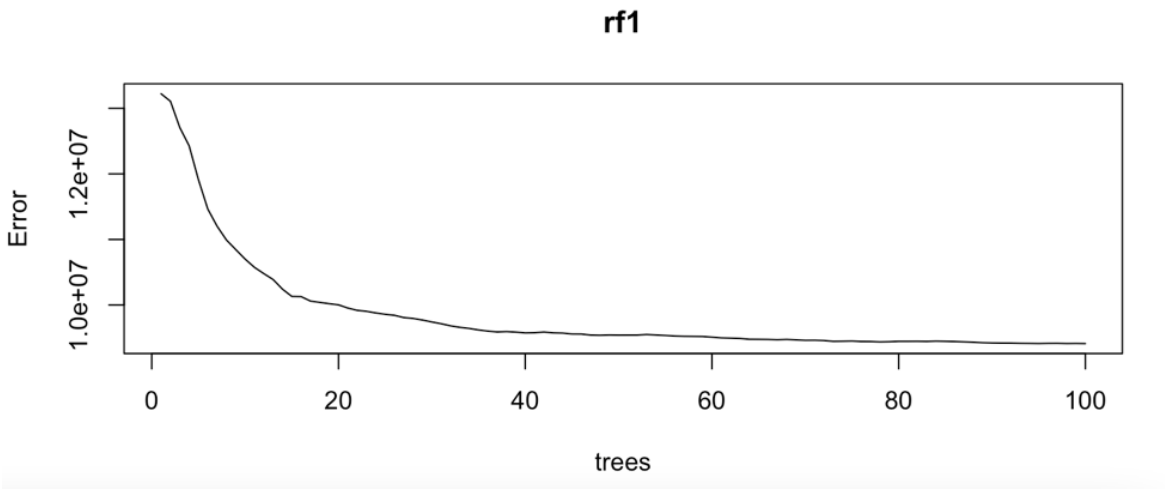

Figure 39: Plot of Error vs Number of Trees with ntree = 100

Figure 39 depicts the error rate with respect to number of trees used in the model. We can see that error rate decreases with the increase in number of trees. With 100 trees, the model shows negligible error. We will now increase the number of trees to 500 and check for results.

```
Call:
 randomForest(formula = Purchase ~ ., data = train_dataset1, ntree = 500,      mtry = 3)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 9309578
                    % Var explained: 62.06
```

Figure 40: Output of random forest algorithm with ntree = 500

From figure 40, we can see that the variance has increased to 62.06% when number of trees is 500. There is an increase of 0.42% from the previous model with ntree = 100 and hence this model performs better. The mean of squared residuals is also decreased from the previous model.

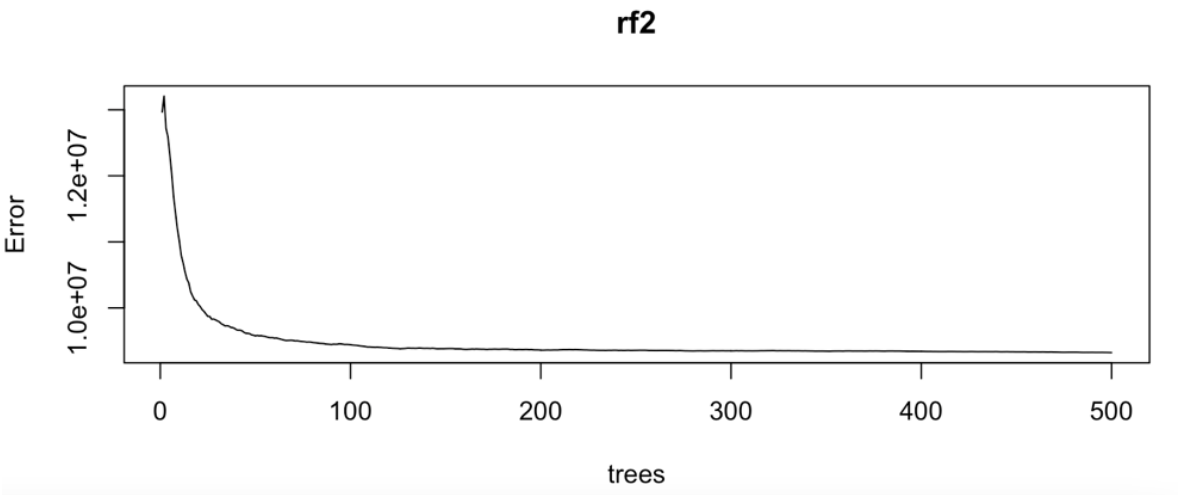

Figure 41: Plot of Error vs Number of Trees with ntree = 500

From figure 41, we can see that the error rate has further reduced as compared to the previous model. From this we can tell that increasing number of trees gives better result for this dataset. We will further run the model with 1000 trees and check if this assumption still holds true. If the assumption is true, the percentage of variance should increase and mean residuals should decrease.

```
Call:
 randomForest(formula = Purchase ~ ., data = train_dataset1, ntree = 1000,      mtry = 3)
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 3

          Mean of squared residuals: 9298220
                    % Var explained: 62.1
```

Figure 42: Output of random forest algorithm with ntree = 1000

From figure 42, we can see that the percentage of variance explained is 62.1% which is slightly higher than previous model with 500 trees . There is only an increase of 0.04% with the double the number of trees which is not very useful. But the residual error has decreased which is a good sign. Increasing the number of trees increases the computation time and is expensive in real time. Since there is no significant difference from the previous model, we compute RMSE and MAE values to check if increasing the number of trees is really beneficial to the model improvement.

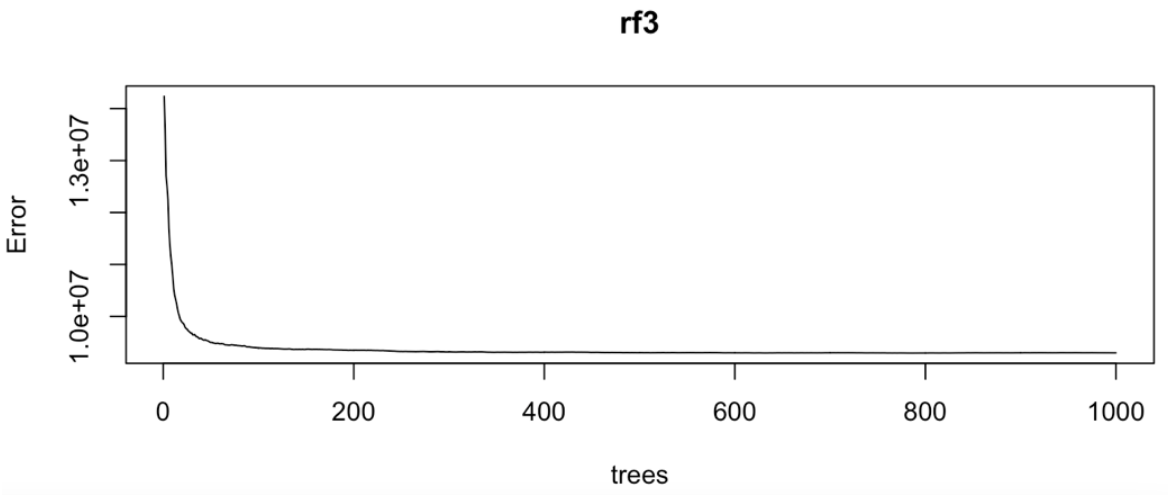

Figure 43: Plot of Error vs Number of Trees with ntree = 1000

Figure 43 depicts the error rate for a model with 1000 trees. We can see that error rate is quite constant after 500 trees. Increasing the number of trees to 1000 does not improve the error rate as such. We will now compute the RMSE and MAE values for all 3 models and compare the results to determine the best model for this dataset.

Number of trees	**RMSE**		**MAE**	
	Training	Test	Training	test
100	2345.05	3037.156	1749.38	2290.677
500	2342.321	3041.92	1749.948	2288.85
1000	2342.469	3038.6	1749.72	2285.694

Table 3: RMSE and MAE values for random forest model with 100, 500 and 1000 trees

From table 3, we can see that the model with 1000 trees provides the best result as it has the lowest Root mean square error and Mean absolute error. However, the values are not significantly higher than that of 100 trees. We can also observe that the difference between training and testing values for both RMSE and MAE are quite high. This indicates that the prediction is not very accurate. When we are using a large set of data, increasing the number of trees makes the model unnecessarily complex and it is hard to understand. If the error rates are not highly impacted by increasing the number of trees, it is best to not implement it as degrades the model from being cost and time effective. Because of the above reasons, I feel the model with 100 trees is more efficient. Further we will compare the results from all three algorithms to select the best model for this dataset.

After all the algorithms are applied and models are developed with different parameters and configurations, results from these models are compared to choose the best model which provides good accuracy with minimal residual.

5 Results and Evaluation

In this section we will compare the results from multiple linear regression, support vector machine and random forest algorithm. We will compare the RMSE and MAE values to decide the best model for this dataset. We use the metrics library in R to compute these values. This section also comprises of plots between different variables to describe the relationships and pattens in the data frame.

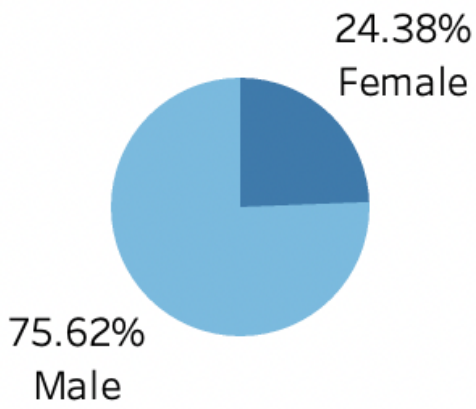

Figure 13: Pie chart representing Gender ratio

Figure 13 depicts the percentage ratio of Male and Female users in the dataset. As we can see, the store predominantly has male users constituting of almost 76% of total customers. The percentage of female users in only 24%. We can say that the store has more male shoppers.

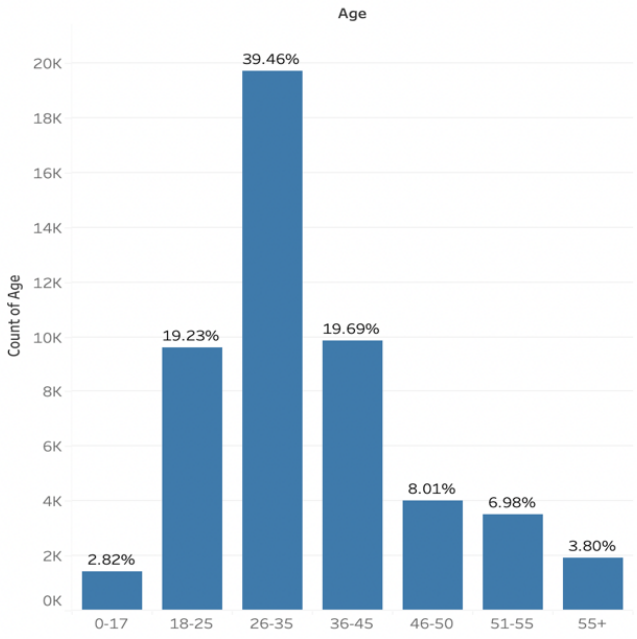

Figure 14: Percentage distribution of Age ranges

From figure 14 we can see that most of the users fall under the age range of 26-35 years. Least amount of users are of age 0-17 years. The age ranges 18-25 and 36-45 also has decent percentage of shoppers. From this, we can say that majority of the users fall under the age of 18 to 45.

