PREDICTING DELINQUENT CUSTOMER

MBA6693 – Business Analytics

Group Project

Kilfoil, David

Kumbhare, GS

Sharma, Shubh

OBJECTIVE & APPROACH

- Problem Statement:
  - **CredX** is a leading credit card provider that gets thousands of credit card applications annually. However, CredX has suffered an increase in credit losses due to an increase in the number of defaulters.
  - The company management has chosen to mitigate the credit risk and minimize the losses by acquiring new customers that should not default.

- Objective:
  - The aim of this project is to help CredX identify the right customers using predictive models. We are supplied the data of past CredX customers, and we need to determine the factors affecting credit risk plus create strategies to mitigate the credit risk and assess the financial benefit of our chosen model.

APPROACH

- The problem at hand is a binary supervised classification problem.
- We will build three models for the datasets provided:
  - **Demographic Model**
    - In order to understand the role played by the demographic data of a customer, we will build a classification model solely with the demographic data set. We will use the **Support Vector Machines** technique to build this model.
  - **Combined Data Set Models**: We will create the following two models with the combined demographic and credit data:
    - A model using **Logistic Regression**. The predictors will be selected on the basis of their **WOE** and **IV** values.
    - A second combined data model using **Random Forest**. The dimensions will be reduced with the help of **Principal Component Analysis**.

STEPS

We followed the following steps as we progressed through this project:

1. **Data Description**
2. **Data Cleansing and Preparation**
3. **Exploratory Data Analysis**
4. **Data Transformation and Model Building**
5. **Model Evaluation**
6. **Cost Benefit Analysis**

DATA DESCRIPTION

- There are two data sets involved in this problem:
  - **Demographic Data**
    - This information is extracted from the data provided by the credit card applicants at the time of their application for a credit card
    - This primarily contains the customer-centric information on applicant's age, gender, income, marital status, education level, profession and number of dependents
    - This dataset also has data about whether or not a particular customer has previously defaulted. This data is identified by the column name 'Peformance_Tag'. A value of 0 means non-default and 1 means default.
  - **Credit Data**
    - This data is provided by a credit bureau and provides monetary/credit details of the customer.
    - Some of the variables in this data are: Number of times 30 Days Past Due or worse in last 3/6/12 months, Outstanding Balance, and Average Credit Card Utilization.

DATA DESCRIPTION

- **Nature of Data**
  - The demographic data consists of 71295 records with 12 variables.
  - The credit data also contains 71295 records, but with 19 variables.
  - The primary key in both datasets is Application ID and this will act as the foreign key for the other data set.
  - As mentioned, Perfomance_Tag is the dependent/target variable and denotes whether or not a customer has defaulted.
  - **0 stands for non-default and 1 stands for default.**

DATA CLEANSING & PREPARATION

Data Quality

Some variables in the two data sets have NA or missing values. There are also some other issues with the data. The following tables illustrate this, with the left and right tables for demographic and credit data respectively.

Predictor	No. of NA values	Other flaws in Data
Application ID	-	3 duplicate ID's each replicated twice
Age	-	65 records with Age < 18
Income	-	81 records with Income < 0
Gender	2	
Marital Status	6	
No. of Dependents	3	
Education	119	
Profession	14	
Type of Residence	8	
Performance Tag	1425	

Predictor	No. of NA values	Other flaws in Data
Application ID		3 duplicate ID's each replicated twice
Average Credit Card utilization in last 12 months	1058	
No. of trades opened in last 6 months	1	
Presence of Open Home Loan	272	
Outstanding Balance	272	

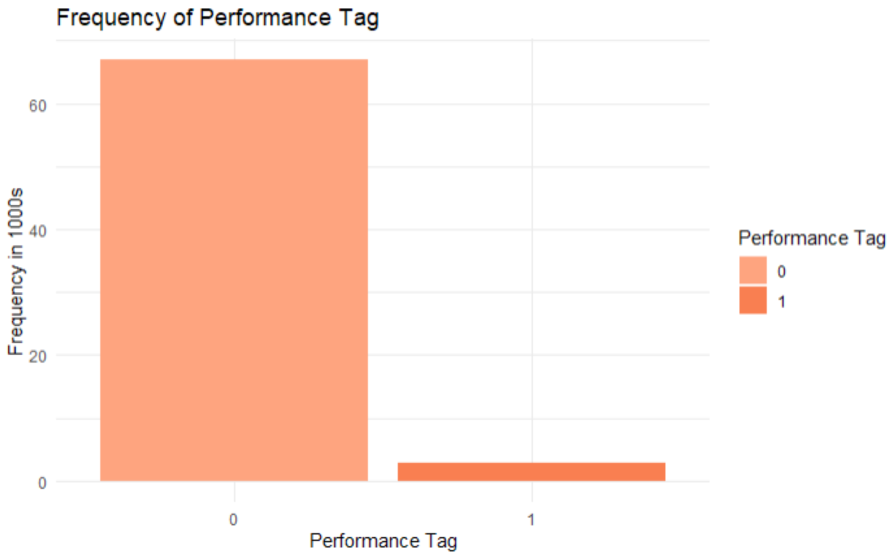
DATA CLEANSING & PREPARATION

- In order to mitigate the errors and flaws in the data as described, we have chosen the following methods:
  - We have removed the instances of the data with duplicate Application IDs. We cannot really identify which records actually correspond to a particular Application ID. In order to avoid this ambiguity, we have eliminated the rows with Application IDs that are duplicated.
  - We have also removed the rows that have missing values for the aforementioned predictors/variables.

