Northeastern University
College of Professional Studies

Final Report

Course #: ALY 6040
Course Name: Data Mining Applications
Professor: Justin Grosz

Report submitted by **Yesha Nagar**
Ana Paniagua
Prisca Kakpo
Harsh Samani

Academic Term: Winter 2019 - Term B
Date of Submission: 16/05/2019

Introduction

Black Friday is the most popular time of the year where shoppers take advantage of shopping deals just before the holidays. Retailers offer a significant number of discounts to attract buyers. The origin of Black Friday is based on an accounting term when records were kept in ink with red signifying a loss in profits and black signifying a profit. Retailers generally operate in the red (unprofitable) throughout the year and depend heavily on the holiday season sales to end the year in the black with a profit.

The purpose of this project is to observe and analyze the consumer behavior and trends of a Black Friday customer. The dataset contains variables that have both numerical and categorical data. The variables include userID, productID, gender, age, occupation, city category, stay in current city years, marital status, product category 1, product category 2, product category 3 and purchase. This dataset will be useful in determining consumer purchase behaviors and how they interact with different products. Performing analytical operations such as selecting models on this dataset could be challenging as we explore the retail sector of numerous brands, customer behavior, sales and profit margins, the trend line of most popular products and extensive data for the shopping industry.

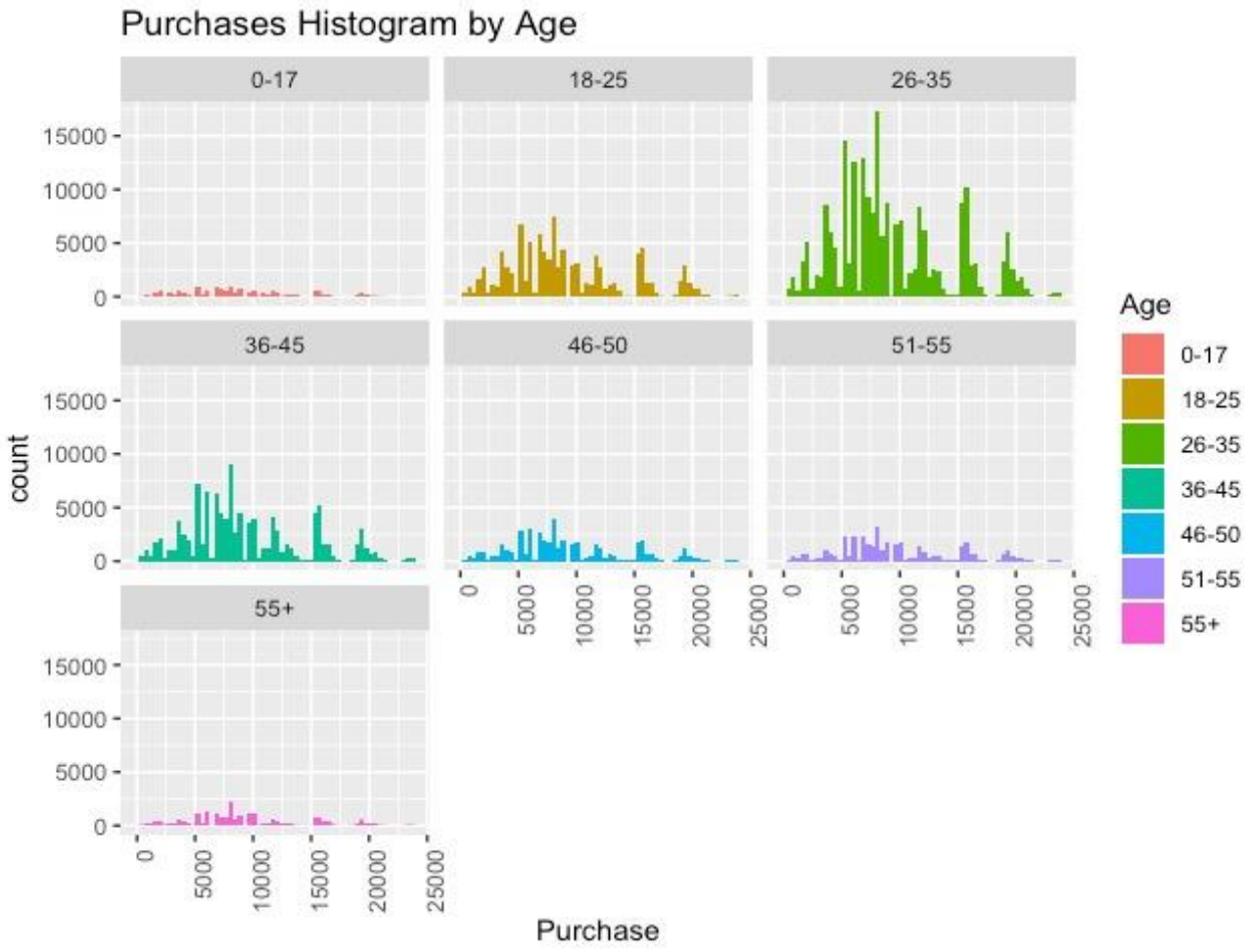
Although there are significant price cuts during this time, retailers are able to make profits, as buyers will purchase in larger quantities. As members of the group are more familiar with Black Friday we decided to work with the dataset. Additionally, we ensured that the dataset has attributes that will require us to clean the data. We found that there were missing values in the dataset which will be ideal for data cleaning.

The group decided to use a dataset from Kaggle titled, "A Study of Sales Trough Consumer Behavior". The dataset represents a sample of 550,000 retail transactions in a store. The dataset

contains variables that have both numerical and categorical data. The variables include userID, productID, gender, age, occupation, city category, stay in current city years, marital status, product category 1, product category 2, product category 3 and purchase. This dataset will be useful in determining consumer purchase behaviors and how they interact with different products. Performing analytical operations on this dataset could be very challenging as we can explore the retail sector of numerous brands, the behavior of customers, the margin of sales and profit, trend line of most popular products and a lot of data for the shopping industry.

