**ALY 6015 Black Friday Project Report**

**By**

**Parveen A**

**Tuhena Sen**

**Northeastern University**

**Course: ALY6015 Intermediate Analysis**

**Professor: Dr. Roy Wada**

**Jun 28th,2020**

**Introduction**

In this assignment we will be analyzing holiday shopping “ Black Friday”. Black Friday is a special day for shopping as stores keep interesting discounts to clear their stuff. Here we are going to analyze the customer behaviour based on current black friday sales data (from Kaggle) and then predict the amount of sales that will happen in next black friday sale. In other words, we will analyze the possible outcomes based on the purchase that customers have made.

Overview of data:

The dataset has been taken from (Kaggle,2018). The observation on the Black Friday data set consists of 537577 rows and 12 predictor variables listed in the summary statistics.

This report will explore the likely amount of purchase by customers during Black Friday Sale. The variable of interest in this report is the purchase amount between the period 2005 and 2013. This data also contains demographic information on customers that shopped at ABC Private Limited for the previous years. This demographic information includes age, gender, occupation, city category, stay in current city and marital status. It also contains information on the products purchased, such as their ID and different product category information. Using this information, we have built various machine learning models that will predict the purchase amount based on the customer’s demographics and the categories of the products. In the end, ABC Private Unlimited will have information that they can use to offer customers products that will be appropriate for them.

**Summary Statistics of the Black Friday Data Set**

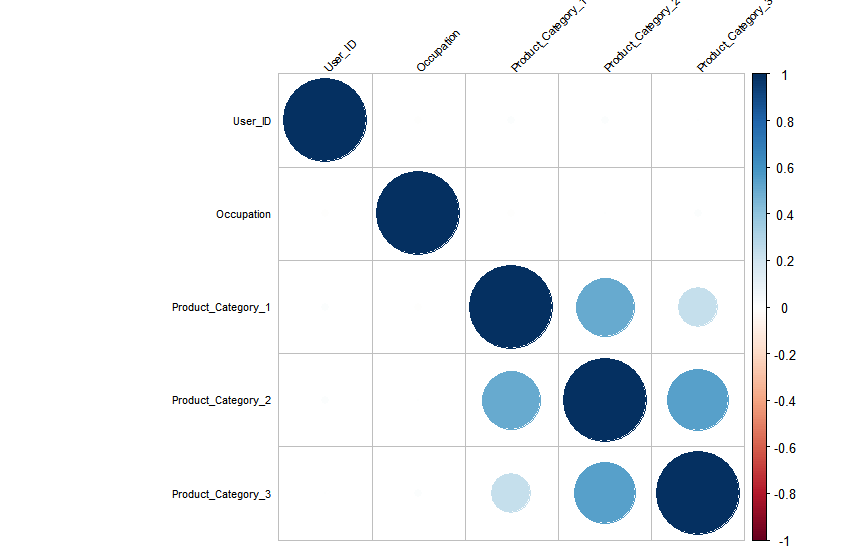
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **User\_ID** | **Product\_ID** | **Gender** | **Age** | **Occupation** |
| 1001680: 1025 | P00265242: 1858 | F:132197 | 0-17 : 14707 | 4 : 70862 |
| 1004277: 978 | P00110742: 1591 | M:405380 | 18-25 : 97634 | 0 : 68120 |
| 1001941: 898 | P00025442: 1586 |  | 26-35: 214690 | 7 : 57806 |
| 1001181: 861 | P00112142: 1539 |  | 36-45 : 107499 | 1 : 45971 |
| 1000889: 822 | P00057642: 1430 |  | 46-50 : 44526 | 17 : 39090 |
| 1003618: 766 | P00184942: 1424 |  | 51-55 : 37618 | 20 : 32910 |
| (Other):532227 | (Other) :528149 |  | 55+ : 20903 | (Other):222818 |
|  |  |  |  |  |
| **City\_Category** | **Stay\_In\_Current\_City\_Years** | **Marital\_Status** | **Product\_Category\_1** | **Product\_Category\_2** |
| A:144638 | 0: 72725 | 0 : 317817 | Min. : 1.000 | 8 : 63058 |
| B:226493 | 1:189192 | 1 : 219760 | 1st Qu.: 1.000 | 14 : 54158 |
| C:166446 | 2: 99459 |  | Median : 5.000 | 2 : 48481 |
|  | 3: 93312 |  | Mean : 5.296 | 16 : 42602 |
|  | 4: 82889 |  | 3rd Qu.: 8.000 | 15 : 37317 |
|  | NA |  | Max. :18.000 | (Other):124975 |
|  | NA |  | NA | NA's :166986 |
|  |  |  |  |  |
| **Product\_Category\_3** | **Purchase** |  |  |  |
| 16 : 32148 | Min. : 185 |  |  |  |
| 15 : 27611 | 1st Qu.: 5866 |  |  |  |
| 14 : 18121 | Median : 8062 |  |  |  |
| 17 : 16449 | Mean : 9334 |  |  |  |
| 5 : 16380 | 3rd Qu.:12073 |  |  |  |
| (Other): 53569 | Max. :23961 |  |  |  |
| NA's :373299 | NA |  |  |  |