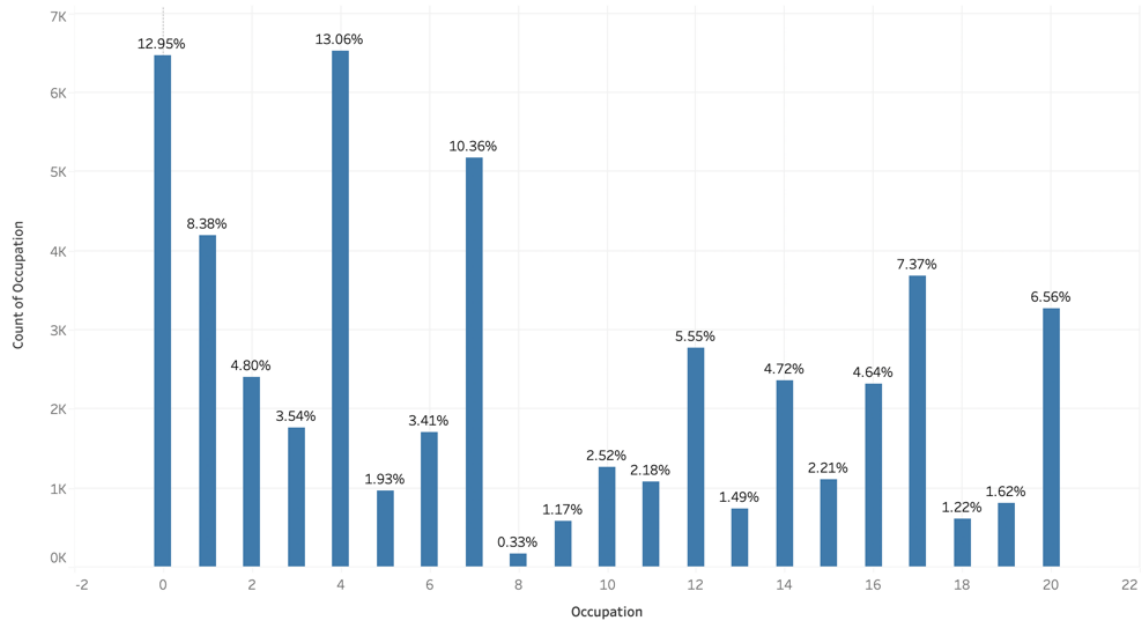

Figure 15: Percentage distribution of Occupation of users

From figure 15, we can see that occupation category 4 and 0 has the most users with 13.06 and 12.95% distribution. Occupation category 7 also has around 10% of users. Occupation category 8 has the least number of users with 0.33%. We can see that users are distributed across all the occupation categories and there is no strong insight on the purchase behaviour.

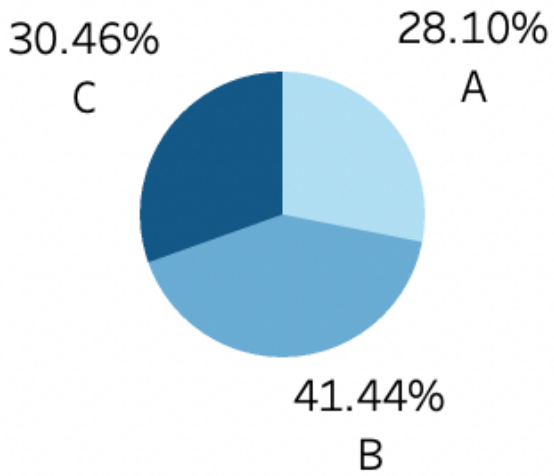

Figure 16: Percentage distribution of users in different cities

From figure 16, we can see that City B has highest users with 41.44% distribution and City A has least users with 28.10% distribution. From this, we can say that store in City B needs proper inventory planning as the number of customers is high here.

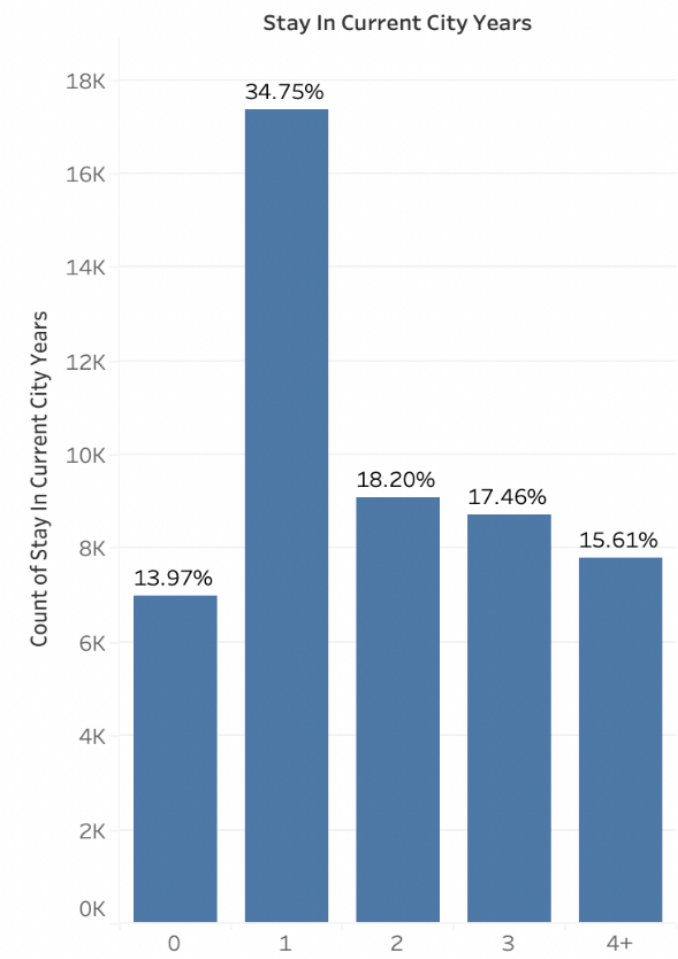

Figure 17: Percentage distribution of number of years users staying in current city

Figure 17 depicts the period of percentage of users staying in the current city. Most of the users have lived for 1 year making 35% of the total users. Lowest users are staying in the city for less than a year. Around 18% of users have lived in the city for 2 years and 17% of users have lived for 3 years. From this graph we can say that users who have just moved in to the city are less likely to purchase from the store. We can also say that the users tend to move to a different city after 2 years of staying in the current city.

Since Purchase is the dependent variable in the dataset, I will plot it against all the other independent variables to gain useful insights.

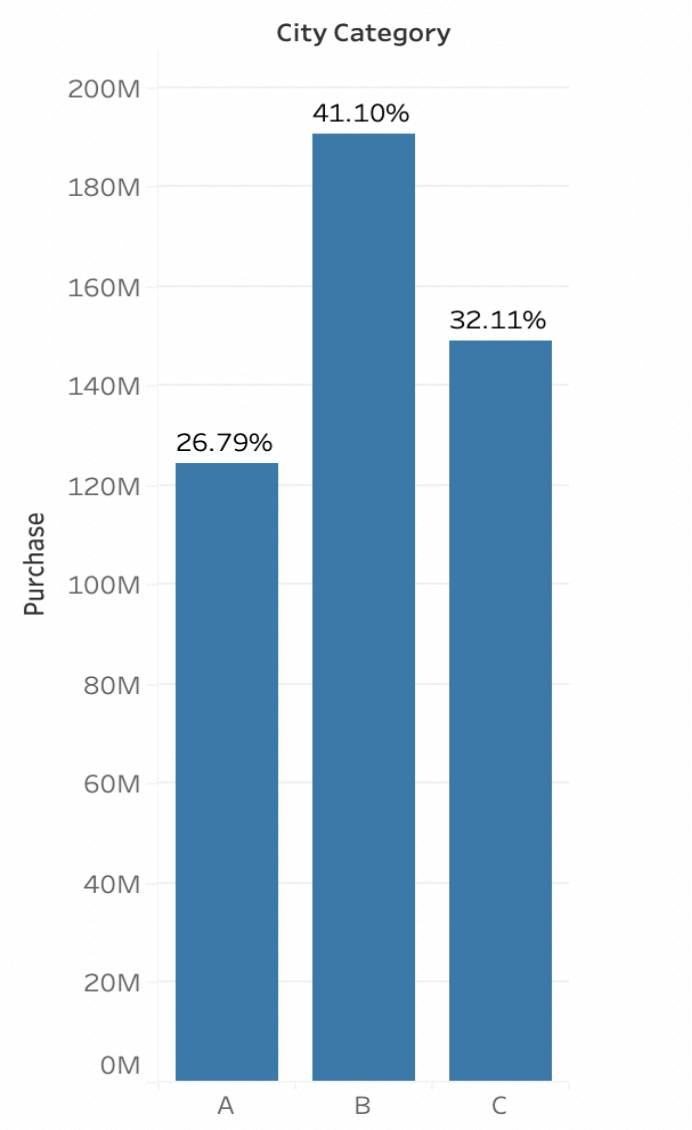

Figure 18: City category vs Purchase

From figure 18 we can see that highest amount of purchase is done by City B with 41% followed by City C with 32% of total purchases. City A has least purchase constituting 27% of total purchases. From this, we can say that the store in City B is going to have the highest purchase and the store owner needs to make necessary arrangements for the same.