Analysis

This is a huge dataset and it does not require much of cleaning. There are two columns that contain N/A values in a huge amount. Product Category 2 contains 166,986 NA values and Product Category 3 had 373,299 NA values. That is 30% and 67.8% of data respectively. Hence, we discard these columns as they may cause a problem during regression and make a poor influence on data. Other columns such as User ID and Product ID have redundant data but counts of only 0.01% of the total dataset and hence it can be ignored.

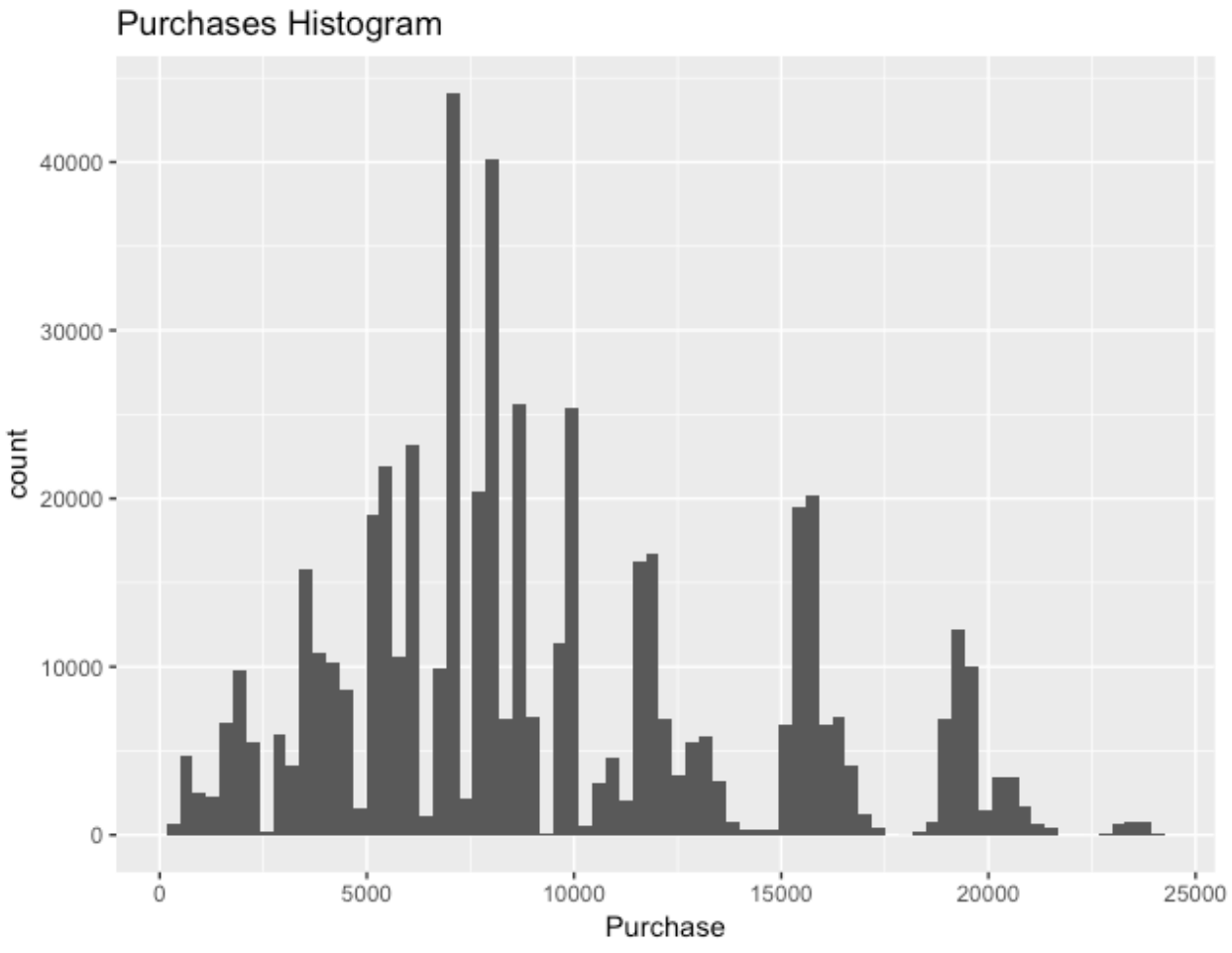


Purchases made by age are observed to notice trends in consumer behavior. Demographic data is necessary to analyze so as to determine if this influence purchasing behavior. The diagram above shows the purchases made by the age group. The age groups with the least amount of purchases are between ages 0-17 and over the age of 55. Based on the distribution above, it can be deduced that the age group with a significant purchasing power is from ages 26-35 years followed by the age group 36-45 years. By observing the age groups with a significant buying power, retailers will be able to maximize sales by marketing their products to a targeted age group of buyers 26-35 years which yield higher sales.