The dimension of the Black Friday data set is 537577 rows and 12 variables listed above in the summary statistics.

**Exploratory Analysis**

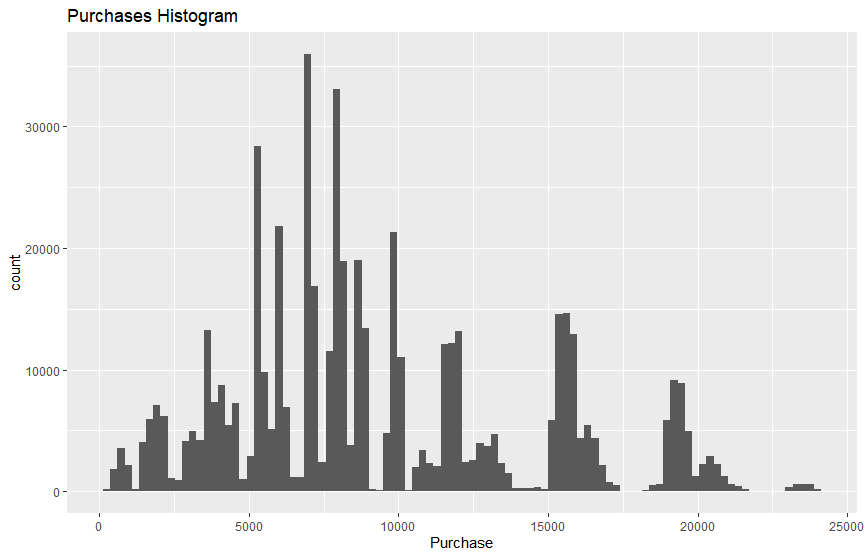
The very first step in analyzing data set is to understand the packages and load the respective libraries. Next step is to find and impute the missing values in the data set. The predictor variables Product\_Category\_2 and Product\_Category\_3 have missing values and are imputed to 0 based on the correlation matrix.

