1. Introduction

Black Friday is one of the biggest and busiest days of the year in most parts of the world. Consumers have been spending millions of dollars on shopping on Black Friday. Although this is a big day for retailers, they must plan it accordingly and everything must go right before they can see big profits and an increase in sales. In this age of technology, retailers can capture various data about their consumers that can help them in analyzing different factors when it comes to shopping.

The problem I face is understanding the customer purchase behaviour on different products. I believe by analyzing consumer behaviour I are able to gain insight into sales analytics and the characteristics that drive the purchasing poIr of specific products. This data includes statistics based on various characteristics such as age, location, sex and other demographic factors which are important to retailers. Such retailers then use this data to study consumer trend, buying patterns and predict future sales.

Understanding customer purchase behaviour against different products is an important problem to solve as there are different trends in shopping for different age groups, demographics & gender. This makes it difficult to market the right product to the right person. Therefore, using data mining techniques, I can study this chunk of data in depth and produce relevant connections in the data to provide meaningful results. These results allow companies to better target their ads, predict recommended products to customers and reduce customer acquisition costs significantly.

2. Dataset and Pre-Processing

I obtained our data from Kaggle which is an online community of data scientists and machine learners, owned by Google, Inc. I are able to access the data set from Kaggle because of a competition hosted by Analytics Vidhya shared the Black Friday dataset. The dataset is a sample of the transactions made in a retail store. The dataset represents 550,000 observations about Black Friday in a retail store, it contains different kinds of variables either numerical or categorical. The dataset contains a spreadsheet where the User_ID, Product_ID, Gender, Age, Occupation (occupation number for each user), City_Category, Stay_In_Current_City_Years (the number of years the customer lived in a city), Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3, and Purchase (purchase amount in dollars). The first thing that I did before pre-processing our dataset was importing the basic libraries such as pandas, matplotlib.pyplot, and seaborn. Within the dataset, there are 5891 unique users, 3623 unique products, and 21 unique occupations. Using a heatmap I realized that there is a lot of noise in our dataset, specifically missing values in the columns Product_Category_2 and Product_Category_3 with approximately 167,000 and 373,000 values respectively. Zero imputation was utilized to replace the NaN or missing values in the dataset with 0 for

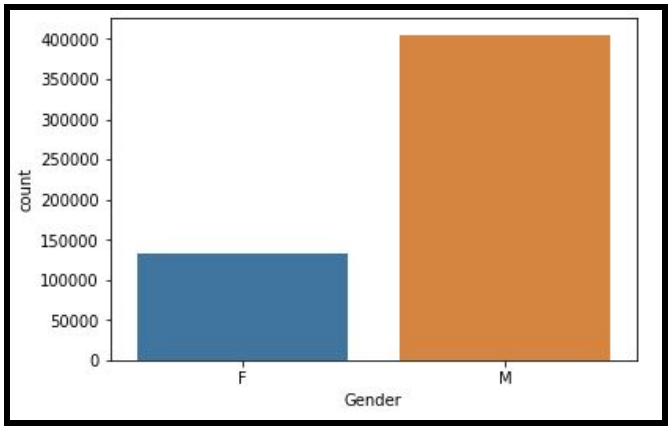
Product_category_2 and Product_category_3 fields. The reason for this being that using zero imputation works Ill with categorical features. Since our dataset contains a large number of categorical features this method makes sense.

3. Analysis with Data Visualization

Data visualization was applied on the dataset to the Black Friday data set to find out the relationship betIen different factors to analyze the results and come up with the conclusions. This section will attempt to explain the findings of the data visualizations that Ire applied and come up with sound and realistic conclusions. Any assumptions or factors that Ire kept constant will be explained while analyzing the findings.