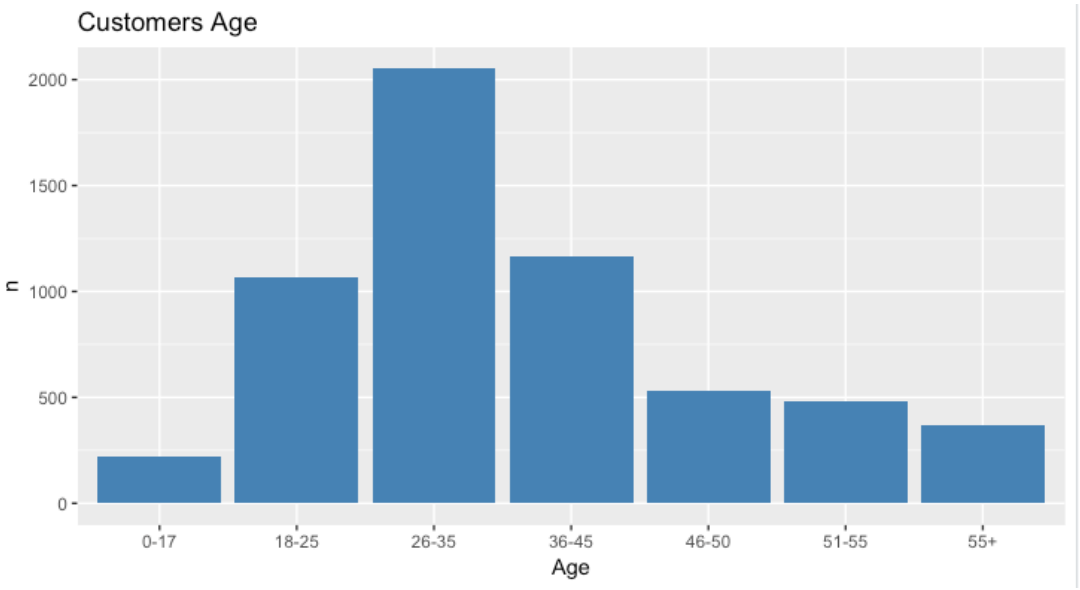
Initial Findings

Before conducting our analysis, the structure of the dataset was observed, missing values were plotted in the dataset using the data explorer library function. Histograms from the following were plotted: age, gender, city, the gender of buyers in the cities and occupation categories. The mode was also calculated for product category 1 and 2 to assess which product categories were more popular.

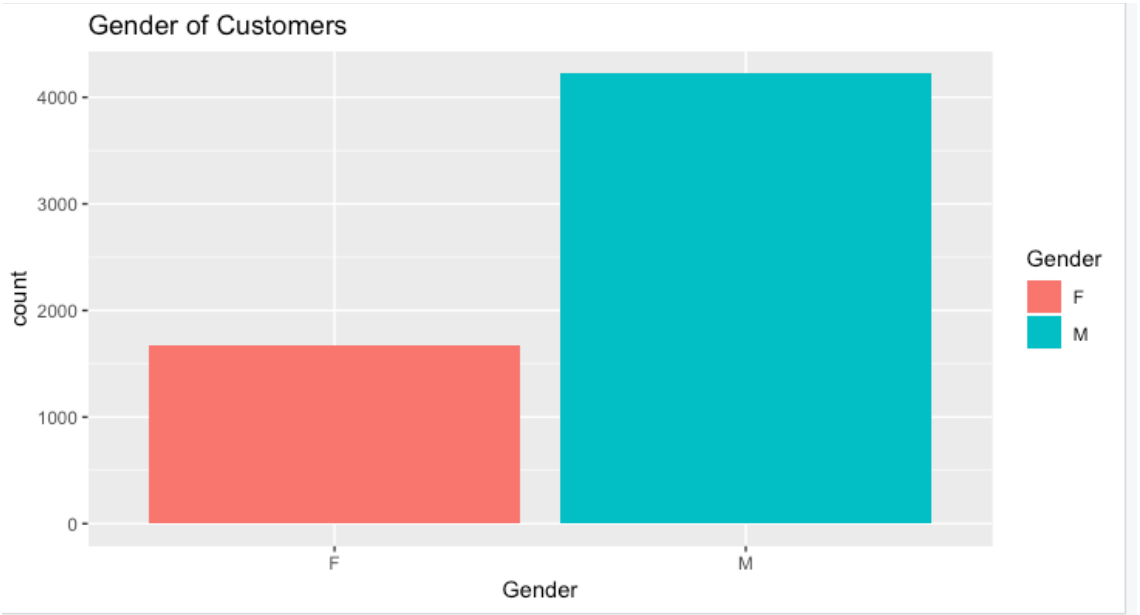
The following graph shows the histogram of purchases. It can be seen that the majority of customers spend between $6,000 and $10,000 in the purchase of items.



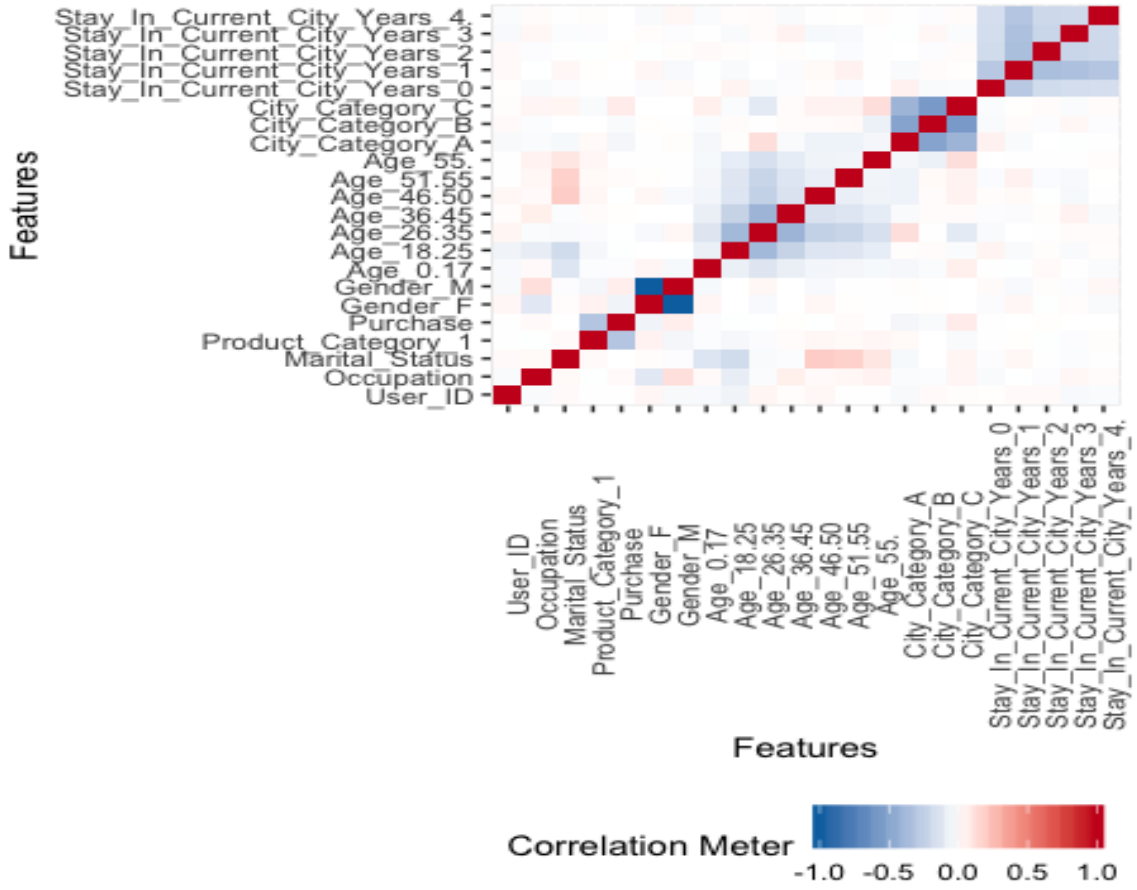
In our EDA we found that the majority of our customers fall under the age range of 26 -35. Followed by the age range of 36 - 45 and 18 -25. This is an important finding as it tells us what age range most of our frequent customers fall under.