EXPLORATORY DATA ANALYSIS

Performance Tag

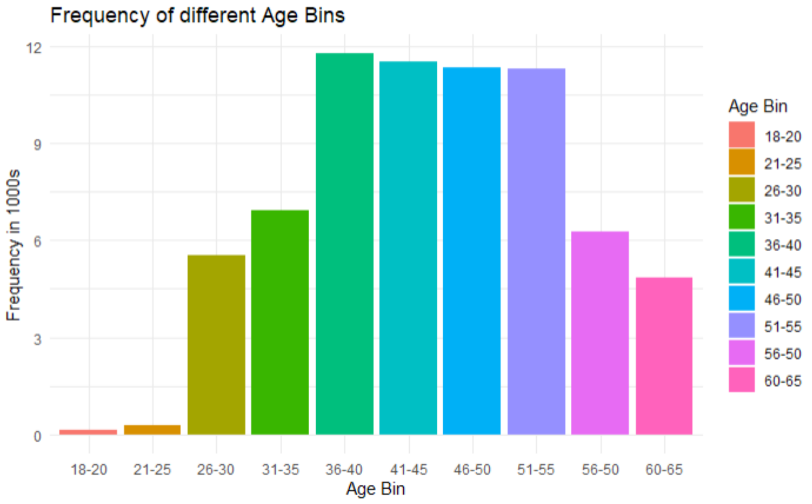
- This plot shows the number of rows with the Performance Tag value 0 and 1.
- There are 95.78 percent non-defaulters, and the remaining 4.22 percent are defaulters.
- This shows that the data set is highly skewed and imbalanced

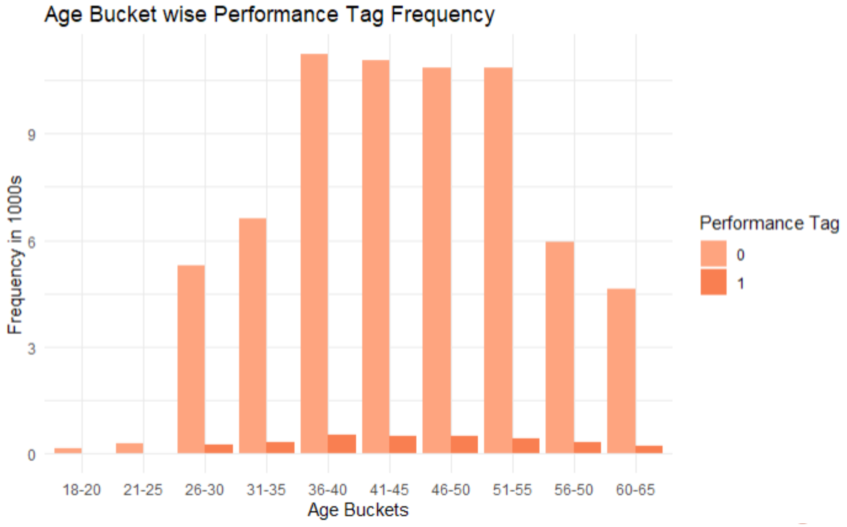


EXPLORATORY DATA ANALYSIS

Age

- The first plot shows the distribution of age of customers in various age ranges. We can see that most customers are in the age range of 36-40, and the fewest in the range 18-20.
- The second plot shows that the age range 36-40 has the most defaulters.

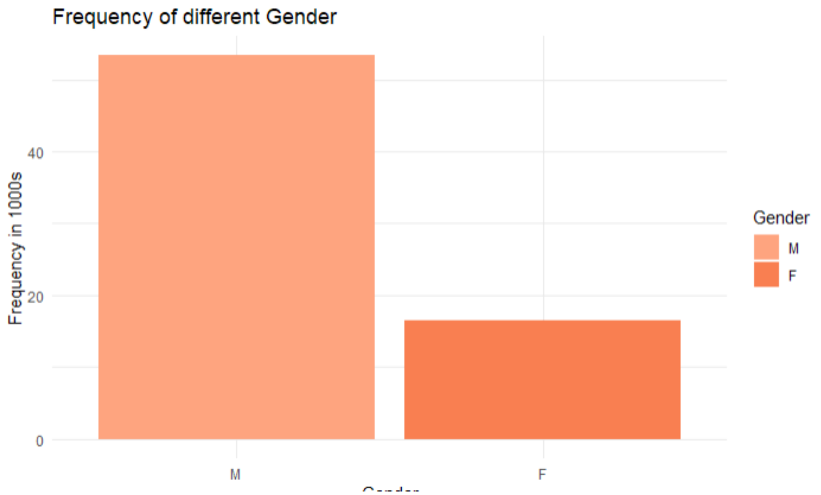

Frequency of different Age Bins

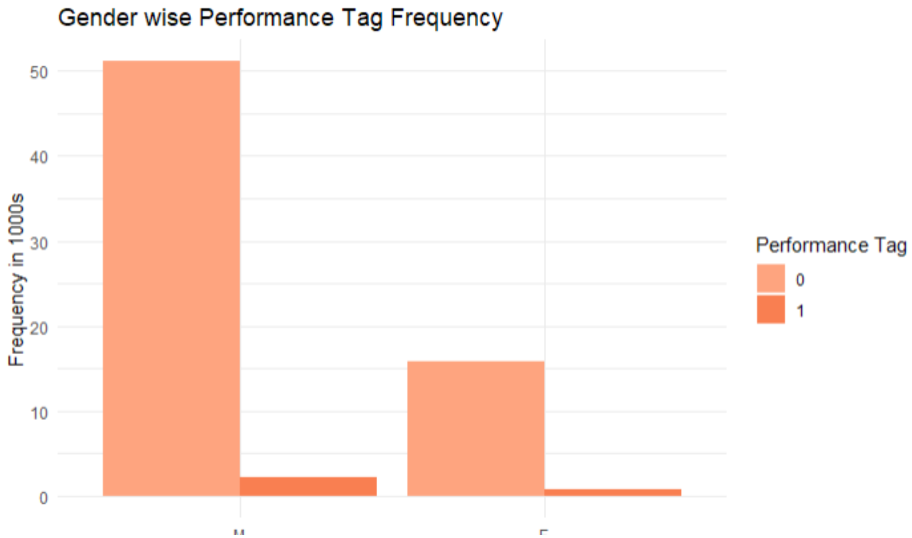

Age Bucket wise Performance Tag Frequency

EXPLORATORY DATA ANALYSIS

Gender

- The first plot shows the distribution of the two genders recorded in our data. We see that most customers are male. This is could be an implication of the fact that women are often under-represented in the employment sector.
- The second plot shows that most defaulters are men. However, women have a higher percentage of default.