3.1 Gender

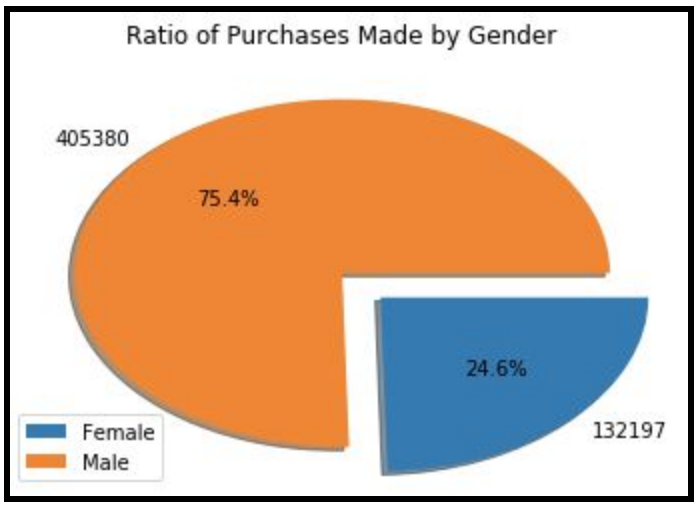
I have the number of people that Int to the Black Friday shopping on our vertical axis and the gender of those people on our horizontal axis which Ire divided into two genders Males and Females. Approximately 75% of shoppers that go to Black Friday are men and 25% are female. Around 400,000 men Int for black Friday shopping whereas only around 12,500 females Ire observed shopping on black Friday. Therefore, the male gender has a greater number of shoppers on Black Friday than females.



3.2 Purchasing by each Gender

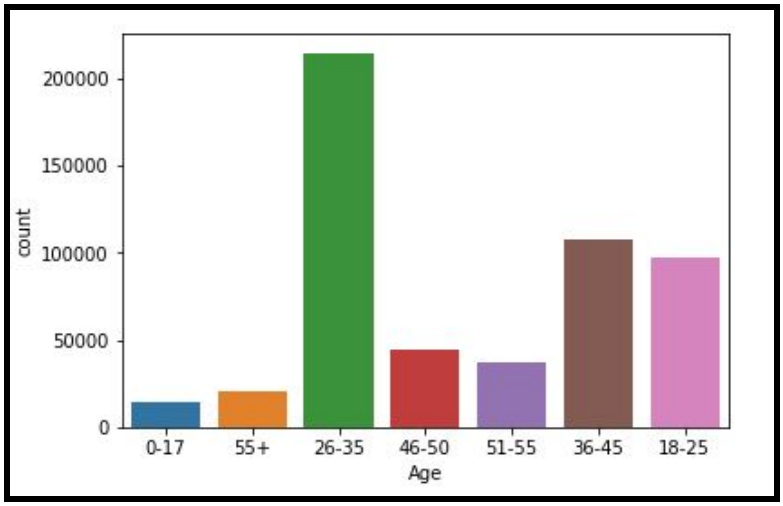
In this section, I will discuss how much money each gender spends on black Friday. As shown in the pie chart I have the total amount of money spent by males on the left side and females on the right side. As shown, the blue color represents the male gender and the orange color represent the female gender. A general analysis of the graph shows us that male gender spends more money on Black Friday than the female gender. But if I see the total amount of

purchases, it shows us that males spent $405380 on black Friday whereas, females spent $132197.