During our EDA we wanted to understand more about the Black Friday shoppers, we are able to find that most of the customers were males and they almost double the female customers.



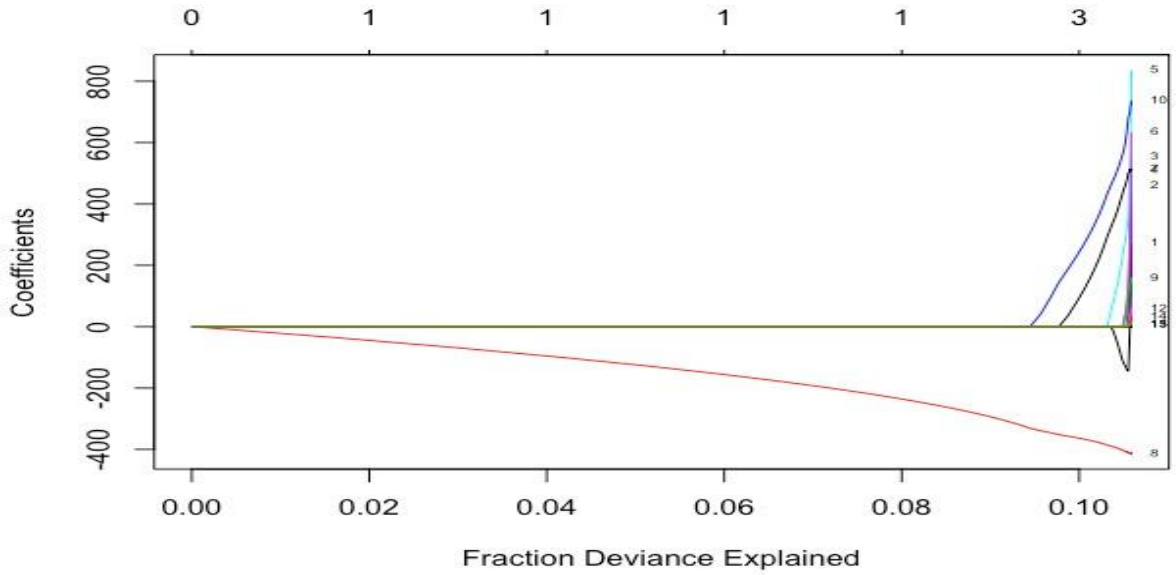
The graph below shows the correlation plot of all the variables.


Features
Stay_In_Current_City_Years_4.
Stay_In_Current_City_Years_3
Stay_In_Current_City_Years_2
Stay_In_Current_City_Years_1
Stay_In_Current_City_Years_0
City_Category_C
City_Category_B
City_Category_A
Age_55.
Age_51.55
Age_46.50
Age_36.45
Age_26.35
Age_18.25
Age_0.17
Gender_M
Gender_F
Purchase
Product_Category_1
Marital_Status
Occupation
User_ID
User_ID
Occupation
Marital_Status
Product_Category_1
Purchase
Gender_F
Gender_M
Age_0.17
Age_18.25
Age_26.35
Age_36.45
Age_46.50
Age_51.55
Age_55.
City_Category_A
City_Category_B
City_Category_C
Stay_In_Current_City_Years_0
Stay_In_Current_City_Years_1
Stay_In_Current_City_Years_2
Stay_In_Current_City_Years_3
Stay_In_Current_City_Years_4.
Features
Correlation Meter
-1.0 -0.5 0.0 0.5 1.0