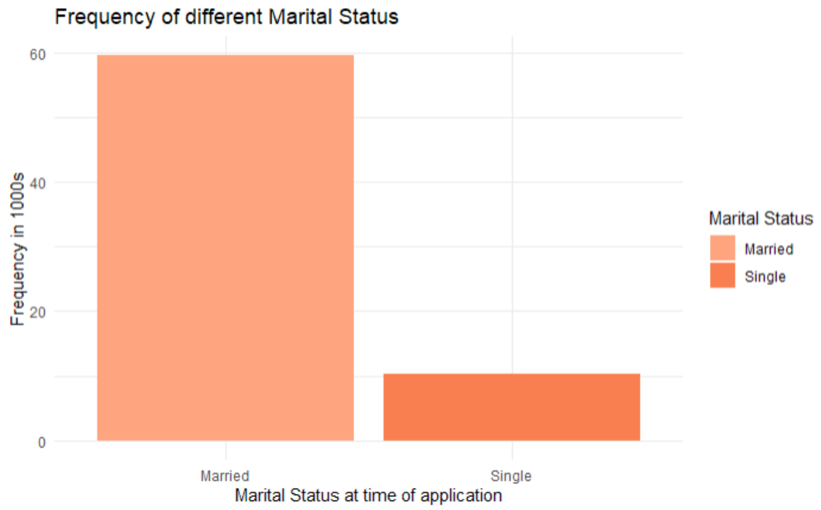

Frequency of different Gender

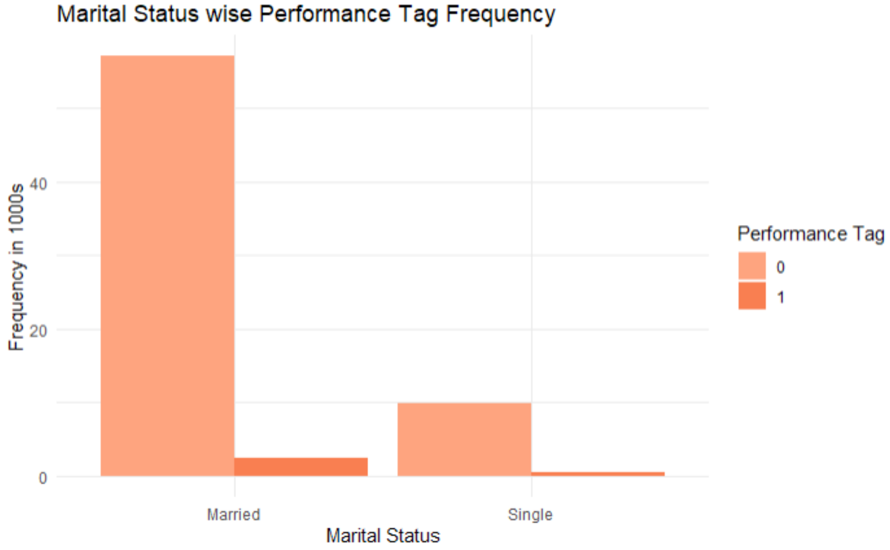

Gender wise Performance Tag Frequency

EXPLORATORY DATA ANALYSIS

Marital Status

- The first plot shows the distribution of marital status of customers in our data. We see that most customers are married. Married people often have more financial responsibilities, and this could explain a higher number of credit card applicants.
- The second plot shows that most defaulters are married customers, compared to singles.