**Identifying Multicollinearity**



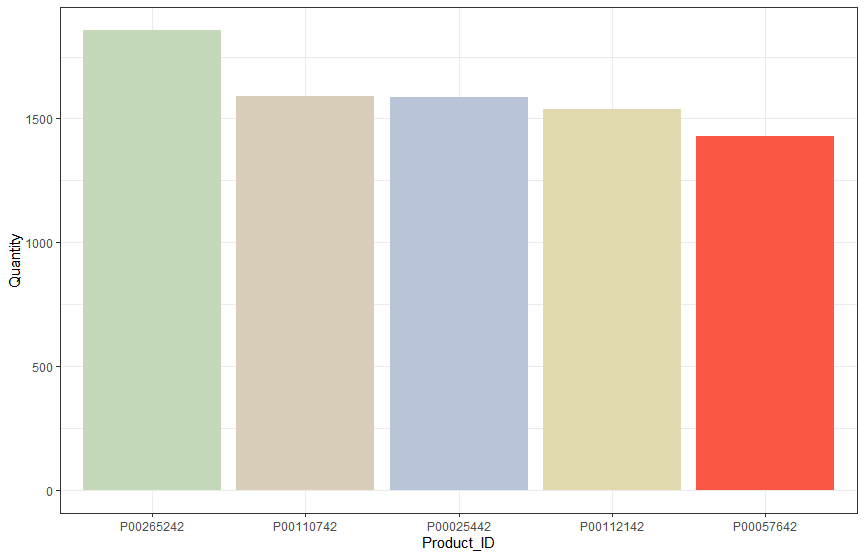
Since the correlation between Product\_Category 1,2,3 are strong, Product\_Category 2 and 3 with NA values are dropped in order to avoid multicollinearity.

**Histogram of Purchase :**



We can see that from the histogram the average purchase rate is between 5k to 10k.

**Purchase by Top 5 Product IDs:**



We can see that the highest purchase quantity of products belongs to

Product\_ID’s: ‘P00265242’ followed by 'P00110742','P00025442','P00112142','P00057642'.

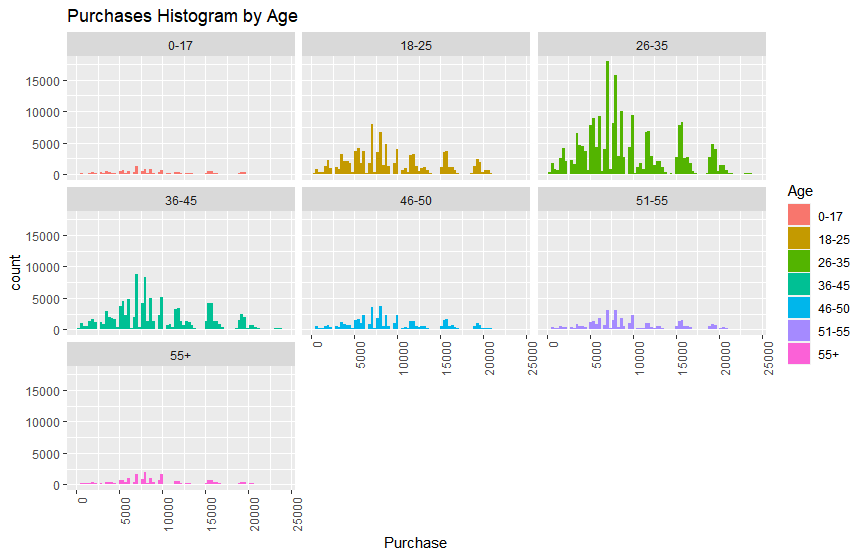
**Bar Chart of Purchase by Gender**

From the below chart its very clear that Male percentage stands highest in purchasing products.