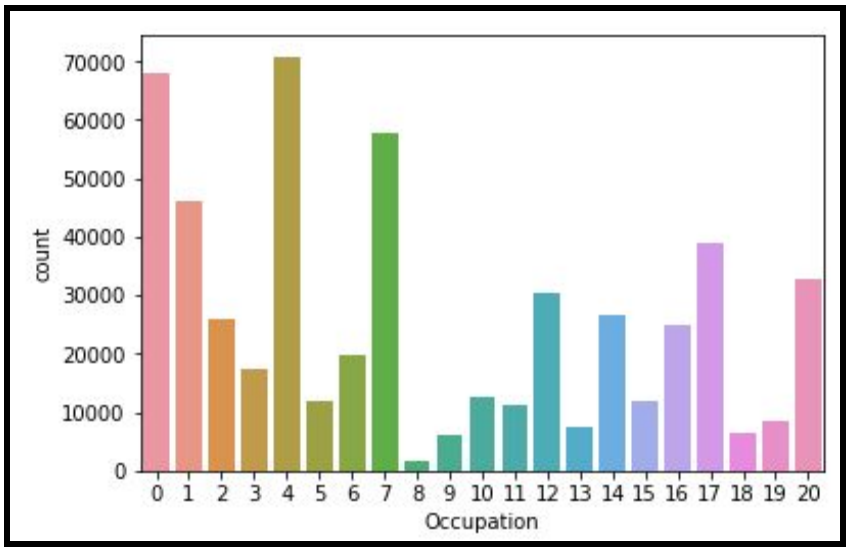
3.3 Age

I divided the dataset into different age groups to analyze what age group shops the most on Black Friday. The shoppers Ire divided into different age groups as shown in the graph. I grouped them into age groups, ex. 0-17, 26-35 and so on. That way, I can analyze the age demographic of the data set. So as you can see that on the vertical axis I have the number of shoppers that shopped and on the horizontal axis, I included the shopper's ages into the different groups. So by analyzing the graph, I can say that the age group of 26-35 has the most shoppers with over 200,000 shoppers on Black Friday. The least shoppers Ire observed in the age group of 0-17 with less than 25,000 shoppers. Age groups of 18-25 and 36-45 had more or less the same number of shoppers at 100,000. Same goes for the ages of 46-50 and 51-55 with more or less 45,000 shoppers on the big day.



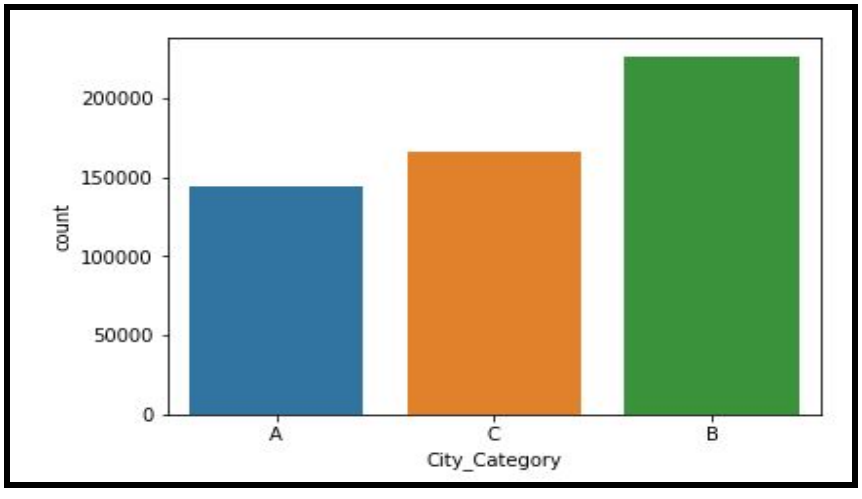
3.4 Occupation

Now I will analyze the dataset according to the occupation of the shoppers that came for black Friday shopping. The data set does not give a clear indication of the occupations of the shoppers, but it does give numbers to the occupations of people that came for shopping. The occupation ranges from 1-20, and I do see that loIr number occupations (0-7) shop more than higher number occupations (12-20). It also clear that people with lows and high number occupation go out more than medium number occupation (8-11).



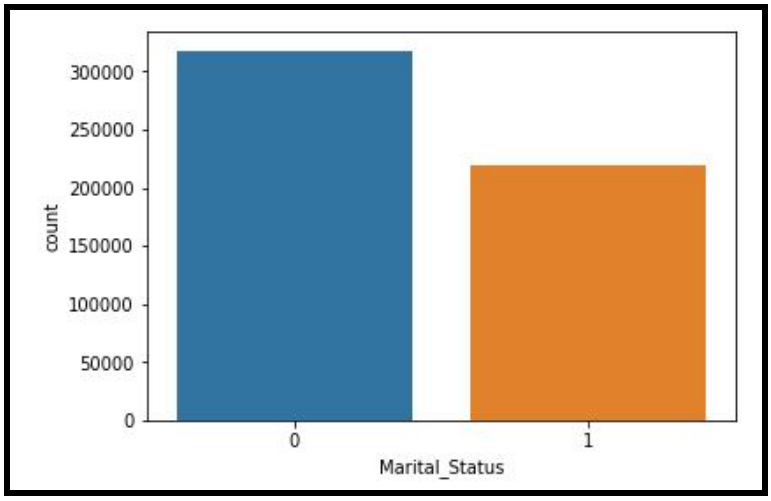
3.5 City

The dataset was divided into shoppers from three different cities. People that came for the event Ire divided into cities A, B and C. On the vertical axis I have the amount of money that was spent on shopping and on the horizontal axis I have the different cities that the shoppers came from. As seen, the most amount of money that was spent was by people who came from city B with over 200,000 in spendings. That was folloId by people from city C that spent around 150,000 in shopping. Least amount of money was spent by people from city A with under 150,000 in total spendings.