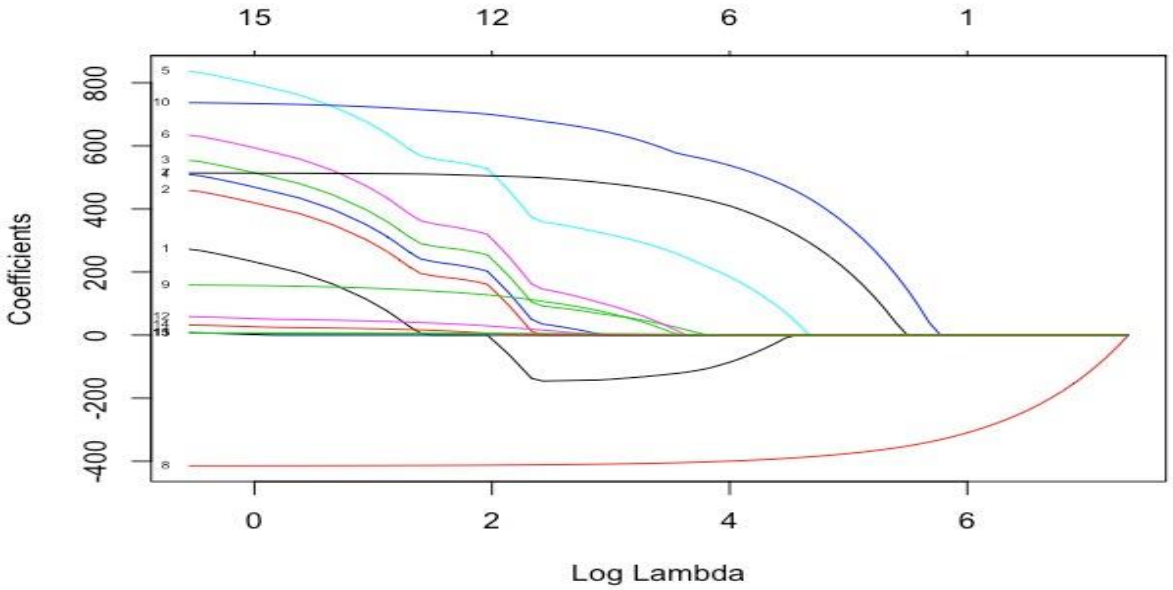
Model Selection

We decided to perform the multi linear regression model and lasso regression models in addition to the regression model. Prior to selecting the multi-linear regression model, we considered performing the logistic regression model on the dataset, we found that this cannot be performed. The logistic regression model cannot be performed on the dataset because it is more appropriate for datasets that can be used to explain the relationship between a dependent binary variable and one or more nominal or ordinal independent variables. Unfortunately, the black Friday dataset does not possess these attributes. Additionally, a heat map of the data is not possible as the dataset needs to be an integer. However, a correlation plot was performed to observe the variables that are highly or least correlated with the dataset.

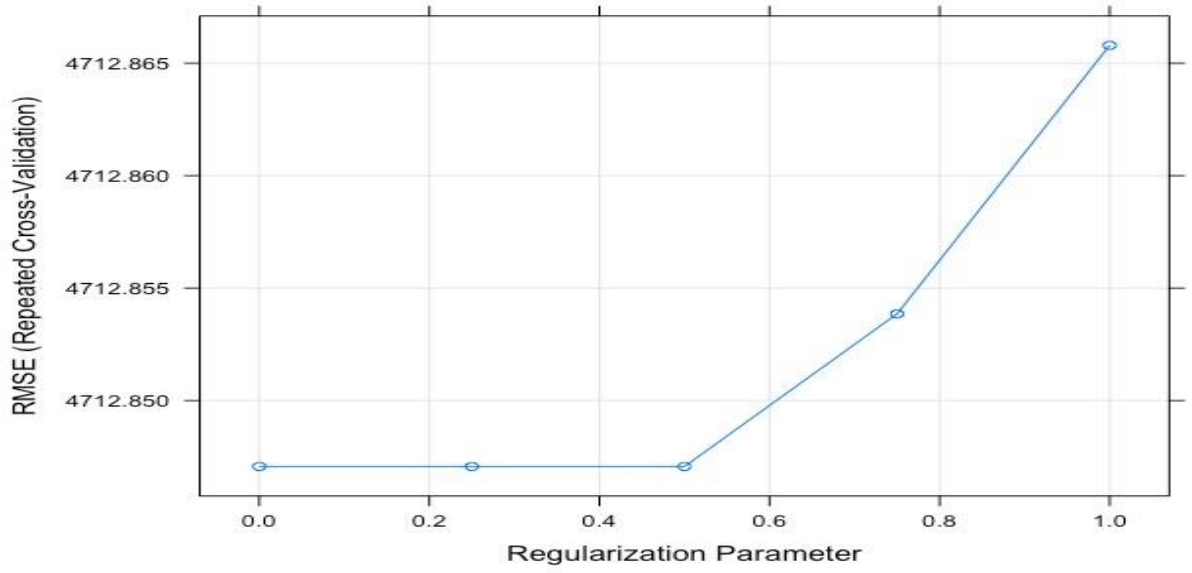
Due to the nature of the dataset, a lasso regression model was used to analyze the dataset as overfitting problems were encountered. The lasso regression model is appropriate for datasets that have multicollinearity which describes instances where near-linear relationships exist among the independent variables. The lasso regression model shrinks or regularizes coefficients in order to improve prediction accuracy.



The graph above shows that about 10% of the deviance which is similar to R-squared is explained by three variables. The coefficient keeps on increasing after 0.8 and there is a sudden downfall and coefficient becomes highly inflated and the reason behind this is overfitting. Overfitting which fails to describe the relationships between variables. The overfitting problem depicted in the dataset will be addressed via regularization.