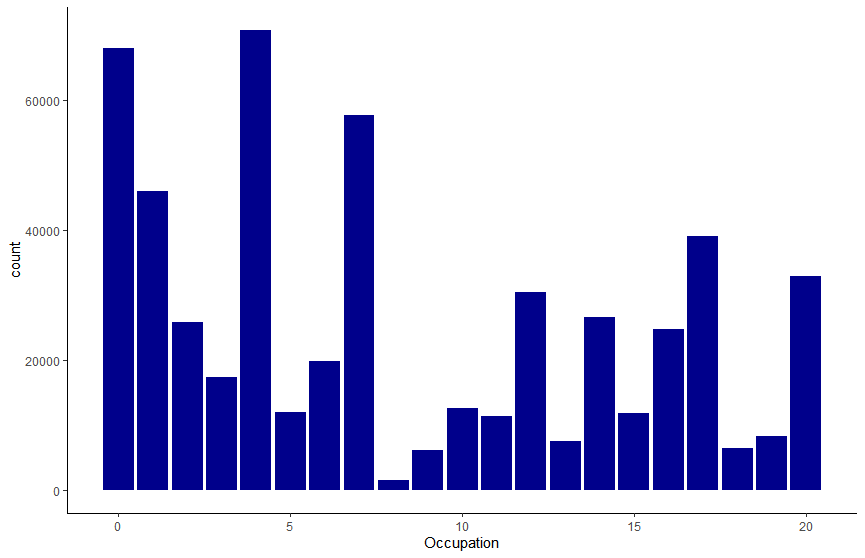
**Histogram of Purchase by Age Group:**

The histogram below depicts that highest purchase is done by the younger generations of age group 26-35 years. Lowest purchase is done by old age people as they will be less likely to purchase gadgets and other electronic items during sales.



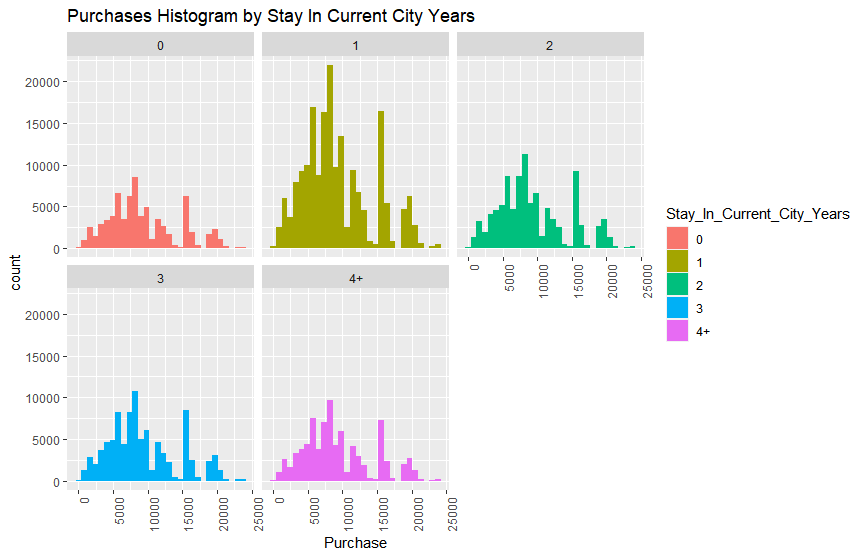
**Bar Chart of Purchase by Occupation:**

We can see that the highest purchase is done by the people belonging to Occupation category 0 ,5,7. And the lowest purchase is recorded by the occupation category 8.



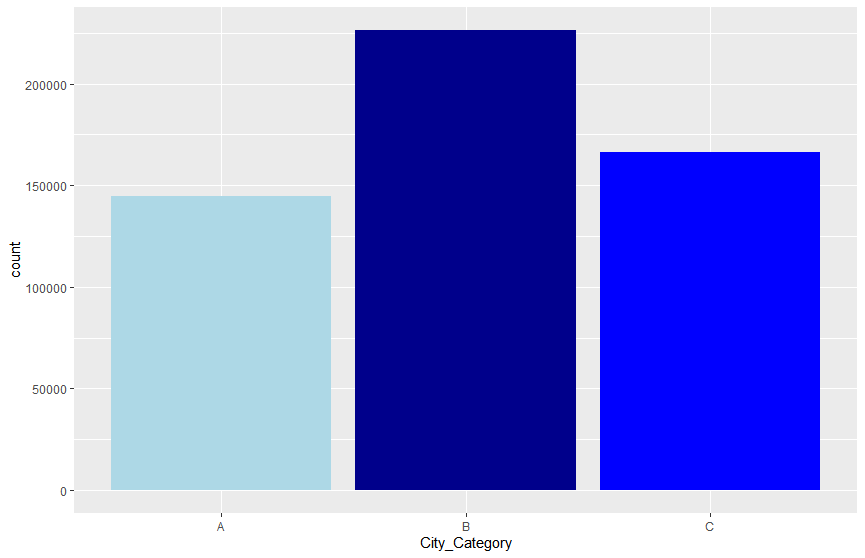
**Histogram of Purchase by Stay in current city years:**

People who are residing in the current city for a year has the highest quantity of purchase of products.