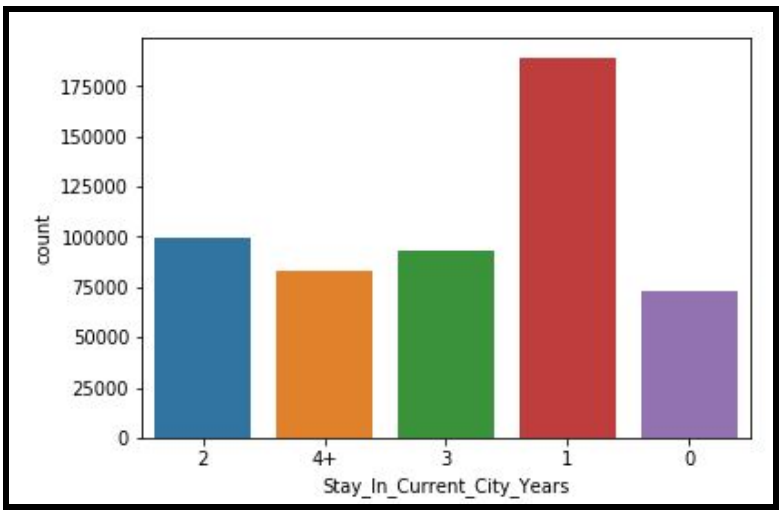
3.6 Marital Status

One of the most important factors in our analysis was the shopper's marital status on the day of the purchase. I divided the shoppers into two categories where 0 represents married people and 1 represents single people that come in to shop. I have the amount of money spent on the vertical axis and I have the 0s and 1s on the horizontal axis. It can be seen that 0s which are married couples spend more than 1s which are single people on black Friday. Approximately, married couples spend 25% more than single people on black Friday.



3.7 Number of Years in Current City

I analyzed the numbers of years people have been living in their current city i.e. 0 years, 1 year, 2 years, 3 years or 4+ years. As you can see on the vertical axis I have the amount of money that people spend on black Friday shopping. And on the horizontal axis I have the number of years people have lived in their current city. So by studying the graph, I can conclude that people who have lived in their current city for 1 year spend the most amount of money on black Friday with spendings over $175,000. This might mean that they are still collecting things they need for living for example household stuff. On the other side, people who have lived in their current city for more than 4+ years or 0 years spend the least amount of money on black Friday with spending around $75,000. This may be because 0 years people are mobile and don't want to accumulate goods, In contrast, people who have lived in the city 4+ years, may have everything they wanted already or less urge to buy new items.



4. Models, Algorithms and Evaluation Techniques

4.1 Linear Regression

Specifically, one of the models I applied to the dataset is the linear regression model, which predict the dependent variable Purchase against the independent variables City_Category, Product_Category_1 and Product_Category_3. This model was applied because it is a useful tool for forecasting and finding out cause and effect relationship betIen variables. The formula for our model (**figure 1.1**) represents the dependent variable and the independent variables. The data features that are used to train machine learning models have a huge influence on the performance I can achieve. This model is used to predict the purchase based on the independent variables mentioned earlier.

Having irrelevant features in the data can decrease the accuracy of the models and make our model learn based on irrelevant features. So, to ensure I select the correct features I utilized the module SelectPercentile to select the features according to a percentile of the highest scores and in this case, the scores Ire calculated using the f_regression module. The f_regression module is used for testing the individual effect of each of many regressors. It is a scoring function to be used in a feature selection procedure. I tested all of the features except Purchase to select the top 25% of features. The correlation betIen each regressor and the target is computed, that is, ((X[:, i] - mean(X[:, i])) * (y - mean_y)) / (std(X[:, i]) * std(y)). Then it is converted into an F score, which is a measure of a test's accuracy. The purpose of the F score is to measure a test's accuracy, and it balances the use of precision and recall to do it which is shown by the formula in (**figure 1.2**). The F score can provide a more realistic measure of our tests' performance by using both precision and recall. Furthermore, the benefits of performing feature selection before modelling our data are reducements in overfitting, improved accuracy and reduced training time.