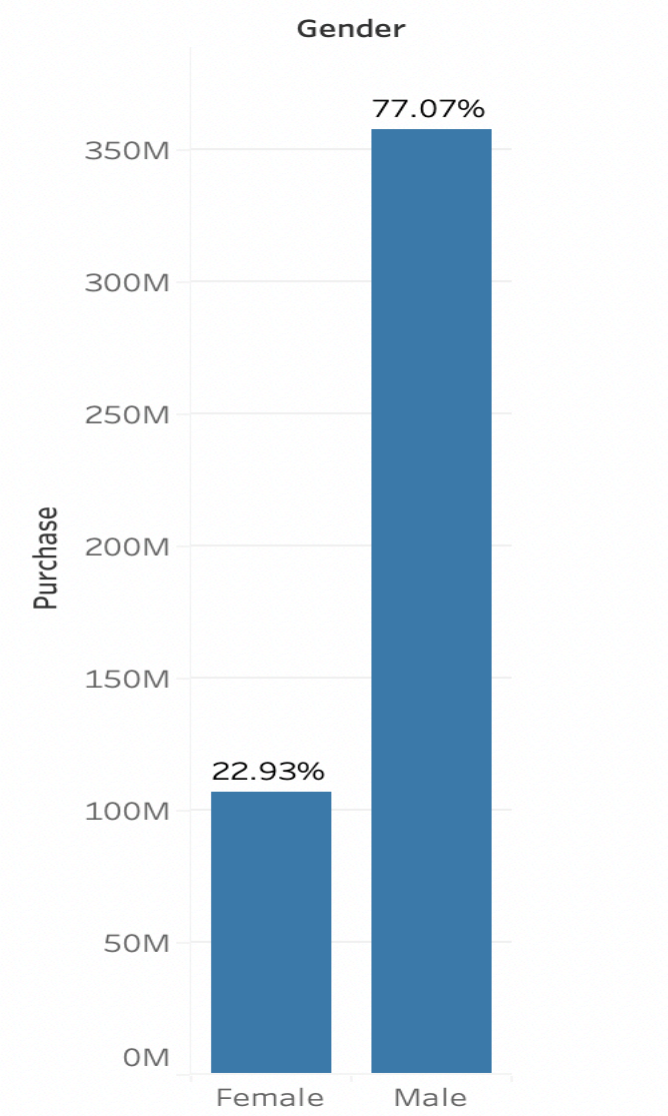

Figure 19: Gender vs Purchase

From figure 19, we can see that male users are making more purchases than female users. The purchases made by male users is 3 times more than that of female users. Since most of the customers are males, stores can plan their marketing strategy to target them. The stores can also come up with strategies to attract more female shoppers.

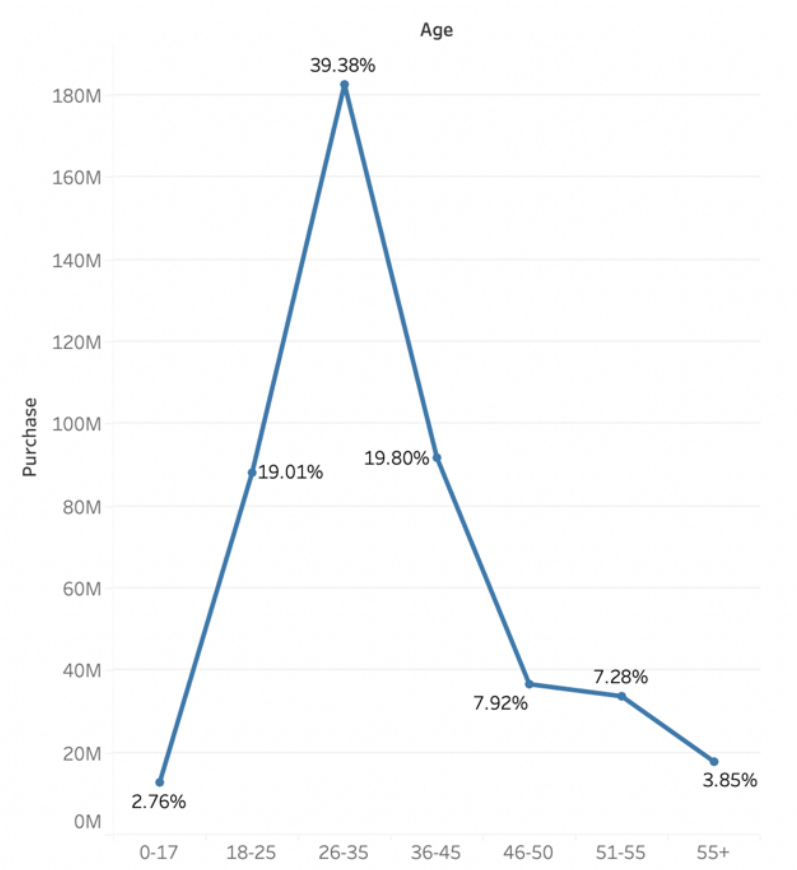

Figure 20: Age vs Purchase

Figure 20 depicts the percentage of purchases made by each age group. Age group 26-35 years has made the highest purchase and lowest purchase of 2.76% is done by age group 0-17 years. We can infer that youths who are settled with a job are making most of the purchases. The store can come up with strategies to target this segment of users. Most of the purchases are done by people who fall under the age range of 18 to 45. Any user who is younger than 18 years or older than 45 years tend to purchase less.

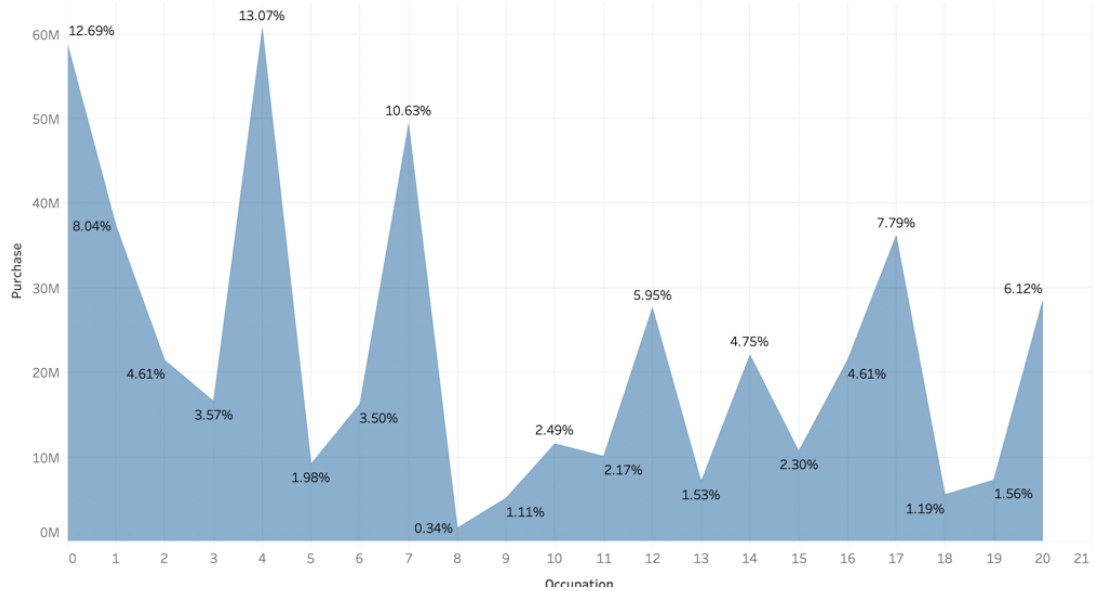

Figure 21: Occupation vs Purchase

From figure 21, we can see that maximum purchase of 13.07% is done by occupation category 4 followed by occupation category 0 with 12.69% purchase. Lowest purchase is done by occupation category 8 with a value of 0.34% followed by occupation category 9 and 18 with values 1.11% and 1.19% respectively.

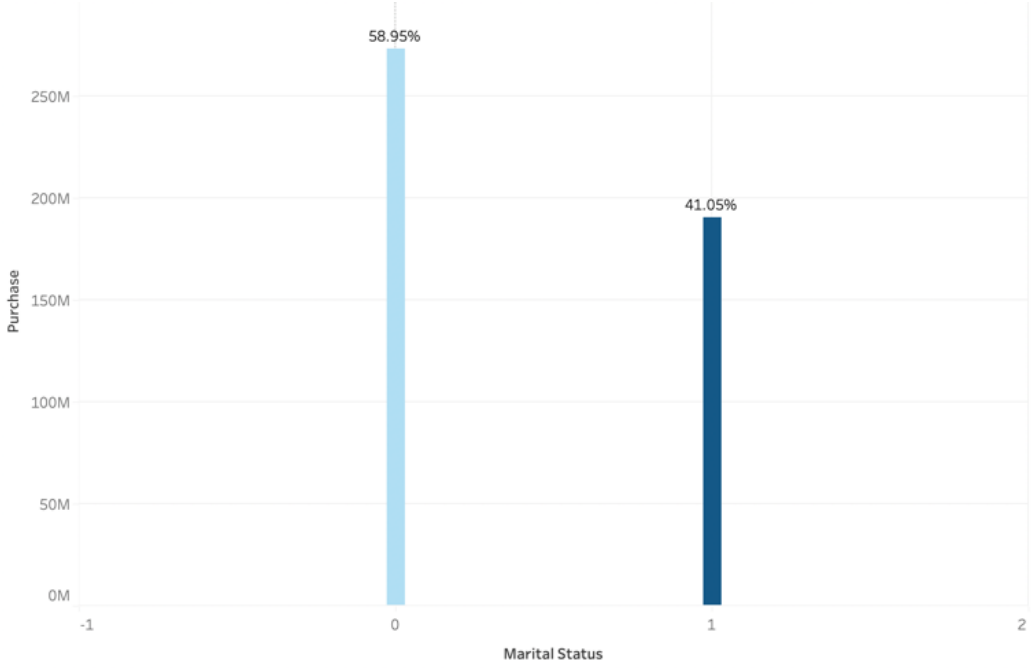

Figure 22: Marital Status vs Purchase

In figure 22, 0 represents single people and 1 represents married people. Majority of the purchases, that is, 59% is done by users who are single. The store can come up with marketing strategies to target this set of users. Married users also make a purchase of 41% which is a good percentage so we can tell that marital status does not highly influence the purchase pattern.