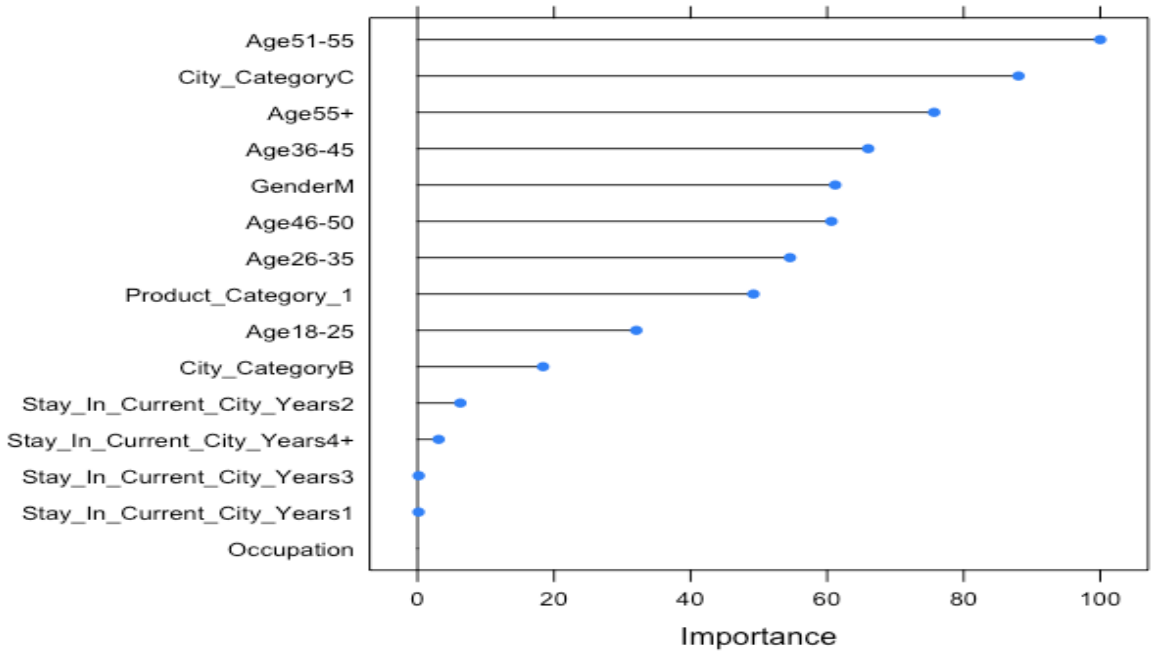
Based on the above graph we can see that log lambda above 8 the coefficient is zero and we start relaxing lambda the coefficient starts to grow. As the coefficient starts to grow, we can observe that some value becomes larger and larger on the top of the chart we can see that there is no independent variable.



The above graph is calculated based on Root Mean Square Error (RMSE) and it is calculated based on Repeated Cross-Validation. As we can see for higher values of lambda the error increases so the best value for lambda, we get is 0.5

Variables Used

The variables we used during our analysis include purchase, product category 2 and product category 3, age, gender, city and occupation variables.



The graph below shows the importance of each of the variables. Based on the level of importance they were selected or not selected from the prediction model. The graph above further explains the degree of importance of each variable. Age group 51-55 variable is the most significant, followed by the variable City_CategoryC comparatively less significant than other variables. On the contrary, the variables at the bottom are least relevant such as a stay in current city years and occupation. Evaluating variables that are more relevant than others is crucial to predicting purchases based on variables that drive this. This will give insights to retailers which consumer group they should target to maximize Black Friday purchases for example.

The graph above explains the degree of importance of each variable. The chart above shows that the variable in the Age group 51-55 is the most important, followed by the variable

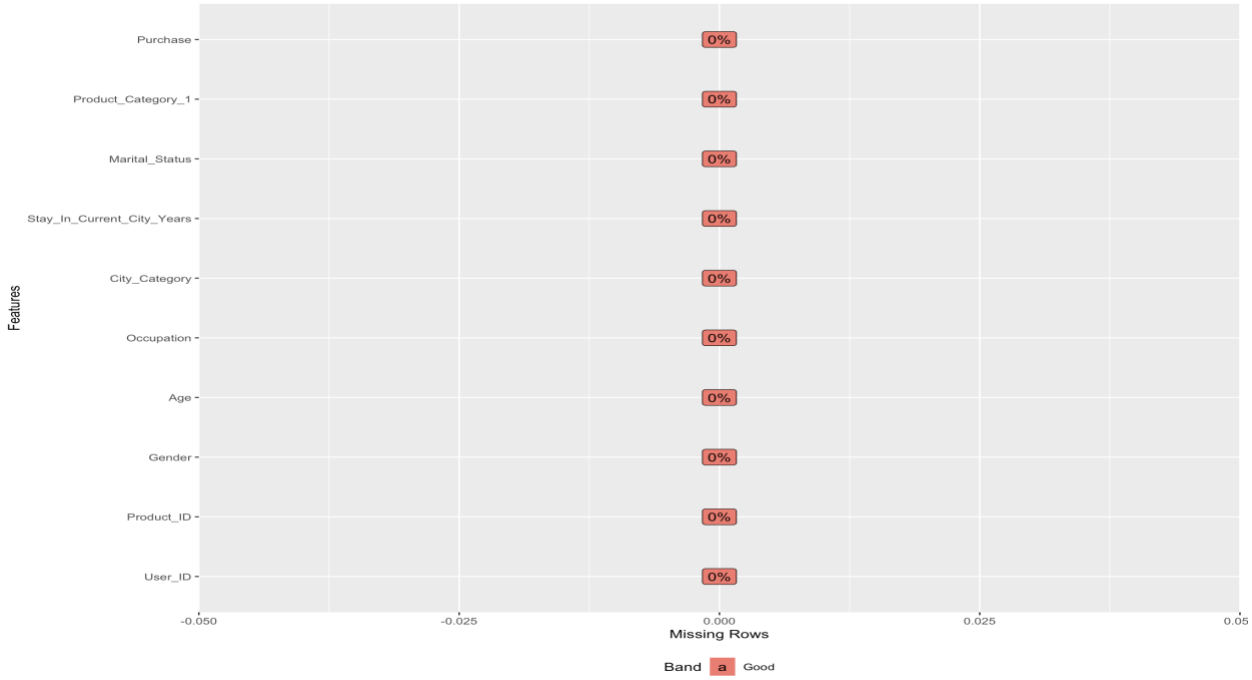
City_CategoryC comparatively less important than other variables. On the contrary, the variables at the bottom are least important such as a stay in current city years and occupation. It is important to assess which variables are more relevant than others as this gives insights to what retailers need to pay attention to in order to maximize Black Friday sales.