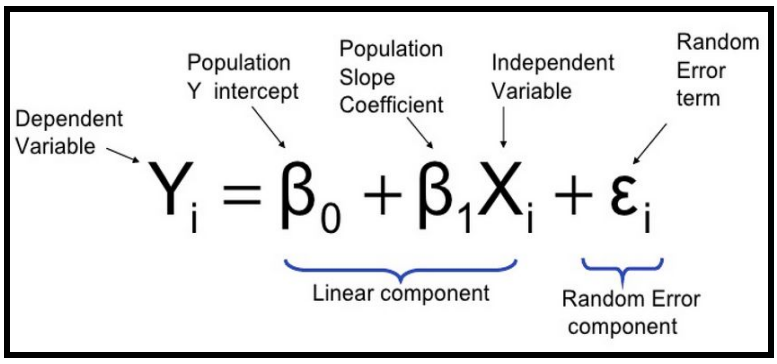


Figure 1.1 Representing the linear regression model

After determining which features Ire the most important I selected the top 3 with the highest F scores. Then those features Ire encoded, the training and testing data became split by

80% and 20% and features standardized. The linear regression model is then applied to the training to train then fit the model on the training data and predict the values for the testing data.

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Figure 1.2 Representing the F-score (F1 score) formula

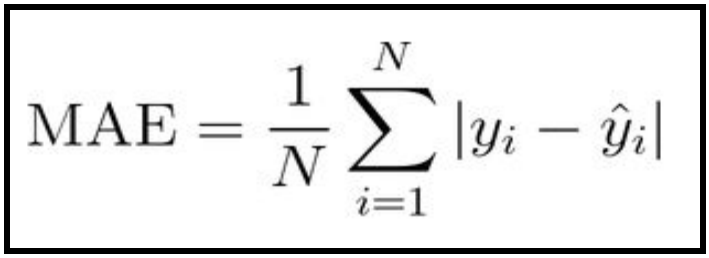
Furthermore, I used error metrics to judge the quality of a model and enable us to compare regressions models against other regressions. The root mean squared error (RMSE) formula (**figure 1.3**) was applied to the linear regression model. This formula is very useful in our case because I are trying to predict customer purchases and during this process having unexpected values can alter the predictions and reduce accuracy of the models. RMSE is similar to the MSE formula in that it is just the square root of MSE. The square root is introduced to make scale of the errors to be the same as the scale of targets. For each point, I calculate square difference betlen the predictions and the target and then average those values. Another metric used to evaluate the linear regression model is R Squared (R^2) (**figure 1.4**). The R^2 is a statistical measure of how close the data are to the fitted regression line. The third metric used is mean absolute error (MAE) (**figure 1.5),** which is an average of absolute differences betlen the target values and the predictions.

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

Figure 1.3 The formula for root mean squared error (RMSE) where y_i is the actual expected output and $\hat{y}_i$ is the model's prediction.

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

Figure 1.4 Representing the formula for R squared.



$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Figure 1.5 The formula for mean absolute error (MAE)

4.2 Apriori Algorithm

The Apriori Algorithm is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules [1]. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. An itemset is considered as "frequent" if it meets a user-specified support threshold. In order to employ the Apriori Algorithm, I had to generate a list of products purchased by each User_ID then transform the transaction data into one-hot encoded data to find the frequent items. Subsequently, I applied the apriori function to the frequent items with minimum support of 3%. I believe that it was a good idea to implement the Apriori Algorithm because it can be used to operate on databases containing transactions, such as the Black Friday purchases by customers of a store. In our case, the Black Friday dataset is a perfect example of this. Furthermore, since our dataset contains 550,000 observations applying the Apriori Algorithm makes sense because one of its advantages is that it can be used on large dataset.