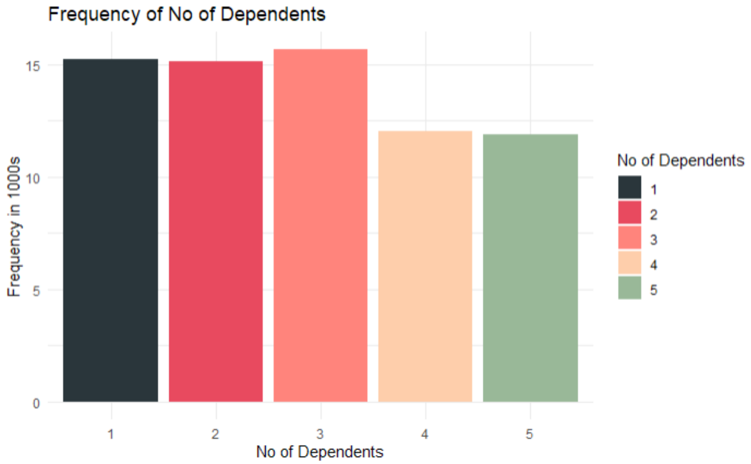


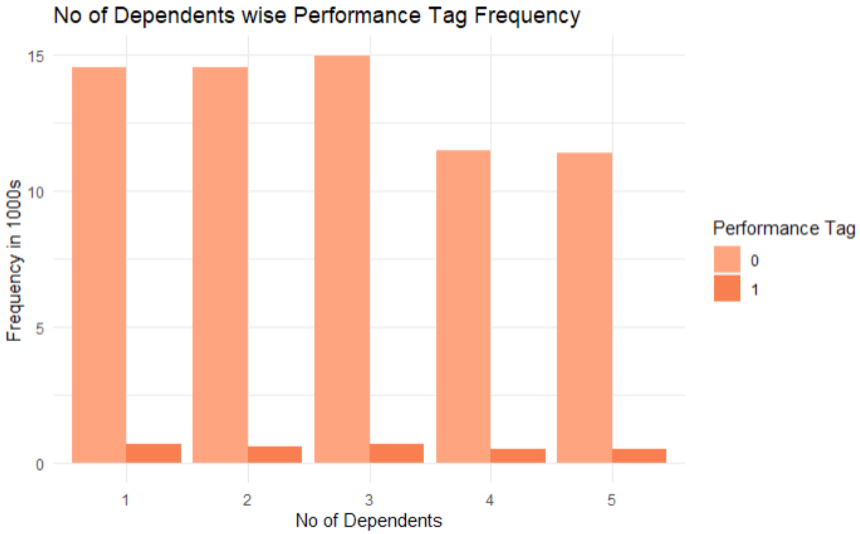


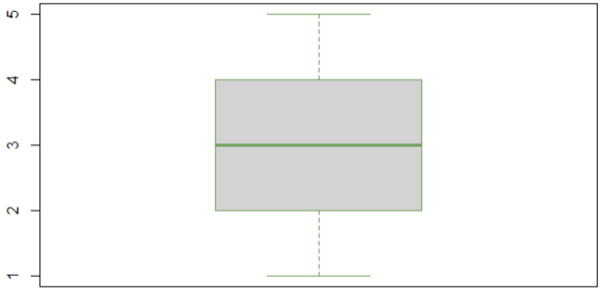
EXPLORATORY DATA ANALYSIS

No. of Dependents

- The boxplot below highlights the quartiles (2 & 4) and median (3) for Number of Dependents. There are no outliers
- The first plot on the right shows the distribution of customers on the basis of number of dependents. We can see that most customers have three dependents.
- The second plot shows that the most defaulters are customers who have three dependents.



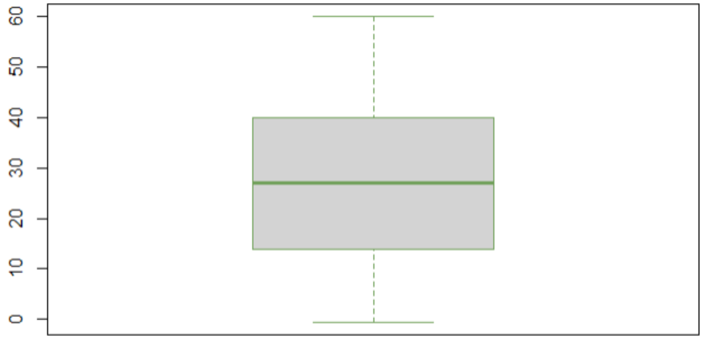


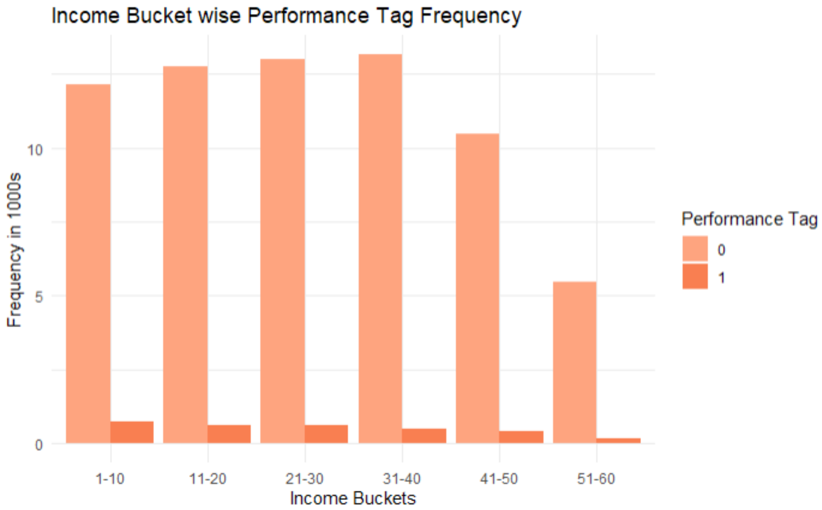


EXPLORATORY DATA ANALYSIS

Income

- The first plot shows the quartiles and median for Income. There are no outliers.
- The second plot shows that most defaulters are customers who were in the Income range of 1,000 – 10,000. However, the largest grouping of customers are in the income range 31,000 – 40,000.