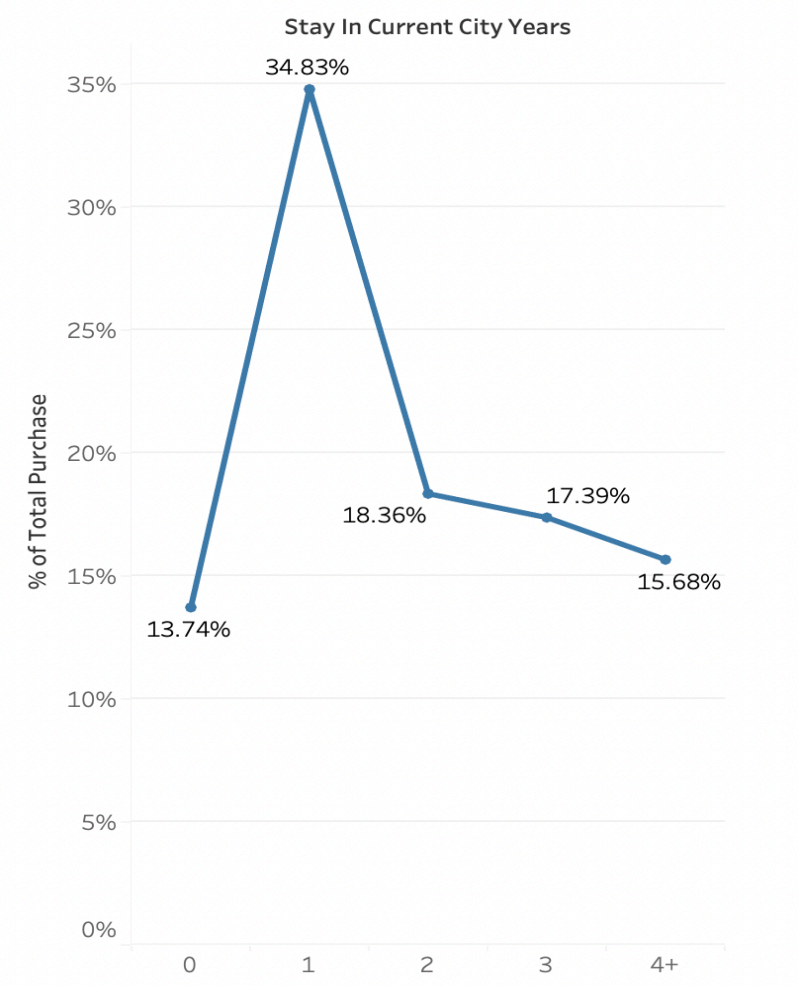

Figure 23: Stay_In_Current_City_Years vs Purchase

From figure 23, we can see that 35% of purchases are from users who have lived in the city for 1 year. Lowest purchase of 14% is done by users who have lived in the city for less than a year. We can also infer that once a user gets adjusted to the city, they tend to purchase less when compared to purchase behaviour of 1 year users.

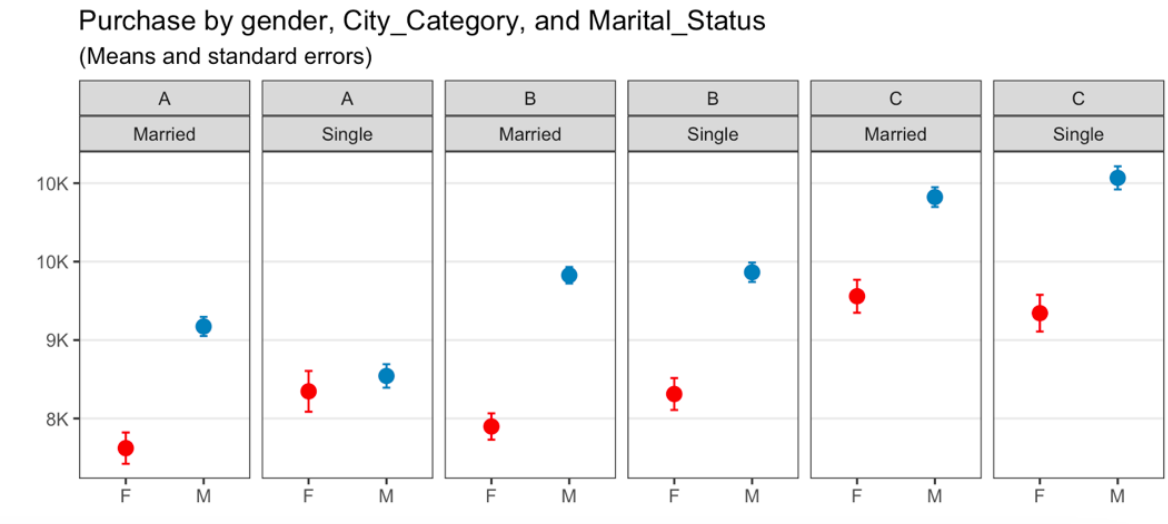

Figure 24: Purchase pattern by Gender, City Category and Marital Status

Figure 24 depicts the purchase behaviour of users with respect to gender, city category and marital status. In City A, we can see that married men are purchasing more and the least purchase is done by married women. Single people in City A mostly spend around 8000 to 9000 on shopping. In City B, both married and single men tend to purchase more than that of women. Marital status does not play any significant role in City B. In City C, single men tend to purchase more. They spend at least 10000 on shopping. In all the three cities, married women does least amount of purchase. Also the significant observation is that men purchase more than women.

All the visualisations are done using Tableau and R programming language. As mentioned in section 2.4, Tableau provides user friendly approach to plot the graphs. Many interesting relationships were found using visualisations which proved the importance of graphical representation as mentioned in sections 2.3 and 2.4.

We will now discuss the results of different models from sections 4.2, 4.4 and 4.6.

Algorithm	**RMSE**		**MAE**	
	Training	Test	Training	Test
Multiple linear regression	4663.757	4670.785	3577.372	3577.564
Support vector machine	4176.846	4189.219	3001.031	3032.828
Random forest	2345.05	3037.156	1749.38	2290.677

Table 4: RMSE and MAE values for multiple linear regression, SVM and Random forest models

Table 4 depicts the RMSE and MAE values for all the 3 algorithms used in this research. I have chosen model4 from multiple linear regression algorithm which consists of 5 dependent variables namely Age, Gender, City_Category, Product_Category_1 and Product_Category_2 along with the dependent variable Purchase. For support vector machine algorithm, I have chosen radial kernel over the polynomial kernel as it provides better accuracy which has been discussed previously. For random forest algorithm, I have chosen the model with 100 trees as it provides better result as discussed in the section 4.6.

From table 4, we can see that error rate for multiple linear regression is very high. It has the highest RMSE value of 4670.785 and MAE value of 3577.564 for the testing dataset. Random forest algorithm has the lowest RMSE value of 2345.05 and MAE value of 1749.38 for the training dataset. However, we cannot say that Random forest is the best algorithm because the difference in the RMSE and MAE values for training and testing dataset is quite high. The difference between RMSE value of random forest algorithm for training and testing data is 692.106 whereas, the difference between MAE value for training and testing data is 541.297. This huge difference in the values makes the prediction weak and unreliable Bajaj & Purvika (2020).

The error rate of SVM algorithm lies between that of multiple linear regression and random forest models. The difference between RMSE value of training and testing data is 12.373 and the difference between MAE value of training and testing data is 31.797. We can see that the difference between training and testing values of both RMSE and MAE are very minimal. This makes the model more reliable as the predictions are accurate.

In comparison of all the 3 models with algorithms multiple linear regression, support vector machine and random forest, we can say that SVM model with radial kernel provides most accurate prediction and random forest model gives lowest mean residual. Even though random forest model gives minimal error rate than other algorithms, we cannot consider it to be the best model as the prediction accuracy is lower than the other models. In real time, choosing a model with low prediction accuracy results in overestimating or underestimating sales which will result in a huge loss for the company. Hence we cannot consider random forest model. Because of all these reasons, the model obtained from support vector machine algorithm works best for this dataset. The SVM model provides a good prediction accuracy and has comparatively lower error rates.

In section 2.1, we discussed the study by Odegua (2020) where analysis was performed on 5k records. In this study, the number of records were increased to 50k and this proved the results to be better. For random forest model, the MAE was decreased to 2290.677 from 4091.78 thus making our model efficient. Research by Kaneko and Yada (2016) in section 2.1 describes the importance of categories. In this project, variables such as City, Products and Stay_In_Current_City_Years had categories which helped in providing targeted results thus proving the importance of categories. In section 2.1, Zhang, et al., (2016), Setiawan, et al. (2017) and Ezhilarasan & Ramani (2017) have performed analysis on online data. In this project, same process was followed but with an offline retail store data. We can see that the end result for both online and offline data are similar. Wu (2018) used 550k records to perform the analysis. XGBoost algorithm had the least residual and linear regression had high residual. In this project, with a reduced data size of 50k records, linear regression still had high residual but random forest model performed better. Random forest model was also run with 100, 500 and 100 trees to compare the results.

In section 2.2, Rajeshwari & Sushma (2021) conducted the research using Black Friday Sales dataset with 550k observations and 12 features. This study performed the same analysis but implemented thorough data cleaning and feature selection to improve the results. Multiple linear regression model in this project had a RMSE of 4663.757 whereas, in section 2.2 it was 4617.99. So our model did not perform better for this algorithm. Random forest model in this project had a RMSE value of 2345.05 and in section 2.2, it had a value of 3062.72. The error rate has decreased drastically in our model and hence increasing the accuracy.

In section 2.5, the study by Bajaj & Purvika (2020) proved Random Forest algorithm to have better result than other algorithms. In this project, even though random forest model had the least error rate, SVM model performed better as the difference between training and testing RMSE and MAE values were small.

In section 2.6, the study by Guo, et al., (2013) proposed to make feature selection to improve the accuracy of the model. In this project, features selection was performed using correlation and variables with insignificant p values were removed. This helped the model to perform better by improving the RMSE and MAE values.

In section 2.9, as proposed by Ramachandra, et al., (2021), both RMSE and MAE have been used to check the model accuracy in this project. The values were calculated for both training and testing data and the results were verified.

5.1 Limitations

The limitations associated with this study are described in this section.

The number of observations in the dataset has been reduced to 50K from 550K records since the RStudio application was getting automatically terminated during the modelling phase. As discussed in section 2.1 of literature review, reducing the size could affect the accuracy of the model.

Variables such as Occupation, Product_Category_1, Product_Category_2 and Product_Category_3 are masked which restricts the planning of marketing strategy. Unmasking these variables might result in making better business decisions.

These are some of the limitations of this project.

5.2 Future Enhancements

There is always a scope for improvement and this section describes some the observations noted that might help in enhancing the model.

The modelling can be done using all 550K records. This might help in improving the prediction accuracy.

For this study, the first 50K records have been selected for modelling. The observations can be selected at random to check if the model performs better.

Other regression algorithms such as Aripori, Ridge regression, Arima, Prophet, etc can be applied. This will give us more insight on which model to use to predict sales.

In this study, missing values are imputed using mean value of the variable. Other method such as replacing NA's with 0 can be implemented to check is the accuracy improves.

These are some of the improvements which could help in increasing the model accuracy.

6 Conclusion

Sales prediction modelling typically involves identifying the main factors that impact sales performance, determining their relative importance, and creating a model that can accurately predict future sales based on this information. This enables decision makers to understand the key drivers of sales, assess the performance of their existing stores, and identify opportunities for new, profitable stores through accurate forecasting. By understanding the main drivers of sales performance and using accurate forecasting, businesses can make informed decisions about their sales and operations (Tewari, 2023). The results obtained from our model helps the businesses to manage the inventory, staffs and other activities by predicting the upcoming sales.

The aim of this research was to predict sales of a retail store during Black Friday. The dataset by Sammari (2020) on Kaggle was used for this purpose. 50,000 observations with 12 features were used for developing the model. In this study, machine learning (ML) algorithms are used to forecast how much a buyer will likely spend on the upcoming "Black Friday" sale. It has been demonstrated that intriguing trends can be found in a dataset by using interpretive data exploration. According to the study, machine learning approaches build good forecasting model that can be utilised in businesses, and store managers may assess their customers to more effectively attract shoppers and boost Black Friday sales. According to the study, pre-processing the data is necessary to create a useful dataset for creating the prediction model.