Optimization Selection

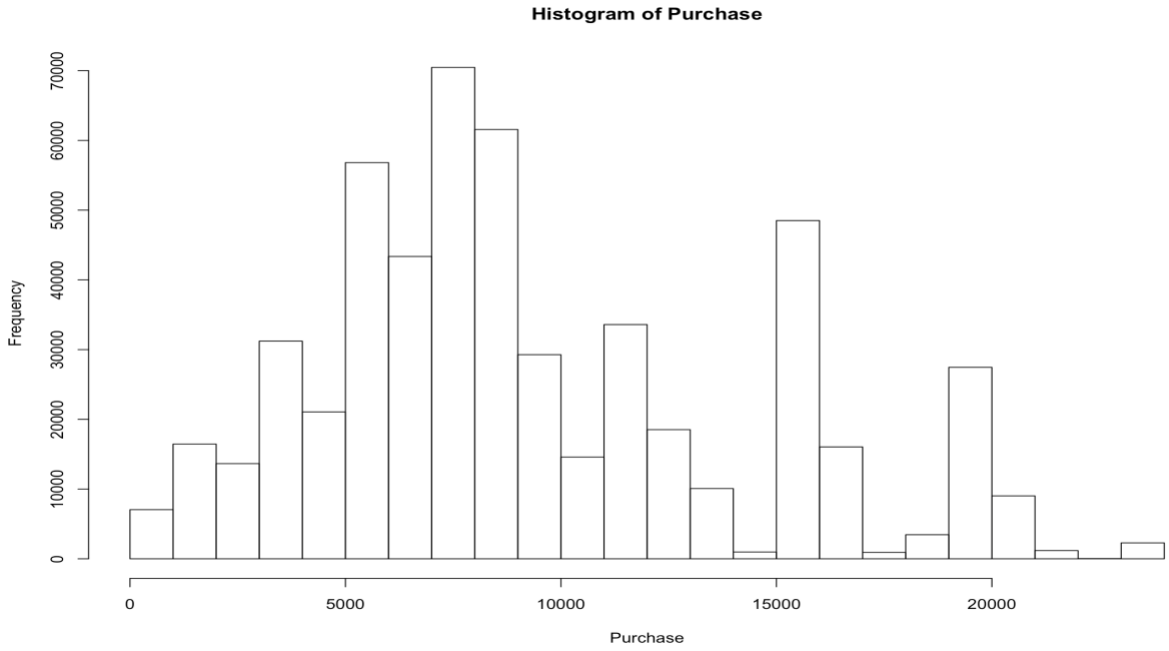
We performed the lasso for regularization to improve the accuracy of the prediction. Furthermore, in order to optimize the model, we decided to use decision trees and the principal component analysis (PCA). The PCA is used to reduce the dimensionality of a large dataset consisting of several variables correlated with each other while keeping the most important variation in the small dataset. In order to obtain better performance for our decision tree, we use the PCA first and then performed a decision tree. The decision tree is used for decision making as it allows to see to further analyze the consequences of a decision.

Findings

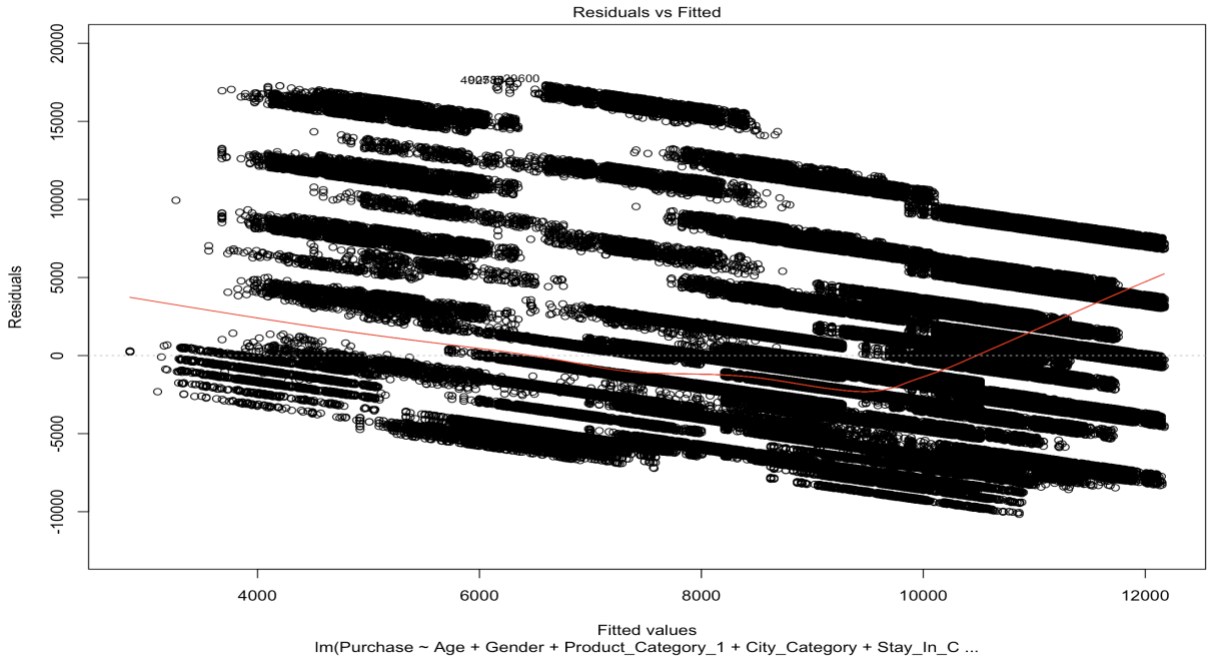
The variable used for this analysis were purchase, Product_Category_1, Marital_Status, Stay_in_Current_City_Years, City_Category, Occupation, Age, Gender, Product_ ID, and User_ID.