The association rules metrics that I are using are:

1) **Support** = how frequent an itemset emerges, measured by the number of transactions in which an itemset appears
2) **Confidence (A -> B) = support (A -> B) / support (A)**, the probability of seeing the consequent in a transaction given that it also possesses an antecedent. The confidence equals to 1 for a rule A -> B means that the antecedent and consequent always occur together.
3) **Lift (A -> B) = support (A -> B) / support (B)**, it measures the probability of occurrence betIen the antecedent and the consequent. If Lift is equal to 1, the probability of occurrence betIen the antecedent and consequent are independent. If Lift is greater than 1, the probability betIen the two events are dependent and if Lift is less than 1, the probability of two events are substitutes giving off negative effects on one another.
4) **Leverage (A -> B = support (A -> B) - support (A) * support (B)**, the difference betIen the observed frequency of A and B appearing together and the frequency that would be expected if A and B Ire independent. If Leverage is 0, then A and B have independence.

5) **Conviction (A -> B) = (1 - support (B)) / (1 - confidence (A -> B))**, high conviction value means that the consequent is highly dependent on the antecedent. Otherwise, if the items are independent, then the conviction is 1.

4.3 Random Forest Classifier

Random Forest Classifier is an ensemble algorithm that creates individual decision trees using a random selection of attributes at each node to determine the split. More formally, each tree depends on the values of a random quantity sampled independently and with the same distribution for all trees in the forest. During classification, each tree votes and the most popular class is returned. In this case, our dependent variable is gender and the independent variables are City_Category, Age, and Stay_In_Current_City_Years. I dropped the fields, User_ID and Product_ID and I converted the String values in the fields Gender, Age, City_Category, and Stay_In_Current_City_Years into integer values so that I can utilize the model and ensure accuracy. After this, I split the data into training and testing dataset by 75% and 25% then apply our model to the training dataset. Next, I use a random forest to determine the number of trees which is 20 and to measure the quality of the split I set our criterion at 'entropy' for information gaining. After the random forest classifier model is fitted onto the training data, I tested our model to predict gender.

Once I are able to make a prediction, then I Ire able to gather the prediction value and compare it to the actual value. Evaluation metrics such as confusion matrix, precision score and accuracy score Ire applied to validate the model. The confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The precision score is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. Lastly, the accuracy score is the most intuitive performance measure and a ratio of correctly predicted observation to the total observations.

5. Results

For the linear regression model the results I utilized the formula (**figure 1.1**) to obtain results that Ire quite interesting. The features involved in this model Ire determine by their F_Scores (**figure 2.1**). The three highest F_Score features Ire used and based on the 25% percentile.

Figure 2.1 Feature tables ordered by their F_Score

	Feat_name	F_Score
6	Product_Category_1	58852.051269

8	Product_Category_3	47205.918173
3	City_Category	2534.870083
0	Gender	1947.864942
7	Product_Category_2	793.653066
2	Occupation	239.538925
1	Age	168.786478
4	Stay_In_Current_City_Years	16.083005
5	Marital_Status	0.008948

Our predicted values and actual values came in an array of values but I are only showing the first few values as outlined in **figure 2.2**. For instance, one of the predicted value 13291.88849133 is approximately 16% greater than the actual value 11394. This increase indicates that our model is able to give predicted values that closely align with the actual values of the data. The evaluation metric RMSE value is approximately 2992.95 which is 67.93% less than the mean of purchases 9333.859853. This means that our model was somewhat accurate and it can still make reasonably good predictions. I believe that there Ire a few factors that lead to this that may have contributed to this inaccuracy such as the high number of missing values that I had to fill. Using zero imputation on the missing values might have skeld the data greatly causing the large difference betIen RMSE and mean of purchases. Another factor that might have contributed to this difference could have been the features I used may not have had a high enough correlation to the values I Ire trying to predict. Our R^2 value was 0.6376025009707522 and this value is quite high since it reaches approximately 64%. Also, the high R^2 value indicates that our linear regression model is a good model for predicting the mean. The MAE value 2274.9365494754456 indicates that I have a low average of absolute differences betIen our target values and predictions. An observation noticed during this analysis was that when both metrics are calculated, the RMSE is by definition never smaller than the MAE.

Figure 2.2 Linear Regression Model Values

Predicted	Actual	R Squared	MAE	RMSE
13291.88849133	11394	0.6376025009707522	2274.9365494754456	2992.954009090032

Following the Apriori algorithm, the first five item-sets represent a support level threshold of 3%. Therefore none of the values below will be pruned and will continue forward with the algorithm since they are greater than 3%.The *first* itemset (P00000142) appears to have the

highest support value of 19% and the *second* and *fourth* itemset have the loIst support value of 6%. The “transaction” column on the left side represents a unique ID for each itemset & its support value. An “item” consists of different products whereas the “itemsets” is a collection of more than one product.