The dataset was divided into training and testing data in the ratio 70:30. Machine learning algorithms such as multiple linear regression, support vector machine (SVM) and Random forest were applied. Multiple linear regression applied to 4 models with different variables. These variables were selected based on the p value. Model4 with features Age, Gender, City_Category, Product_Category_1 and Product_Category_2 gave the best result. 2 models with radial kernel and polynomial kernel were developed using SVM algorithm. Model with Kernel radial provided better result. 3 models with 100, 500 and 1000 trees were developed using random forest algorithm. All the 3 models had somewhat similar result and hence the model with 100 trees was considered to be the best as it was cost and time effective. Increasing the number of trees increases the computation time and will turn out to be expensive in real time and hence it is not recommended. Random forest model with 100 trees provided the least residual rate but the difference between training and testing data for RMSE and MAE was too huge which will result in inaccurate prediction. It is a risky move for businesses to select that model. Multiple linear regression provided the highest residual rate for both training and testing data and hence it is not considered to be the best model. SVM model with radial kernel provided lesser error rate when compared with multiple linear regression model. SVM model also had minimal difference between training and testing data for both RMSE and MAE and hence this model is chosen to be the best model to predict sales during Black Friday. This model provides better result than the earlier models described in section 2.1.

By using the SVM model with radial kernel, the shop owner can foresee the sales on the day of Black Friday. This helps them to manage the inventory by providing details of the products that are most likely to have maximum sales and products that are least likely to be purchased. It gives details about which city is going to have maximum sales and which age group makes the most purchases. Appropriate marketing strategies can be implemented to increase the sales and make a profit during Black Friday.

7 Bibliography

SM Wood, S. B., 2006. Convenience store sales forecasting - art before science?. *European Retail Digest,* Volume 15, pp. 15-18.

R. Praveen, D. P. K. A. P. S. a. G. S. S., 2022. Sales prediction using machine learning. *AIP Conference Proceedings ,* 2444(1).

GoCardless,2021.[Online] Available at: https://gocardless.com/guides/posts/importance-sales-forecasting/ [Accessed 2022 August].

Sriram,Kothandaraman,2021.*Aviso.*[Online] Available at: https://www.aviso.com/blog/sales-forecasting [Accessed 20 December 2022].

R. Caruana, A. N.-M., 2006. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on machine learning ICML '06,* pp. 161-168.

Diane, N. S. A. a., 2015. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal Production Economics,* Volume 170, pp. 321-335.

Doganis, P., 2006. Time series sales forecasting for short shelf life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering,* Volume 75, pp. 196-200.

Odegua, R., 2020. Applied Machine Learning for Supermarket Sales Prediction.

Wu, C.-S. M. a. P. P. a. G. S., 2018. *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS).* s.l.:IEEE.

Ostwal,K.,2019.*Kaggle.*[Online] Available at: https://www.kaggle.com/kkartik93/black-friday-sales-prediction?select=train.csv
[Accessed 21 December 2022].

Wajgi, K. S. a. R., 2016. Data analysis and visualization of sales data. *World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave),* pp. 1-6.

W.K. Wong, Z. G., 2010. A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics,* Volume 128, pp. 614-624.

Maharjan, M., 2019. Analysis of Consumer Data on Black Friday Sales Using Apriori Algorithm. *SCITECH Nepal,* 14(1), pp. 17-21.

Hotz,N.,2022.[Online]Available at:https://www.datascience-pm.com/crisp-dm-2/
[Accessed 23 December 2022].

Brian,A.,2021.*What is CRISP–DM Methodology?.*[Online] Available at: https://insideaiml.com/blog/What-is-CRISP-–-DM-Methodology%3F-250 [Accessed 23 December 2022].

Nadali, A. &. K. E. &. N. H., 2011. Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology,* Volume 6, pp. 161-165.

Schröer, C. &. K. F. &. M. G. J., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science,* Volume 181, pp. 526-534.

Sammari,S.,2020.*Black Friday Sales Prediction.*[Online]Available at: https://www.kaggle.com/code/midouazerty/black-friday-sales-prediction/notebook [Accessed 20 June 2022].

Miller, R., 2019. *Data Preprocessing: what is it and why is important.* [Online] Available at: https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/ [Accessed 26 December 2022].

Tranmer, M. a. M. E., 2008. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR),* 5(5), pp. 1-5.

Eberly, L., 2007. *Multiple Linear Regression.* s.l.:Humana Press.

Halls-Moore, M., 2014. *Support Vector Machines: A Guide for Beginners.* [Online] Available at: https://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners/ [Accessed 27 December 2022].

Kvalheim, O. M. e. a., 2018. Determination of optimum number of components in partial least squares regression from distributions of the root-mean-squared error obtained by Monte Carlo resampling. *Journal of Chemometrics,* 32(4), p. 2993.

Tableau.com, 2019. *What Is Data Visualization? Definition, Examples, And Learning Resources.*[Online] Available at: https://www.tableau.com/learn/articles/data-visualization [Accessed 31 December 2022].

Hennessey, J., 2014. A picture can be worth more than a thousand words: Integrating data visualization into an undergraduate econometrics course. *Conference on Teaching and Research in Economic Education,* pp. 28-30.

Shaptunova, Y., 2017. *Tableau software review: Pros and cons of a BI solution for data visualization.*[Online] Available at: https://www.sam-solutions.com/blog/tableau-software-review-pros-and-cons-of-a-bi-solution-for-data-visualization/ [Accessed 31 December 2022].

Louis,F.R.B.o.S.,2019.*WhatisFRED?.*[Online]Availableat:https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/ [Accessed 31 December 2022].

Rouse,M.,2012.*Infographics.*[Online]Availableat:https://www.techtarget.com/whatis/definition/infographics [Accessed 31 December 2022].

Fradkov, A. L., 2020. Early history of machine learning[J]. *IFAC-PapersOnLine,* 53(2), pp. 1385-1390.

Yin J, F. V., 2020. A systematic review on business analytics[J]. *Journal of Industrial Engineering and Management,* 13(2), pp. 283-295.

Dalrymple, D. J., 1987. Sales forecasting practices: Results from a United States survey. *International journal of Forecasting,* 3(3-4), pp. 379-391.

Abdulkareem N M, X. A. M., 2021. Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business,* 5(2), pp. 128-142.

Sharkawy, A. N., 2020. Principle of neural network and its main types. *Journal of Advances in Applied & Computational Mathematics,* Volume 7, pp. 8-19.

Miller A S, B. B. H., 1992. Review of neural network applications in medical imaging and signal processing. *Medical and Biological Engineering and Computing,* 30(5), pp. 449-464.

O'Shea K, N. R., 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458.*

Mao, Q., 2022. Sales Prediction Based on Machine Learning Scenarios. *BCP Business & Management,* Volume 23, pp. 922-930.

Xu-Ying Liu, J. W. a. Z.-H. Z., 2009. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems Man and Cybernetics,* 39(2), pp. 539-550.

Enko, S. D. a. B., 2005. Is Combining Classifiers with Stacking Better Than Selecting the Best One. *Machine learning,* 54(3), pp. 255-273.

Tewari, A., 2023. Sales Prediction for Franchise Stores Using Artificial Neural Networks..

Minsky, M. & Papert, S., 1969. Perceptron: An Introduction to Computational Geometry. *The MIT Press: Cambridge, MA, USA,* p. 2.

Kong, J. & Martin, G., 1995. A backpropagation neural network for sales forecasting. *In Proceedings of the ICNN'95—International Conference on Neural Networks,* Volume 2, pp. 1007-1011.

Thiesing, F. & Vornberger, O., 1997. Sales forecasting using neural networks. *n Proceedings of the InternationalConference on Neural Networks (ICNN'97),* pp. 2125-2128.

Chen, C.-Y.et al., 2010. The study of a forecasting sales model for freshfood. *Expert Syst,* Volume 37, pp. 7696-7702.

Vhatkar, S. & Dias, J., 2016. Oral-care goods sales forecasting using artificial neural network model. *Procedia Comput.Sci,* Volume 79, pp. 238-243.

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecast,* Volume 30, pp. 1030-1081.

Cincotti, S., Gallo, G., Ponta, L. & Raberto, M., 2014. Modeling and forecasting of electricity spot-prices: Computationalintelligence vs classical econometrics. *AI Commun,* Volume 27, pp. 301-314.

Wang, P.-H., Lin, G.-H. & Wang, Y.-C., 2019. Application of Neural Networks to Explore Manufacturing Sales Prediction. *Applied Sciences,* Volume 9, p. 5107.

Jumaa, A. K. & Omar, H. K., 2019. Big data analysis using apache spark mllib and hadoop HDFS with scala and java. *Kurdistan Journal of Applied Research,* 4(1), pp. 7-14.

Amruta Rajeswari, A. & Sushma, K. V., 2021. *Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques.* s.l.:s.n.

Baba, N. & Suto, H., 2000. Utilization of artificial neural networks and GAs for constructing an intelligent sales prediction system. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN,* Volume 6.

Bajaj & Purvika, 2020. Sales Prediction Using Machine Learning Algorithms. *nternational Research Journal of Engineering and Technology (IRJET),* 7(6).

Biau, G. & Scornet, E., 2016. A random forest guided tour. *TEST,* Volume 25, pp. 197-227.

Bohanec, M., Borštnar., M. K. & Robnik-Šikonja., M., 2017. Explaining Machine Learning Models in Sales Predictions. *Expert systems with applications,* Volume 71, pp. 416-428.

Brynjolfsson, E., Hitt, L. M. & Kim, H. H., 2011. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?.

Chai, T. & Draxler, R. R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development,* 7(3), pp. 1247-1250.

Chipman, H. A., I., G. E. & E., M. R., 1998. Bayesian CART model search. *Journal of the American Statistical Association,* 93(443), pp. 935-948.

Cutler, A., Cutler, D. & Stevens, J., 2012. Random Forests. *Zhang, C., Ma, Y. (eds) Ensemble Machine Learning,* pp. 157-175.

H. V. Ramachandra, B. G., Rajashekar, A. & Patil, H., 2021. Machine Learning Application for Black Friday Sales Prediction Framework. *2021 International Conference on Emerging Smart Computing and Informatics (ESCI),* pp. 56-61.

Punam, K., Pamula, R. & Jain, P. K., 2018. A Two-Level Statistical Model for Big Mart Sales Prediction. *International Conference on Computing, Power and Communication Technologies (GUCON),* pp. 617-620.

Hearst., M. A. et al., 1998. Sup-port vector machines. *IEEE Intelligent Systems and their Applications,* 13(4), pp. 18-28.