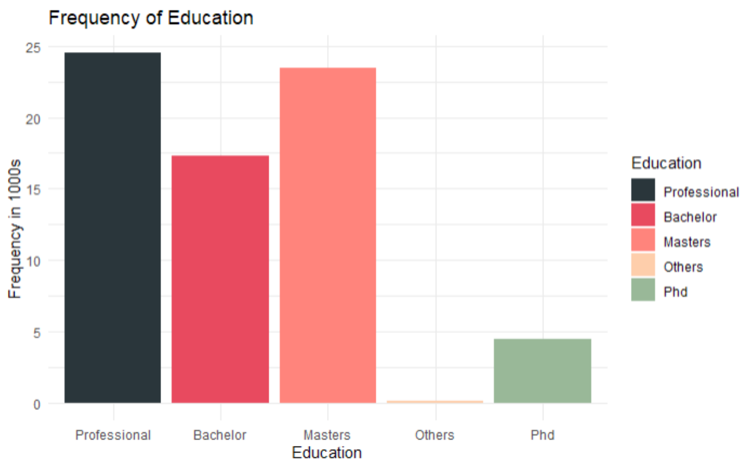


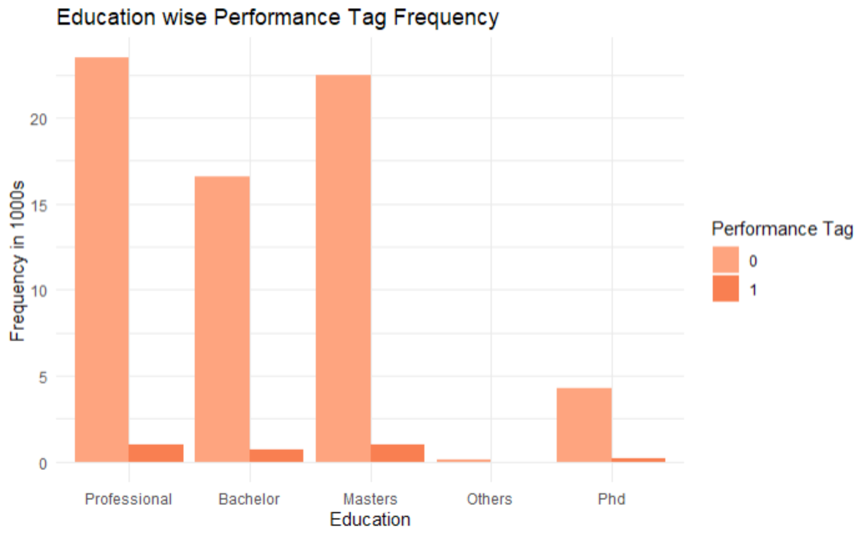


EXPLORATORY DATA ANALYSIS

Education

- The first plot shows that professional degree holders form the largest chunk of customers.
- The second plot shows that most defaulters are customers who have a professional degree. However, the highest percentage of defaulters is among customers who with a Masters degree. Graduate life crisis!

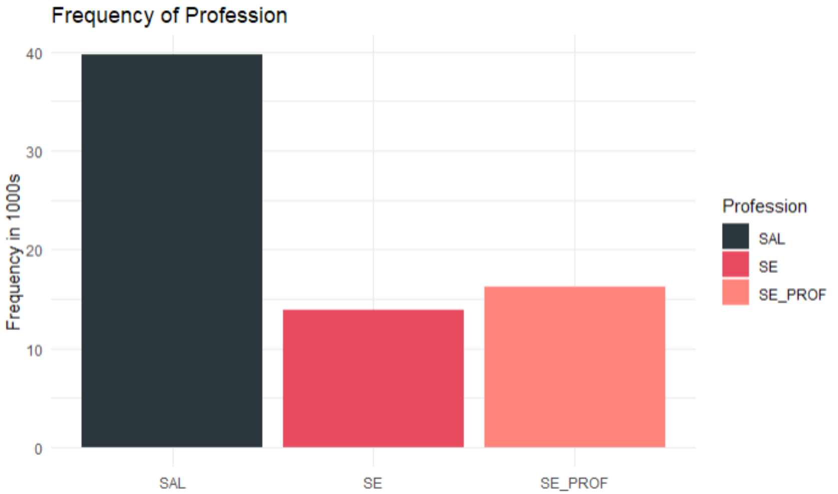


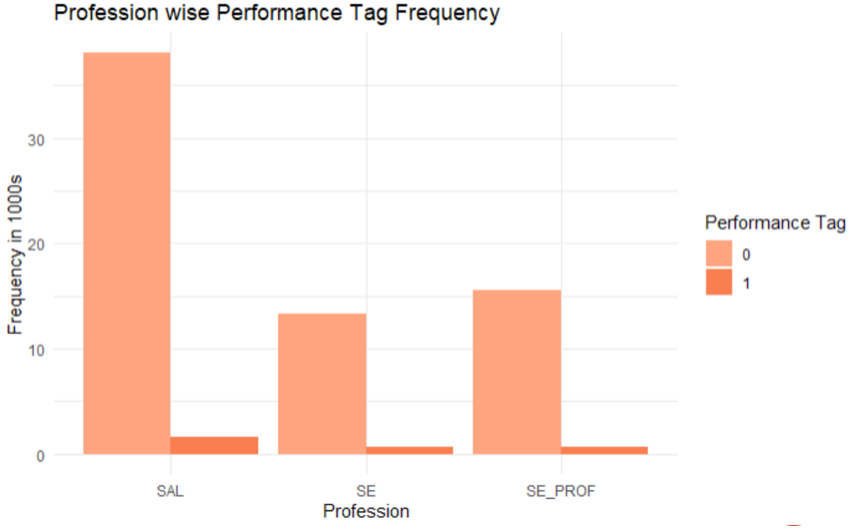


EXPLORATORY DATA ANALYSIS

Profession

- The first plot shows that salaried professionals form the largest chunk of customers among professionals.
- The second plot shows that most defaulters are customers are who are salaried professionals. But the overall default percentage is low among salaried individuals as compared to self-employed professionals.