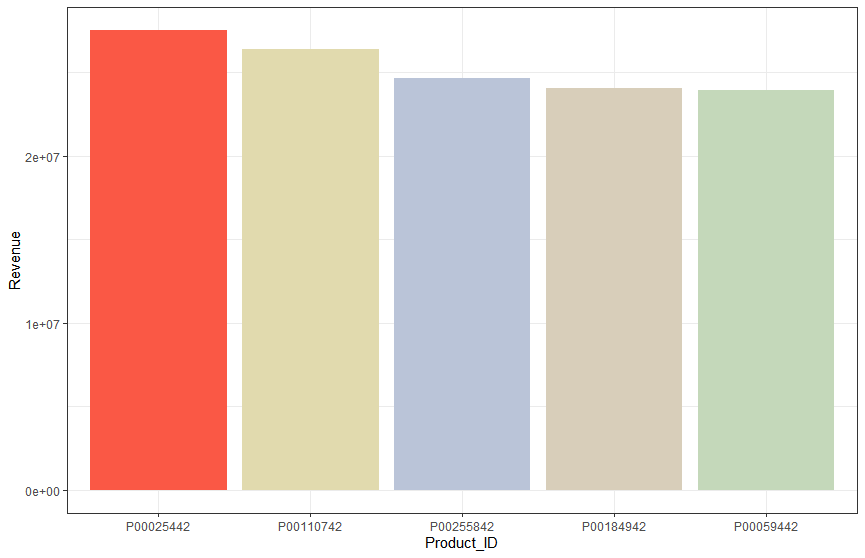
**Density Plot of Purchase by City:**

We can see that City\_Category B records highest quantity of purchased products followed by C and A.



**Revenue by Product ID:**

Since our model needs to be trained on the purchase amount of the data set we are interested to know which Product\_ID yields the highest revenue during the Black Friday sales. Below bar chart lists the top 5 Product\_IDs which account for highest revenue .



**Machine Learning Models**

Once the near zero variance variables are removed, we create a random sample from the data. This sample is created since the data set is huge. Its size may cause trouble while building ML models. The sample size 10% of the data. Once the sample is selected, the next step in building the model can be performed.

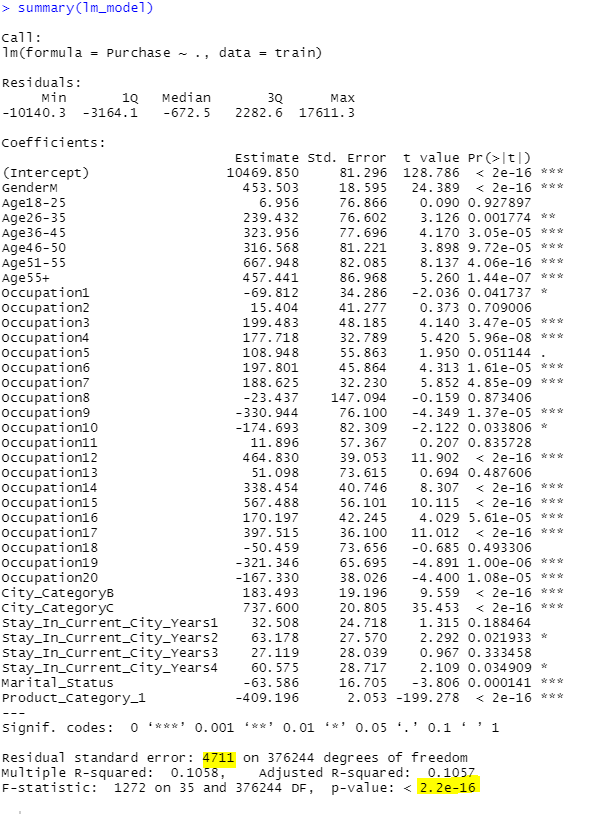
Cross Validation of the Sample data:

Training data 🡪 Includes 80% of the sampled data

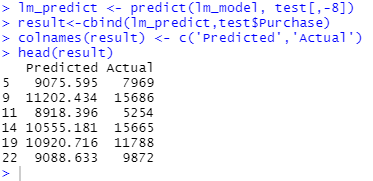
Test data 🡪 Remaining 20% of the sampled data

**Building Linear Regression Model:**

When we apply LM model on the training data we get the below coefficients for the predictor variables. We have used the root mean squared error (rmse) metric for evaluating the model when compared to the Mean Absolute Error, since RMSE punishes large errors.



When we apply the training model on the test data we get the below list of Actual and Predicted purchase values.



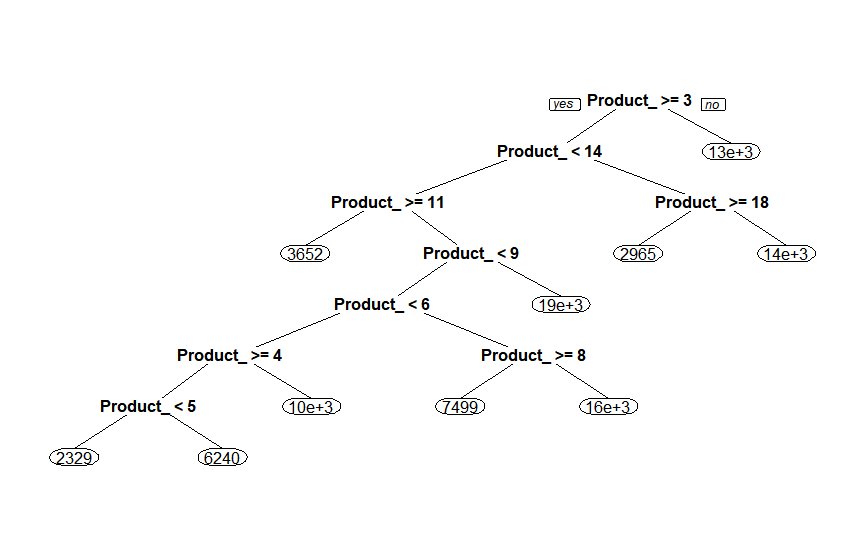
**RMSE of LM Model :**

> sqrt(mean((result$Actual - result$Predicted)^2))

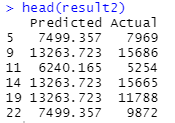
[1] **4695.367**

In order to further improve our model, we will be performing decision tree to predict the purchase values more accurately.

**Building Decision Tree Model**



When we apply decision tree on the test data, we got the below list of predicted purchase amounts against the actual ones.



**RMSE of Decision Tree Model:**

> sqrt(mean((result2$Actual - result2$Predicted)^2))

[1] **3103.864**

**Regression Analysis(Logistic Regression)**

In the next step we will set the data for regression analysis. The dataset here is a sample of the transactions made in retail shops. Particularly, here the issue is a regression problem where we are trying to predict the dependent variable (purchase amount) with help of information of remaining variables.

Initially we have to define the 2 variables we are analyzing. Hence, we have taken

Y as purchase variable and remaining all as X variables.

Now in order to use these variables for the logistic regression we need to change the formats as a factor for all the variables except purchase variables which we have considered as Y. We change the Y variable as categorical that is 0 or 1. 0 means the purchase amount is less hence purchase for that particular product will not be made next year whereas 1 means the purchase amount is more hence purchase for that particular product will be made next year. Now we will take the mean of purchase amount and will set it to 0 for less than the mean value and 1 for more than the mean value.