Assefi, M., Behravesh, E., Liu, G. & P., T. A., 2017. Big data machine learning using apache spark mllib. *Big Data IEEE International Conference,* pp. 3492-3498.

Javed Awan, M. et al., 2021. A big data approach to black friday sales. *Intelligent Automation & Soft Computing,* 27(3), pp. 785-797.

Gönül, M. S., Önkal, D. & Lawrence, M., 2006. The effect of structural characteristics of explanations on use of a DSS. *Decision Support Systems,* pp. 1481-1493.

Madge, S. & Bhatt., S., 2015. Predicting stock price direction using support vector machines. *Independent work report spring,* Volume 45.

Ramasubbareddy., S., T. A. S, S., K, G. & E, S., 2021. *Advances in Intelligent Systems and Computing.* s.l.:s.n.

Batt., S., Grealis., T., Harmon., O. & Tomolonis., P., 2020. Learning Tableau: A data visualization tool. *The Journal of Economic Education,* Volume 51, pp. 317-328.

Meng., X. et al., 2016. Mllib: Machine learning in apache spark. *ournal of Machine Learning Research,* 17(1), pp. 1235-1241.

Xia, Z., Xue, S. & Wu, L., 2020. ForeXGBoost: passenger car sales prediction based on XGBoost. *Distributed and Parallel Databases,* 38(3), pp. 713-738.

Guo, Z. X., Wong, W. K. & Li., M., 2013. A multivariate intelligent decision-making model for retail sales forecasting. *Decision Support Systems,* 55(1), pp. 247-255.

Schröer, C., Kruse, F., Gómez., M. & Jorge., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science,* Volume 181, pp. 526-534.

Wu., C.-S. M., Patil., P. & Gunaseelan., S., 2018. Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data. *EEE 9th International Conference on Software Engineering and Service Science (ICSESS),* pp. 16-20.

Sundararaman, K. .., Parthasarathi, J., Rao., G. S. V. & S, N. K., 2012. Baseline prediction of point of sales data for trade promotion optimization. *2012 International Conference on Communications and Information Technology (ICCIT),* pp. 17-20.

Kaneko, Y. & Yada, K., 2016. A Deep Learning Approach for the Prediction of Retail Store Sales. *016 IEEE 16th International Conference on Data Mining Workshops (ICDMW).*

Zhang., X., Pei., J. & Ye, X., 2016. Demographic transformation and clustering of transactional data for sales prediction of convenience stores. *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA).*

Setiawan., A., Budhi., G. S., Setiabudi., D. H. & Djunaidy., R., 2017. Data Mining Applications for Sales Information System Using Market Basket Analysis on Stationery Company. *2017 International Conference on Soft Computing Intelligent System and Information Technology.*

Ezhilarasan, C. & Ramani, S., 2017. Performance prediction using modified clustering techniques with fuzzy association rule mining approach for retail. *2017 International Conference on Intelligent Computing and Control (I2C2).*

Kripa,Mahalingam.2021.[Online] Available at: https://www.chargebee.com/blog/importance-of-sales-forecasting/[Accessed August 2022].

Joran,2022.[Online] Available at: https://saleslovesmarketing.co/blog/importance-of-sales-forecasting/[Accessed August 2022].

Dave,Roos.2008.[Online]Available at: https://money.howstuffworks.com/sales-forecasting1.htm[Accessed August 2022].

Samantha,Rohn.2022.[Online] Available at: https://whatfix.com/blog/sales-forecasting/ [Accessed August 2022].

Sammari,S.,2020.[Online] Available at: https://www.kaggle.com/code/midouazerty/black-friday-sales-prediction/notebook

Sriram,Kothandaraman.2021.[Online] Available at: https://www.aviso.com/blog/sales-forecasting[Accessed September 2022].

Yun, Z., Weihua, L. .. & Yang, C., 2014. Applying balanced scordcard strategic performance management to CRISP-DM. *2014 International Conference on Information Science, Electronics and Electrical Engineering,* pp. 2009-2014.

Appendix

I. Ethical Approval Form

SAGE-HDR (v3.6 19/10/22)

Response ID	Completion date
956732-956714-100850920	24 Oct 2022, 12:14 (BST)

1	**Applicant Name**	Anusha Chatra Anilkumar
1.a	**University of Surrey email address**	ac02424@surrey.ac.uk
1.b	**Level of research**	Postgraduate Taught (Masters)
1.b.i	**Please enter your University of Surrey supervisor's name. If you have more than one supervisor, enter the details of the individual who will check this submission.**	Philip Murray
1.b.ii	**Please enter your supervisor's University of Surrey email address. If you have more than one supervisor, enter the details of the supervisor who will check this submission.**	p.murray@surrey.ac.uk
1.c	**School or Department**	Surrey Business School
1.d	**Faculty**	FASS - Faculty of Arts and Social Sciences

2	**Project title**	Black Friday Sales Prediction

3	**Please enter a brief summary of your project and its methodology in 250 words. Please include information such as your research method/s, sample, where your research will be conducted and an overview of the aims and objectives of your research.**	The purpose of this project is to predict the sales during Black Friday.The dataset consists of retail store sales transactions that were recorded . The dataset titled "Black Friday Sales Prediction" is accessed from Kaggle which is compiled by Salah Sammari. It can be used to understand the purchase patterns of the customer which helps to forecast sales. First step is to understand the data by visualization which provides the basic insights. Data needs to be preprocessed before applying any methodology. Feature construction, removal of missing or redundant values, etc. are done in this stage. Feature selection is done based on the correlation between the variables. Data is then standardized before modelling to make the comparison more meaningful. Since our aim is to predict sales, which comes under regression problem in the field of machine learning, we will be evaluating various machine learning algorithms such as random forest, linear regression and LSTM (Long short-term memory) based on our literature review. We then apply our learning from the evaluation to achieve the goal of this project which is to predict sales during Black Friday.

4	**Are you planning to join on to an existing Standard Study Protocol (SSP)? SSPs are overarching pre-approved protocols that can be used by multiple researchers investigating a similar topic area using identical methodologies. Please note, SSPs are only being used by 3 schools currently and cannot be used by other schools. Using an SSP requires permission and sign-off from the SSP owner**	NO

5	**Are you making an amendment to a project with a current University of Surrey favourable ethical opinion or approval in place?**	NO

6	**Does your research involve any animals, animal data or animal derived tissue, including cell lines?**	NO

8	**Does your project involve human participants (including human data and/or any human tissue*)?**	NO

9	**Will you be accessing any organisations, facilities or areas that may require prior permission? This includes organisations such as schools (Headteacher authorisation), care homes (manager permission), military facilities, closed online forums, private social media pages etc. This also includes using University mailing lists (admin permission). If you are unsure, please contact ethics@surrey.ac.uk.**	NO

10	**Does your project involve any type of human tissue research? This includes Human Tissue Authority (HTA) relevant, or non-relevant tissue (e.g. non-cellular such as plasma or serum), any genetic material, samples that have been previously collected, samples being collected directly from the donor or obtained from another researcher, organisation or commercial source.**	NO

11	**Does your research involve exposure of participants to any hazardous materials e.g. chemicals, pathogens, biological agents or does it involve any activities or locations that may pose a risk of harm to the researcher or participant?**	NO

12	**Will you be importing or exporting any samples (including human, animal, plant or microbial/pathogen samples) to or from the UK?**	NO

13	**Will any participant visits be taking place in the Clinical Research Building (CRB)? (involving clinical procedures; if only visiting the CRB to collect/drop-off equipment or to meet with the research team (i.e. for informed consent/discussion) select 'NO').**	NO

14	**Will you be working with any collaborators or third parties to deliver any aspect of the research project?**	NO

15	**Are you conducting a service evaluation or an audit? Or using data from a service evaluation or audit?**	NO

16	**Does your funder, collaborator or other stakeholder require a mandatory ethics review to take place at the University of Surrey?**	NO
17	**Does your research involve accessing students' results or performance data? For example, accessing SITS data.**	NO
18	**Will ANY research activity take place outside of the UK?**	NO
19	**Are you undertaking security-sensitive research, as defined in the text below?**	NO
20	**Does your project require the processing of special category1 data?**	NO
21	**Have you selected YES to one or more of the above governance risk questions on this page (Q10-Q20)?**	NO

22	**Does your project process personal data2? Processing covers any activity performed with personal data, whether digitally or using other formats, and includes contacting, collecting, recording, organising, viewing, structuring, storing, adapting, transferring, altering, retrieving, consulting, marketing, using, disclosing, transmitting, communicating, disseminating, making available, aligning, analysing, combining, restricting, erasing, archiving, destroying.**	NO

23	**Are you using a platform, system or server external to the University approved platforms (Outside of Microsoft Office programs, Sharepoint or OneDrive)?**	YES
23.a	**Please list the platforms, systems and/or servers that you are using in your research.**	https://Kaggle.com

23.b	**Are you collecting personal data via a platform, system or server external to the University approved platforms?**	NO
23.c	**Are you collecting special category data3 via a platform, system or server external to the Unviersity approved platforms?**	NO
23.d	**Are you collecting audio and/or video recordings?**	NO
23.e	**You must delete all research data including personal information from the external platform, system or server you are using upon completion of your research. Students: Your supervisor must support the use of the external platform, system or server and must agree to taking on the responsibility to ensure you delete the data upon completion of your study.**	Students: My supervisor has approved the platform, system or server I am using and I will delete all data from external platforms, systems or servers used upon completion of my research and my supervisor agrees to take on the responsibility to ensure this.

24	**Does your research involve any of the above statements? If yes, your study may require external ethical review or regulatory approval**	NO

25	**Does your research involve any of the above? If yes, your study may require external ethical review or regulatory approval**	NO

26	**Does your project require ethics review from another institution? (For example: collaborative research with the NHS REC, the Ministry of Defence, the Ministry of Justice and/or other universities in the UK or abroad)**	NO

27	**Does your research involve any of the following individuals or higher-risk methodologies? Select all that apply or select 'not applicable' if no options apply to your research. Please note: the UEC reviewers may deem the nature of the research of certain high risk projects unsuitable to be undertaken by undergraduate students**	NOT APPLICABLE - none of the above high-risk options apply to my research.

28	**Does your research involve any of the following individuals or medium-risk methodologies? Select all that apply or select 'not applicable' if no options apply to your research.**	NOT APPLICABLE - none of the above medium-risk options apply to my research.

29	**Does your research involve any of the following individuals or lower-risk methodologies? Select all that apply or select 'not applicable' if no options apply to your research.**	NOT APPLICABLE - none of the above lower-risk options apply to my research.