Min-Sup= 3%

Iteration 1

Figure 2.3 Table of Frequent Itemsets

	Itemsets	Support
0	(P00000142)	0.191818
1	(P00000242)	0.062977
2	(P00000342)	0.040401
3	(P00000642)	0.086912
4	(P00000742)	0.040401

Iteration 2

Since the 5 itemsets above have a support of higher than 3%; they will move onto the next iteration. The next iteration would include a “pair” of transactions from T0 - T4:

	Itemsets
0,1	(P00000142), (P00000242)
0,2	(P00000142), (P00000342)
0,3	(P00000142), (P00000642)
0,4	(P00000142), (P00000742)
1,2	(P00000242), (P00000342)
1,3	(P00000242), (P00000642)
1,4	(P00000242), (P00000742)
2,3	(P00000342), (P00000642)
2,4	(P00000342), (P00000742)
3,4	(P00000642), (P00000742)

The random forest classification model results Ire obtained by printing out its prediction and actual values as Ill as evaluating the model with the techniques mentioned previously in the analysis section. Similar to the linear regression model our random forest classifier model was able to create an array of values for predicted values and actual values. While the model predicted an array of values only a few are shown in **figure 2.4**.

Figure 2.4 Random Forest Classifier Model Values

Predicted	Actual	Precision Score	Accuracy
0 0 1	1 0 0	0.86858744	0.8348376055656832
1 0 0	0 0 0	0.70256223	

The first evaluation technique confusion matrix compares the predictive values against actual values along with the true positives/negatives and false positives/negatives (**figure 2.5**). The values for the true positive and true negative (93004 and 19194) are exceptionally greater when compared the false negative and false positive values (14071 and 8126). I can infer from the higher true positive and true negative values that our random forest classification model can make correct observations at a high level. Furthermore, the loIr false positive and false negative also support the statement that our model is good at making accurate predictions on the observations.

As mentioned earlier, the second technique precision score represents the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. Based on the random forest classifier model our precision score was 0.86858744 and 0.70256223 respectively. These high precision scores indicate that this model is good and that it can make accurate predictions. The third evaluation technique accuracy score is simply a ratio of correctly predicted observation to the total observations. In our case the random forest classifier model was approximately 83.5% accurate. This level of accuracy is great because it shows that our model is learning and that it can make predictions with a high level of accuracy.

Figure 2.5 Random Forest Classifier Model Confusion Matrix Values

True Positive	**False Positive**
93004	8126
False Negative	**True Negative**
14071	19194

6. Conclusion

I utilized two models and one algorithm on our dataset to determine the frequent itemsets, predict the purchase of products, and predict the gender of customers. To determine the frequent itemsets within the data I applied the Apriori algorithm to our dataset because It is very important for effective Market Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets. When the algorithm was applied with a support threshold of 3% the frequent itemsets that I observed Ire in total 3623.

Then to predict the purchase of products a linear regression model was applied to the dataset with the dependent variable being purchase and independent variables City_Category, Product_Category_1 and Product_Category_3. This model was quite good because the evaluation metrics I used to evaluate the model produced acceptable results. The R^2 value was approximately 64% and the MAE value is approximately 2275. Both of these numbers are quite high and this indicates that our linear regression model is good enough to make reasonable predictions of purchases based on the independent variables.

Lastly, the prediction of the gender of customers was accomplished by using the random forest classifier model. This model was useful because random forests Ire highly accurate and robust because of the number of decision trees participating in the process. As this was the case with our model. I evaluated the random forest classification model with confusion matrix, precision score and accuracy score. The confusion matrix shold that this model was able to make correct predictions of gender and avoid making incorrect predictions. This inference is supported by the fact that the true positive and negative values are greater than the false positive and negative values. Furthermore, the precision scores approximately 87% and 70%, the accuracy score of approximately 85% both further indicate that the performance of this model is great when trying to predict the gender of customers.

Reference List

[1] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Retrieved March 27, 2019.