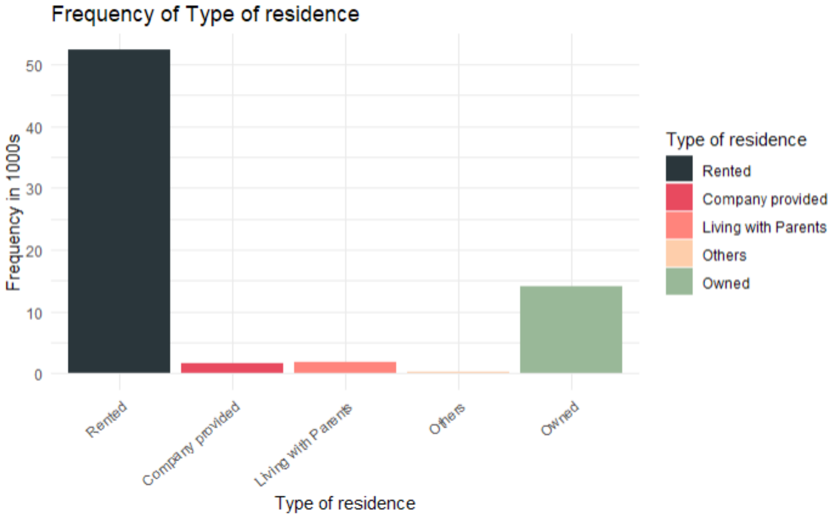




EXPLORATORY DATA ANALYSIS

Type of Residence

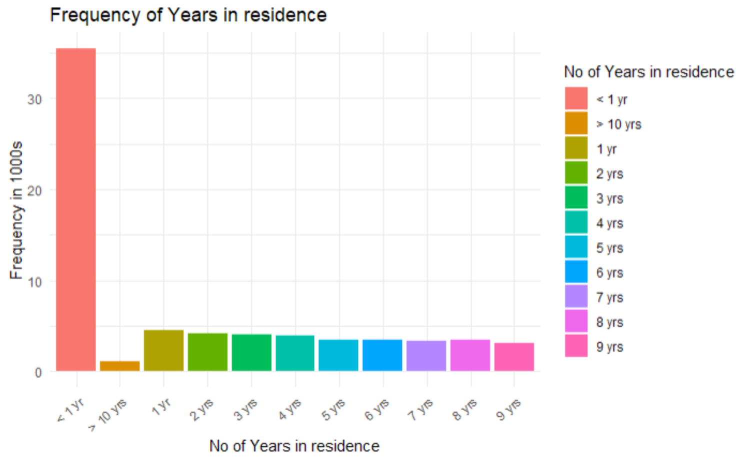
- The plot shows that customers who rented a house defaulted the most. This could be explained by the fact that people who rent a house are early in their careers or maybe do not have a stable income.

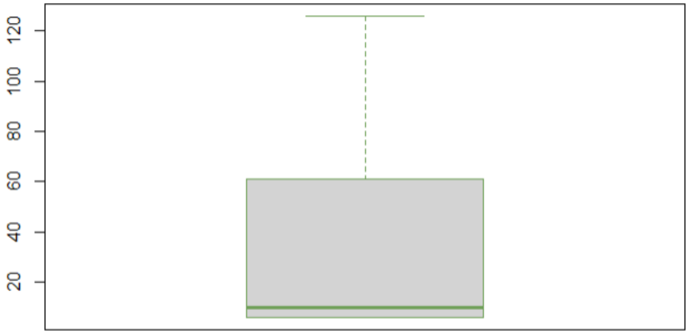


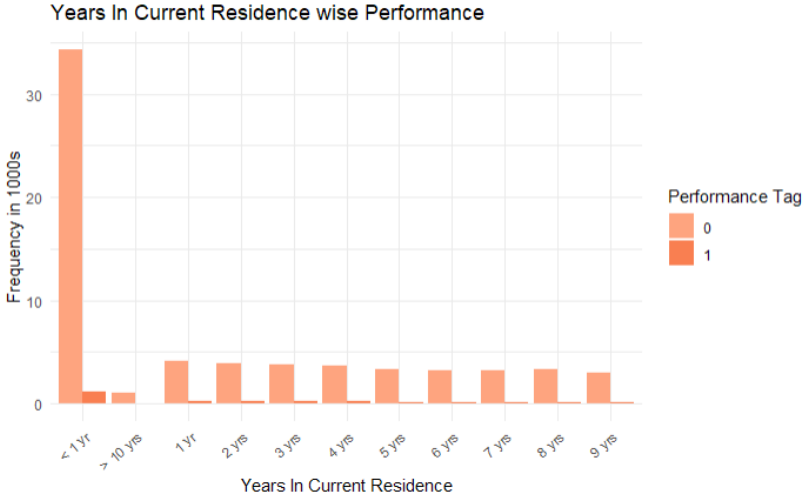
EXPLORATORY DATA ANALYSIS

Years in Current Residence

- The boxplot below highlights the quartiles and median for No. of Years in Current Residence. We can see that the data is skewed and not distributed normally.
- The first plot on the right shows that most customers had lived for less than a year at their current place of residence.
- The second plot shows that most defaulters are customers who have stayed for less than a year at their current place of residence.