30	**Declarations**	• I confirm that I have read the University's Code on Good Research Practice and ethics policy and all relevant professional and regulatory guidelines applicable to my research and that I will conduct my research in accordance with these. • I confirm that I have provided accurate and complete information regarding my research project • I understand that a false declaration or providing misleading information will be considered potential research misconduct resulting in a formal investigation and subsequent disciplinary proceedings liable for reporting to external bodies • I understand that if my answers to this form have indicated that I must submit an ethics and governance application, that I will NOT commence my research until a Favourable Ethical Opinion is issued and governance checks are cleared. If I do so, this will be considered research misconduct and result in a formal investigation and subsequent disciplinary proceedings liable for reporting to external bodies. • I understand that if I have selected 'YES' on any governance risk questions and/or have selected any options on the higher, medium or lower risk criteria then I MUST submit an ethics and governance application (EGA) for review before conducting any research. If I have NOT selected any governance risks or selected any of the higher, medium or lower ethical risk criteria, I understand I can proceed with my research without review and

		acknowledge that my SAGE answers and research project will be subject to audit and inspection by the RIGO team at a later date to check compliance.

31	**If I am conducting research as a student:**	• I confirm that I have discussed my responses to the questions on this form with my supervisor to ensure they are correct. • I confirm that if I am handling any information that can identify people, such as names, email addresses or audio/video recordings and images, I will adhere to the security requirements set out in the relevant Data Protection Policy

II. Research Proposal

RESEARCH PROPOSAL

Name: Anusha Chatra Anilkumar

Student Number: 6720072

Supervisor: Dr Philip Murray

Research Topic: Black Friday Sales prediction

RESEARCH OVERVIEW

Sales forecasting is the practice of estimating how much income a business, group of people, or individual will produce over a certain period. Every organization benefit from using a sales forecast to guide decision-making. It aids in budgeting, risk management, and overall business planning. Businesses can accurately allocate resources and manage the cash flow. With the use of sales predictions, sales teams may discover early warning signs in their pipeline and make necessary course corrections before it's too late. It also aids companies in effectively estimating costs and income, which allows them to foresee both short- and long-term performance Mahalingam *(2021)*. A strong sales forecast will make it easier for potential investors to understand your performance and objectives. They would want to be certain that your business has a planned trajectory and that it generates monthly recurring revenue before they invest. Companies can grow and seize new market opportunities by investing in things like new locations, more staff, or revised plans. An accurate sales prediction will be able to foresee when enough money will be flowing in to cover the costs of the necessary expenditure Joran *(2022)*. A new business's launch also requires careful consideration of sales projections. To buy everything they need to launch, almost all new firms require loans or start-up capital: office space, furnishings, supplies, wages, and marketing. Every business aspires to help and please its clients and investors. A business may continue to support present customers, fund external marketing initiatives, staff enough customer service touchpoints, and more when expectations are met, and internal operations are running well Roos *(2008)*.

Even though there are multiple advantages of performing sales prediction, it comes with a lot of challenges. Sales estimate is severely impacted by unforeseen events. Depending on the company, location, client base, extreme weather, economic crises, civil upheaval, and worldwide pandemics (cough, COVID-19) can all significantly alter your sales projections. Product updates and rebranding also impacts sales prediction. There are high chances of overestimating or underestimating sales which may impact the company adversely. Consumers may be more or less likely to make purchases at times of the year, depending on company's sector, region, and target market. Hence seasonality must be taken into consideration while performing a forecast. If past data is available to use, creating a sales prediction is significantly simpler. Unfortunately, newly established businesses must rely on market analysis and competitive intelligence to make

their forecasts because they lack a significant amount of previous data Rohn (2022).

METHODOLOGY

The dataset consists of retail store sales transactions that were recorded. The dataset titled "Black Friday Sales Prediction" is accessed from Kaggle which is compiled by Sammari *(2020)*. It can be used to understand the purchase patterns of the customer which helps to forecast sales. First step is to understand the data by visualization which provides the basic insights. Data needs to be preprocessed before applying any methodology. Feature construction, removal of missing or redundant values, etc. are done in this stage. Feature selection is done based on the correlation between the variables. Data is then standardized before modelling to make the comparison more meaningful. Since our aim is to predict sales, which comes under regression problem in the field of machine learning, we will be evaluating various machine learning algorithms such as random forest, linear regression and LSTM (Long short-term memory) based on our literature review. We then apply our learning from the evaluation to achieve the goal of this project which is to predict sales during Black Friday.

SIGNIFICANCE OF RESEARCH

Forecasting sales will help the company to get an estimate of the revenue which helps them in budgeting, resource management and managing cash flows. It also helps to take up immediate actions on any foreseeable dip in sales. It also helps in coming up with an effective business plan. The company can make decisions based on the forecast. If the sales meet the target, the company can hire more staff and purchase inventory ahead of time. On the other hand, if sales do not meet the target, company can come up the remediation plans Kothandaraman *(2021)*. Effective sales projections position the company to accurately deliver and predict payments, as well as to obtain better conditions when applying for credit and corporate finance. Revenue estimates are based on predicted short- and long- term performance. Naturally, these advantages provide a strong business justification for sales forecasting Mahalingam *(2021)*.

Further analysis can be done in detail to know which other attribute can be added to get more insights which might improve the accuracy. More detailed product description can also benefit the research. Models can be created for different set of features to see any fluctuation in the result.