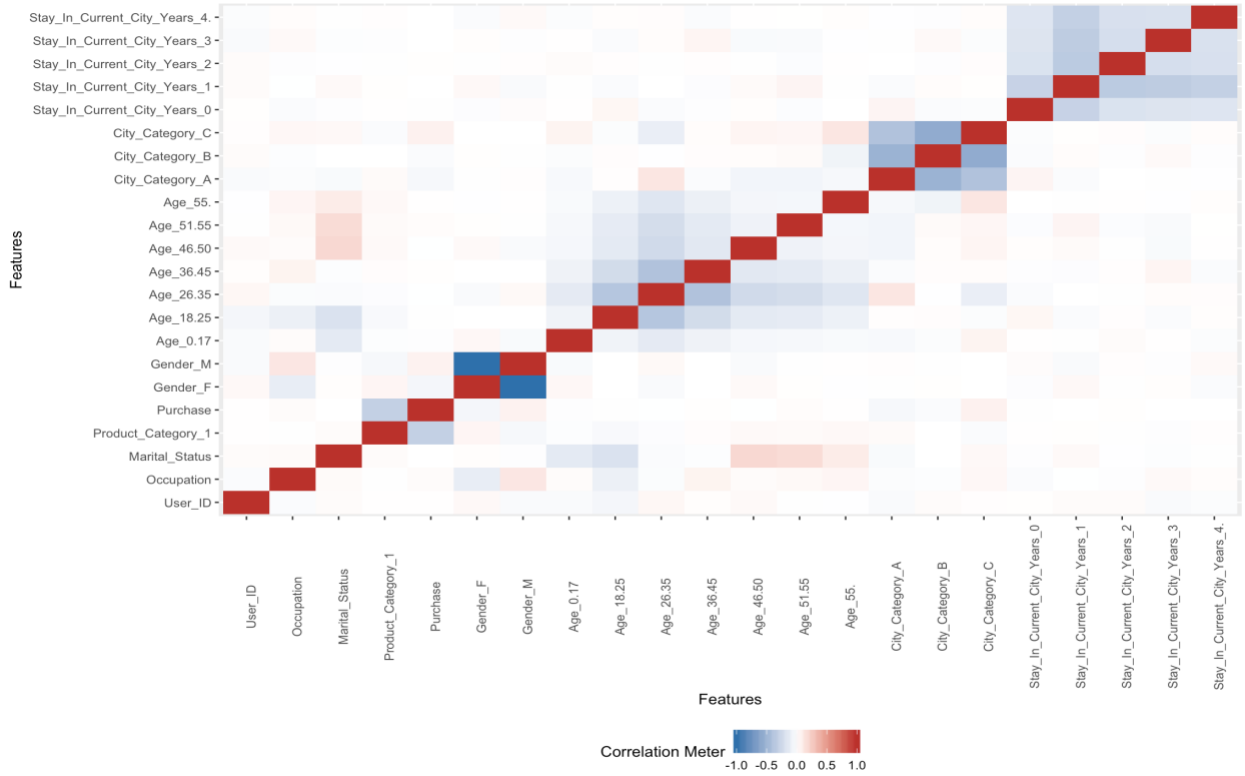
The graph above shows the plot of the missing values. Before drawing conclusions on the dataset, missing values must be checked to prevent the results from being impacted. As there are no missing values in the dataset, we can proceed with our analysis.



The histogram above shows the distribution of purchases made by customers. In the graph above it can be seen that the most customer total purchase was equal to $8,000 and $9,000 following by purchases for a total of close to $5,000. It can be concluded that the majority of customers make purchases of more than 4,000 on black Friday. With this in mind, the retailer can maximize sales on Black Friday by selecting specific products that will be best sellers to increase the frequency of purchases made on this day.



The graph above shows the linear regression performed on the dataset. As there is more than one variable in the dataset, the graph shows a multilinear regression model showing the relationship that exists within the several independent variables and the dataset. A change in the dependent variable, which is purchases can be assessed using this graph to observe how much the dependent variable (purchases) will change when a change in the independent variables is made. Additionally, this will help management which is a retailer in this example, to identify how much purchases are going to increase or decrease for every point increase in the following independent variables: Product_Category_1, Marital_Status, Stay_in_Current_City_Years, City_Category, Occupation, Age, Gender, Product_ID, and User_ID.



The correlation graph above was created to find dependencies between variables in the dataset. The correlation plot assesses which variables are highly correlated and which variables are not, this will help guide which variables should be included or excluded in the models.

Optimization model versus Baseline Model

The optimization model helped to increase a higher accuracy than the one obtained in the original model. The model technique used helped us to increase accuracy. The optimization model did not only provide clear results but also helped to improve the baseline model.

Conclusion

The Black Friday dataset obtained from Kaggle provided the group with a dataset that enabled us to practice data cleaning methods as well as data analysis techniques using the lasso regression model in R. The data cleaning methods we used include removing NA which implied that we removed the two columns namely Product Category 2 and Product Category 3 as discussed in prior paragraphs. This dataset was a good one to look at as we assessed the most effective model to use due to the nature of the dataset. We found that the lasso regression model will be a more effective model to use at it addresses the issue of overfitting in the dataset.

Thus, we can conclude, the occupation and the location of stay don't influence much during a black Friday sale, as it is just one day, and people can commute. The highest purchases are made by the working population which falls under the age group of 26-35 and 36-45. Hence, while considering factors like manufacturing, sales, brands, and utilities one must consider this age group as a priority and make decisions in their favor.

References

Dagdoug, M. (2018, July 25). Black Friday. Retrieved April 20, 2019, from https://www.kaggle.com/mehdidag/black-friday

Filip, S. (2015, March 12). How to make a histogram with ggplot2. Retrieved April 20, 2019, from https://www.r-bloggers.com/how-to-make-a-histogram-with-ggplot2/

Selva, P. Ridge Regression. Retrieved April 20, 2019, from http://r-statistics.co/Ridge-Regression-With-R.html

Josh, E. SQL in R. Retrieved April 20, 2019, form http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/sql.html