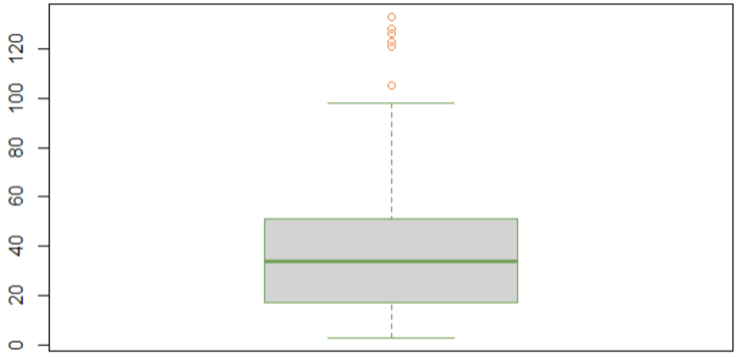


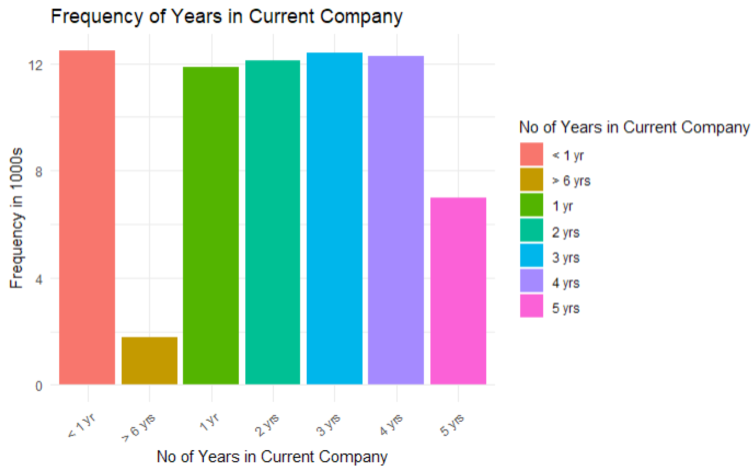


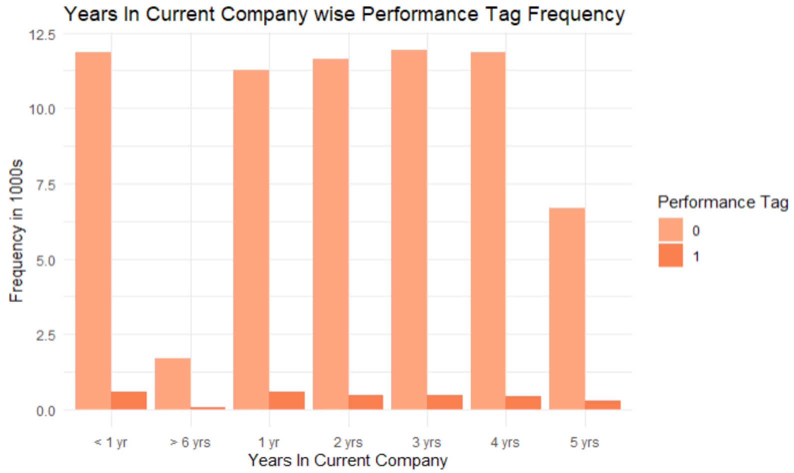
EXPLORATORY DATA ANALYSIS

Years in Current Company

- The boxplot below shows that the data is skewed and there are outliers present.
- The first plot on the right shows that people who have worked for less than one year with their current employer are more likely to apply for a credit card.
- The second plot shows that most defaulters are customers who have been working with their current employer for less than a year.



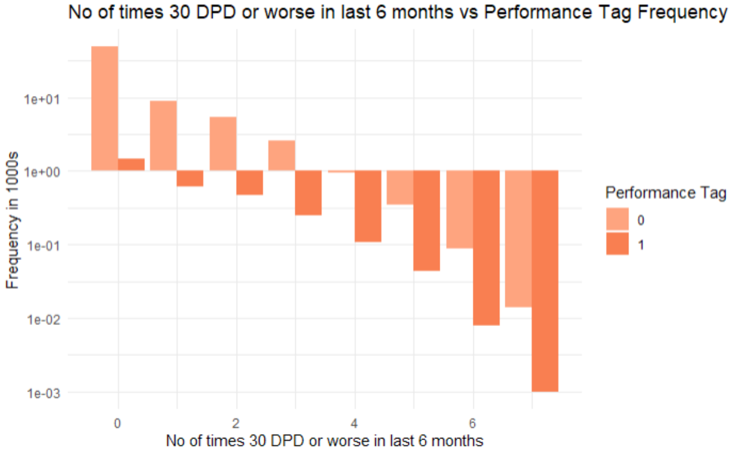


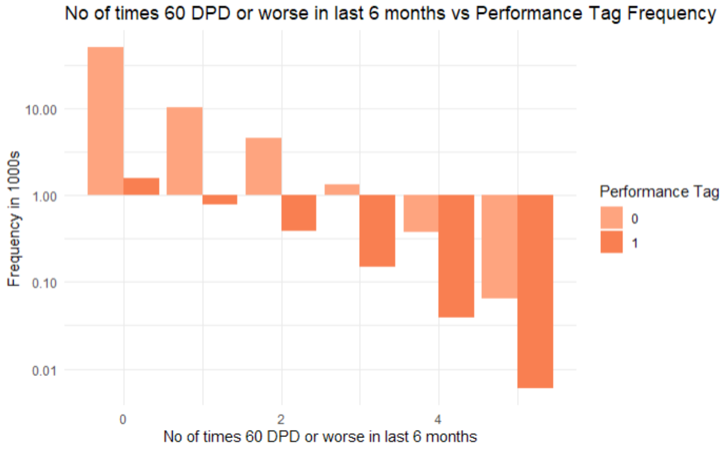


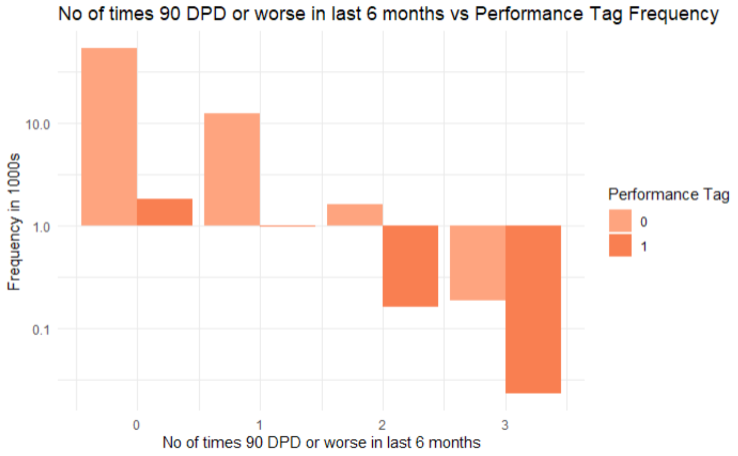
EXPLORATORY DATA ANALYSIS

30/60/90 DPD or Worse in Last 6 Months

- The plot below shows that both the highest number of customers as well as the highest number of defaulters are people who have never gone 30 days past due date in last 6 months.
- The first plot on the right shows that both the highest number of customers as well as the highest number of defaulters are people who have never gone 60 days past due date in last 6 months.
- The second plot shows that both the highest number of customers as well as the highest number of defaulters are people who have never gone 90 days past due date in last 6 months.