III. R Programming Code

```
#Installing libraries
install.packages("tidyverse")
install.packages("corrplot")
install.packages("PerformanceAnalytics")
install.packages("Hmisc")
install.packages("caret")
install.packages("factoextra")
install.packages("DataExplorer")
install.packages("jtools")
install.packages("caTools")
install.packages("randomForest")
install.packages("e1071")
install.packages("scales")
install.packages("ggplot2")
install.packages("Metrics")
#Loading the packages
library(tidyverse)
library(PerformanceAnalytics)
#rcorr function
library(Hmisc)
#To find correlation
library(caret)
#To plot correlation
library(corrplot)
#plot missing values
library(DataExplorer)
#Regression summary
library(jtools)
#To split data
library(caTools)
#Random forest algorithm
library(randomForest)
#SVM algorithm
library(e1071)
#To get label number
library(scales)
#To plot graphs
library(ggplot2)
#RMSE and MAE calculation
library(Metrics)



#Reading dataset
train_data <- read.csv("train.csv")

#Summary of the whole dataset
summary(train_data)
```

```
#Creating a subset of the dataframe with 50000 observations
newdata_train <- train_data[1:50000,]

#Extracting the dataset with 50k records for tableau visualisation
write.csv(newdata_train, "/Users/anusha/Desktop/new_train.csv",
    row.names=FALSE)

#Summary of 50k observations
summary(newdata_train)
#Checking the datatype of each varibale
str(newdata_train)

#Data Visualisation

#Converting categorical variables into factors
newdata_train <-newdata_train %>% mutate(Gender = factor(Gender),
                        Age = factor(Age),
                        City_Category = factor(City_Category),
                        Stay_In_Current_City_Years =
    factor(Stay_In_Current_City_Years),
                        Product_ID = factor(Product_ID))

#Initial data exploration
#Graph for gender ditribution
newdata_train %>%
  group_by(Gender) %>%
  summarise(Count = n(),Percentage=round(n()/nrow(.)*100,2)) %>%
  arrange(desc(Count))
values <- c(37812,12188)
labels <- c("Male", "Female")
colors <- c("steelblue","orange")
piepercent<- round(100 * values / sum(values), 1)
pie(values, labels = piepercent, main = "Gender pie chart",  col = colors)
legend("topright", c("Male", "Female"),
                cex = 0.5, fill = colors)
#In this dataset, 75.6% are male customers

#Graph to plot percentage of customers in each age group
newdata_train %>%
  group_by(Age) %>%
  summarise(Count = n(),Percentage=round(n()/nrow(.)*100,2)) %>%
  arrange(desc(Count))
#In this dataset, approximately 40% of the customers are aged between 26-35 and
    lowest is 3% who are aged between 0-17

#Graph to plot percentage of customers in each occupation category
newdata_train %>%
  group_by(Occupation) %>%
  summarise(Count = n(),Percentage=round(n()/nrow(.)*100,2)) %>%
```

```
  arrange(desc(Count))
#In this dataset, 13.5% of customers fall under Occupation category 4

#Graph to plot percentage of customers in each City
newdata_train %>%
  group_by(City_Category) %>%
  summarise(Count = n(),Percentage=round(n()/nrow(.)*100,2)) %>%
  arrange(desc(Count))
#In this dataset, 41% of the customers are from City B

#Graph to plot percentage of customers staying in the city for different years
newdata_train %>%
  group_by(Stay_In_Current_City_Years) %>%
  summarise(Count = n(),Percentage=round(n()/nrow(.)*100,2)) %>%
  arrange(desc(Count))
#In this dataset, 35% of the customers have lived for 1 year in the city

#Graph of Purchase vs City_Category
options(scipen = 999, repr.plot.width = 14, repr.plot.height = 8)
ggplot(data=newdata_train, aes(x=City_Category, y=Purchase, label = Purchase)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('City Category vs Purchase') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("City_Category") + ylab("Purchase")

#Graph of Purchase vs Gender
ggplot(data=newdata_train, aes(x=Gender, y=Purchase)) +
    geom_bar(stat="identity",fill="steelblue")  +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Gender vs Purchase') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Gender") + ylab("Purchase")


#Graph of Purchase vs Age
ggplot(data=newdata_train, aes(x=Age, y=Purchase)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Age vs Purchase') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Age") + ylab("Purchase")

#Graph of Purchase vs Occupation
ggplot(data=newdata_train, aes(x=Occupation, y=Purchase)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
```

```
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Occupation vs Purchase') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Occupation") + ylab("Purchase")

#Graph of Purchase vs Marital_Status
ggplot(data=newdata_train, aes(x=Marital_Status, y=Purchase)) +
    geom_bar(stat="identity",fill="steelblue") +
geom_text(aes(label= Purchase), vjust=-0.3, color="white", size=3 , position =
    position_dodge(0.9)) +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Marital_Status vs Purchase') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Marital_Status") + ylab("Purchase")

#Graph of Purchase vs Stay_In_Current_City_Years
ggplot(data=newdata_train, aes(x=Stay_In_Current_City_Years, y=Purchase)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Stay_In_Current_City_Years vs Purchase') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Stay_In_Current_City_Years") + ylab("Purchase")

#Graph of age vs product_category1
ggplot(data=newdata_train, aes(x=Age, y=Product_Category_1)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Age vs Product_Category_1') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Age") + ylab("Product_Category_1")

#Graph of age vs product_category_2
ggplot(data=newdata_train, aes(x=Age, y=Product_Category_2)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Age vs Product_Category_2') +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Age") + ylab("Product_Category_2")

#Graph of age vs product_category_3
ggplot(data=newdata_train, aes(x=Age, y=Product_Category_3)) +
    geom_bar(stat="identity",fill="steelblue") +
theme_minimal()+
scale_y_continuous(labels = scales::label_number_si()) +
ggtitle('Age vs Product_Category_3') +
theme(plot.title = element_text(hjust = 0.5)) +
```

```
xlab("Age") + ylab("Product_Category_3")


# create vectors to store the categories and mean purchases
category <- c()
mean_purchase <- c()

# get the unique categories and sort them
categories <- sort(unique(newdata_train$Product_Category_1))

# loop through the categories and calculate the mean purchase for each category
for (e in categories) {
  mean_purchase <- c(mean_purchase,
    mean(newdata_train[newdata_train$Product_Category_1 == e, "Purchase"]))
}
#Plot the graph of mean purchase for each category
ggplot(data.frame(category=categories, mean_purchase=mean_purchase)) +
  geom_col(aes(x=category, y=mean_purchase),fill="steelblue") +
  labs(title="Mean of the Purchases per Category", x="Product Category", y="Mean
    Purchase")

#Graph of mean purchase group by Gender, City_Category and Marital_Status
plotdata <- newdata_train %>%
  group_by(Gender, City_Category, Marital_Status) %>%
  summarize(n = n(),
            mean = mean(Purchase), #mean of purchase
            sd = sd(Purchase),
            se = sd / sqrt(n))

# create better labels for discipline
plotdata$Marital_Status <- factor(plotdata$Marital_Status,
                          labels = c("Married",
                                     "Single"))
# create plot
ggplot(plotdata,
       aes(x = Gender,
           y = mean,
           color = Gender)) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = mean - se,
                    ymax = mean + se),
                width = .1) +
  scale_y_continuous(labels = scales::label_number_si()) +
  facet_grid(. ~ City_Category + Marital_Status) +
  theme_bw() +
  theme(legend.position = "none",
        panel.grid.major.x = element_blank(),
        panel.grid.minor.y = element_blank()) +
  labs(x="",
       y="",
```

```
      title="Purchase by gender, City_Category, and Marital_Status",
      subtitle = "(Means and standard errors)") +
  scale_color_brewer(palette="Set1")

########################################

#Data Pre-processing

#Plot missing values
plot_missing(newdata_train)
#To find missing values in each variable
sapply(newdata_train, function(x) sum(is.na(x)))
#Product category 2 has around 31% of missing values
#Product category 3 has around 70% of missing values and hence it is better to drop
    the variable

#Replacing NA's with mean value of Product_Category_2
mean_prod_2 <- mean(newdata_train$Product_Category_2, na.rm = TRUE)
newdata_train$Product_Category_2[is.na(newdata_train$Product_Category_2)] <-
    mean_prod_2

#Summary of the dataset after imputation of missing values
summary(newdata_train)

#To find outliers in Purchase variable
purchase_outliers <- qplot(y = Purchase, ylab = "Purchase Outliers", data =
    newdata_train, geom = "boxplot", fill = I("gray")) + theme_minimal() +
    theme(axis.title.x = element_text(angle = 90))
purchase_outliers

#Feature Selection
#Dropping unrelated features
df <- subset(newdata_train, select = -c(User_ID, Product_ID,
    Product_Category_3))

#Check variables in the dataframe
glimpse(df)

#Create a duplicate dataframe to make changes
scaled_data <- df

#Converting categorical variables into numeric
scaled_data$Gender <- as.numeric(scaled_data$Gender)
scaled_data$Age <- as.numeric(scaled_data$Age)
scaled_data$City_Category <- as.numeric(scaled_data$City_Category)
scaled_data$Stay_In_Current_City_Years <-
    as.numeric(scaled_data$Stay_In_Current_City_Years)
glimpse(scaled_data)

#Correlation matrix
```

```
cor(scaled_data)

#correlation plot
res <- cor(scaled_data)
corrplot(res, type = "upper",
         tl.col = "black", tl.srt = 45)

#Correlation plot using heatmap
col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(x = res, col = col, symm = TRUE)

#Data modelling

#Set seed to replicate the results
set.seed(123)
#Split the data into training and test in the ratio 70:30
split = sample.split(df, SplitRatio = 0.7)
train_dataset1 = subset(df, split == TRUE)
test_dataset1 = subset(df, split == FALSE)

#Multiple Linear Regression
#Model with all variables
model1 <- lm(Purchase ~ Gender + Age + Occupation + City_Category +
    Stay_In_Current_City_Years + Marital_Status + Product_Category_1 +
    Product_Category_2, data = train_dataset1)
#Summary of the model1
summary(model1)

#Model without Stay_In_Current_City_Years
model2 <- lm(Purchase ~ Gender + Age + Occupation + City_Category +
    Marital_Status + Product_Category_1 + Product_Category_2, data =
    train_dataset1)
#Summary of the model2
summary(model2)

#Model without Stay_In_Current_City_Years and Occupation
model3 <- lm(Purchase ~ Gender + Age + City_Category + Marital_Status +
    Product_Category_1 + Product_Category_2, data = train_dataset1)
#Summary of the model3
summary(model3)

#Model without Stay_In_Current_City_Years, Occupation and Marital_Status
model4 <- lm(Purchase ~ Gender + Age + City_Category + Product_Category_1 +
    Product_Category_2, data = train_dataset1)
#Summary of the model4
summary(model4)

#Model without Stay_In_Current_City_Years, Occupation, Marital_Status and
    Gender
```

```
model5 <- lm(Purchase ~  Age + City_Category + Product_Category_1 +
    Product_Category_2, data = train_dataset1)
#Summary of the model5
summary(model5)

#Confidence intervals for model4
confint(model4)

#Predict purchase values using model4
y1 = predict(model4, train_dataset1)
#Show the purchase values
table(y1, train_data1$Purchase)
#Predict purchase values using model4
y_pred1 = predict(model4, test_dataset1)
#Show the purchase values
table(y_pred1, test_dataset1$Purchase)

#RMSE values foe training and test data
rmse(y1, train_dataset1$Purchase)
rmse(y_pred1, test_dataset1$Purchase)

#MAE values foe training and test data
mae(y1, train_dataset1$Purchase)
mae(y_pred1, test_dataset1$Purchase)

#Model 3 evaluation
#Predict purchase values using model3
y2 = predict(model3, train_dataset1)
#Show the purchase values
table(y2, train_data1$Purchase)
#Predict purchase values using model3
y_pred2 = predict(model3, test_dataset1)
#Show the purchase values
table(y_pred2, test_dataset1$Purchase)

#RMSE values foe training and test data
rmse(y2, train_dataset1$Purchase)
rmse(y_pred2, test_dataset1$Purchase)

#MAE values foe training and test data
mae(y2, train_dataset1$Purchase)
mae(y_pred2, test_dataset1$Purchase)


#------------------------------------------------------------------
#Random Forest Algorithm
#Set set reproducibility
set.seed(123)
#Apply random forest algorithm with 100 trees
rf1 <- randomForest(Purchase~.,data = train_dataset1, ntree = 100, mtry = 3)
```

```
#Print the results of rf1
print(rf1)
#Print the attributes of rf1
attributes(rf1)

#Predict purchase values using rf1
p11 <- predict(rf1, train_dataset1)
#Show the purchase values
cm_train1 <- table (p11, train_dataset1$Purchase)
cm_train1

#Predict purchase values using rf1
p21 <- predict(rf1, test_dataset1)
#Show the purchase values
cm_test1 <- table(p21, test_dataset1$Purchase)
cm_test1

#Graph of rf1
plot(rf1)

# RMSE on training set
rmse(p11, train_dataset1$Purchase)
# RMSE on test set
rmse(p21, test_dataset1$Purchase)
# MAE on training set
mae(p11, train_dataset1$Purchase)
# MAE on test set
mae(p21, test_dataset1$Purchase)

#Apply random forest algorithm with 500 trees
rf2 <- randomForest(Purchase~.,data = train_dataset1, ntreeTry = 500, mtry = 3)
print(rf2)

#Predict purchase values using rf2
p12 <- predict(rf2, train_dataset1)
#Show the purchase values
cm_train2 <- table (p12, train_dataset1$Purchase)
cm_train2

#Predict purchase values using rf2
p22 <- predict(rf2, test_dataset1)
#Show the purchase values
cm_test2 <- table(p22, test_dataset1$Purchase)
cm_test2

#Graph of rf2
plot(rf2)

# RMSE on training set
```

```
rmse(p12, train_dataset1$Purchase)
# RMSE on test set
rmse(p22, test_dataset1$Purchase)
# MAE on training set
mae(p12, train_dataset1$Purchase)
# MAE on test set
mae(p22, test_dataset1$Purchase)

#Apply random forest algorithm with 1000 trees
rf3 <- randomForest(Purchase~.,data = train_dataset1, ntree = 1000, mtry = 3)
print(rf3)

#Predict purchase values using rf3
p13 <- predict(rf3, train_dataset1)
#Show the purchase values
cm_train3 <- table (p13, train_dataset1$Purchase)
cm_train3

#Predict purchase values using rf3
p23 <- predict(rf3, test_dataset1)
#Show the purchase values
cm_test3 <- table(p23, test_dataset1$Purchase)
cm_test3

plot(rf3)

# RMSE on training set
rmse(p13, train_dataset1$Purchase)
# RMSE on test set
rmse(p23, test_dataset1$Purchase)
# MAE on training set
mae(p13, train_dataset1$Purchase)
# MAE on test set
mae(p23, test_dataset1$Purchase)
#---------------------------------------------------------------------

#Support Vector Machine Algorithm

#Radial kernel
svm_rbf1 <- svm(formula = Purchase~.,
                data = train_dataset1,
                type = 'eps-regression',
                kernel = 'radial')

#Summary of svm_rbf1
summary(svm_rbf1)

#Predict purchase values using svm_rbf1
pred11 = predict (svm_rbf1, train_dataset1)
pred11
```

```
#Predict purchase values using svm_rbf1
pred12 = predict (svm_rbf1, test_dataset1)
pred12

# RMSE on training set
rmse(pred11, train_dataset1$Purchase)
# RMSE on test set
rmse(pred12, test_dataset1$Purchase)
# MAE on training set
mae(pred11, train_dataset1$Purchase)
# MAE on test set
mae(pred12, test_dataset1$Purchase)

#Polynomial kernel
svm_rbf2 <- svm(formula = Purchase~.,
                data = train_dataset1,
                type = 'eps-regression',
                kernel = 'polynomial')

#Summary of svm_rbf1
summary(svm_rbf2)

#Predict purchase values using svm_rbf2
pred21 = predict (svm_rbf2, train_dataset1)
pred21

#Predict purchase values using svm_rbf2
pred22 = predict (svm_rbf2, test_dataset1)
pred22

# RMSE on training set
rmse(pred21, train_dataset1$Purchase)
# RMSE on test set
rmse(pred22, test_dataset1$Purchase)
# MAE on training set
mae(pred21, train_dataset1$Purchase)
# MAE on test set
mae(pred22, test_dataset1$Purchase)
```