>bf <- mydata

>str(bf)# original data

>bf$User\_ID <- as.factor(bf$User\_ID)

>bf$Occupation <- as.factor(bf$Occupation)

>bf$Product\_Category\_1 <- as.factor(bf$Product\_Category\_1)

>bf$Product\_Category\_2 <- as.factor(bf$Product\_Category\_2)

>bf$Product\_Category\_3 <- as.factor(bf$Product\_Category\_3)

>bf$Purchase <- as.numeric(bf$Purchase)

>bf$Marital\_Status <- ifelse(bf$Marital\_Status==1,"married", "single")

>bf$Marital\_Status <- as.factor(bf$Marital\_Status)

After changing the format as factor:

Now the data is set for logistic regression. Next, we will take the mean and will set purchase categorical value.

>mean<-mean(bf$Purchase)

>mean

[1] 9333.86

>bf$Purchase <- ifelse(bf$Purchase >= 9333.86, 1, 0)

>bf$Purchase

As we can see that now the purchase value has been changed to 1 or 0 according to the mean.

Now we will separate the data into training and testing sets for running logistic regression.

For this we have separated the dataset as 75% train data and 25 % test data by seed function.

>set.seed(123)

>row.number <- sample(x=1:nrow(bf), size=0.75\*nrow(bf))

>train = bf[row.number,]

>test = bf[-row.number,]

Now we will run the logistic regression:

Initially we will take all the other variables

>logistic\_model <- glm(Purchase ~ Age + Product\_Category\_1 + City\_Category + Marital\_Status + Stay\_In\_Current\_City\_Years + Gender, family = binomial, data=train)

> summary(logistic\_model)

Result:

Now we can fetch some interesting results here we can see that in p value for many of the variables has meaning . It means most of the p values are less than 0.05,

hence has significance which indicates it is a good model.

Like for Age all below variables have meaningful p values means that most buyers lie in this age range.

Estimate Std. Error z value Pr(>|z|)

Age26-35 0.064307 0.031814 2.021 0.043244 \*

Age36-45 0.158684 0.032709 4.851 1.23e-06 \*\*\*

Age46-50 0.127831 0.035765 3.574 0.000351 \*\*\*

Age51-55 0.361430 0.036662 9.858 < 2e-16 \*\*\*

Age55+ 0.300316 0.039761 7.553 4.25e-14 \*\*\*

From the product category we can see many have significant p values means this is a good model and those product categories have more sales.

Now we will predict for next sales.

>pred <- predict(logistic\_model, newdata = test, type = "response")

After predicting the model, we will now check for errors and accuracy.

For accuracy we will plot the ROC curve. More the ROC curve towards 1 more the model is accurate.

>plotROC(test$Purchase, pred)

The graph shows that this model is 92.3 percent accurate.

>misClassError(test$Purchase, pred, threshold = 0.5)

[1] 0.1363

>confusionMatrix(test$Purchase, pred, threshold = 0.5)

0 1

0 69538 7529

1 10785 46543

To check how much false positive and false negative in this model, we carried out the function “ confusion matrix” that shows truthfulness and falseness. We found 7529 false positive and 10785 false negative.

The sensitivity and specificity will tell us the percentage of true positive and false negative. Here we can see that 86 percent is true positive and 86.5% is false negative.

> sensitivity(test$Purchase, pred, threshold = 0.5)# true positive

[1] 0.8607597

> specificity(test$Purchase, pred, threshold = 0.5) #False negative

[1] 0.8657296

**Conclusion:**

As we can observe from the above two model’s decision tree model is computationally intensive when compared to the LM model based on the RMSE error. Feature engineering and model architecture can further improve the performance of our model, but lots of experimentation is needed to figure out the best model.

**Bibliography**

1. Dagdoug, M. (2018, July 25). Black Friday. Retrieved from<https://www.kaggle.com/llopesolivei/blackfriday>
2. Saishurti(2018,March 15) . Logistic Regression. Retrieved from:<https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>
3. Jason (2016, March 15). Linear Regression . Retrieved from:<https://machinelearningmastery.com/linear-regression-for-machine-learning/>