No of times 30 DPD or worse in last 6 months vs Performance Tag Frequency

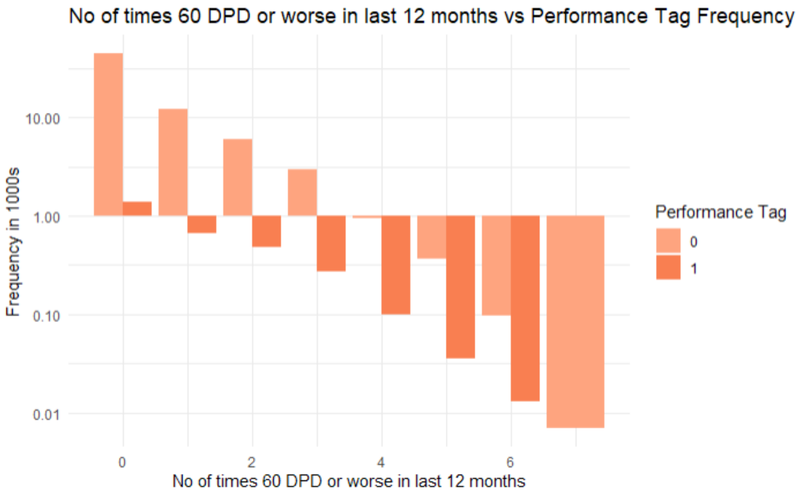

No of times 60 DPD or worse in last 6 months vs Performance Tag Frequency

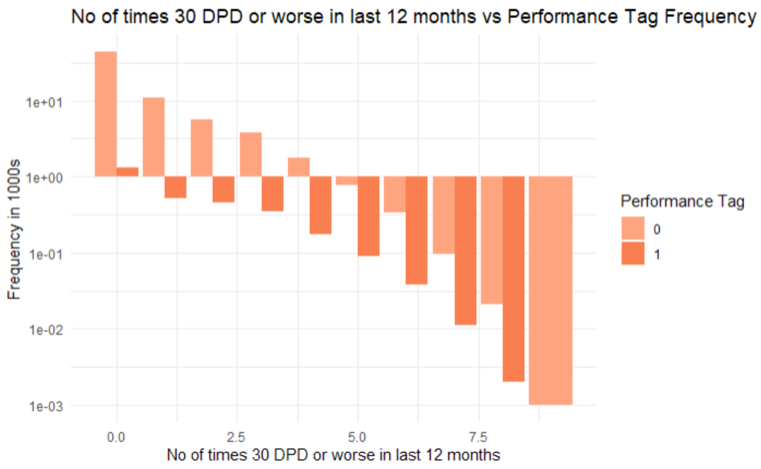

No of times 90 DPD or worse in last 6 months vs Performance Tag Frequency

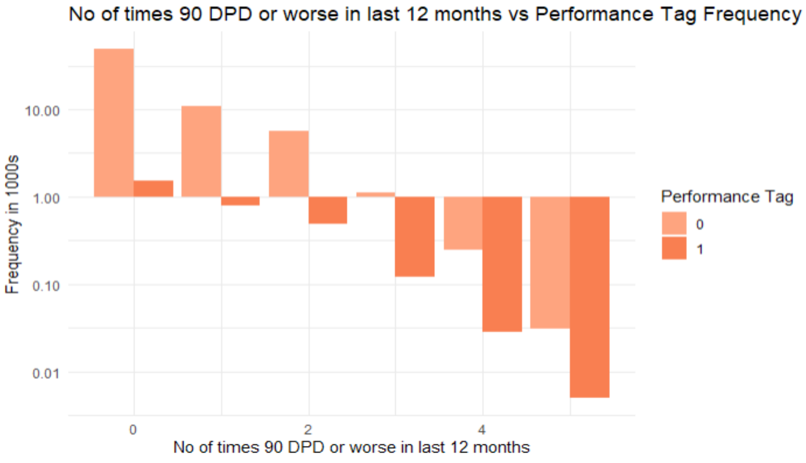
EXPLORATORY DATA ANALYSIS

30/60/90 DPD or Worse in Last 12 Months

- The plot below shows that both the highest number of customers as well as the highest number of defaulters are people who have never gone 30 days past due date in last 12 months.
- The first plot on the right shows that both the highest number of customers as well as the highest number of defaulters are people who have never gone 60 days past due date in last 12 months.
- The second plot shows that both the highest number of customers as well as the highest number of defaulters are people who have never gone 90 days past due date in last 12 months.


No of times 60 DPD or worse in last 12 months vs Performance Tag Frequency

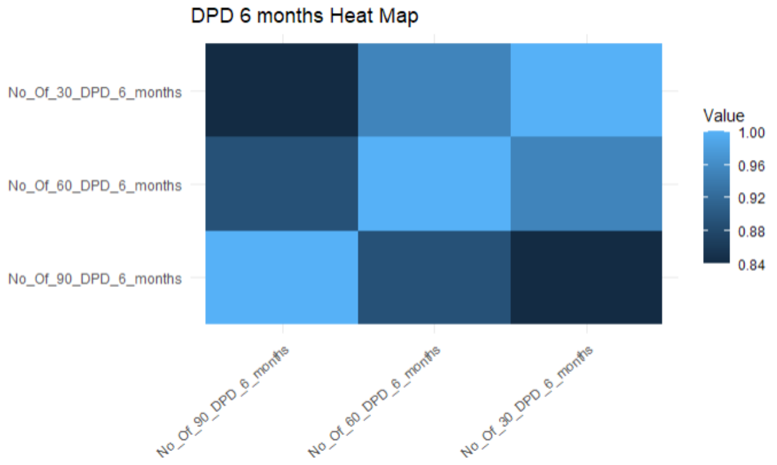

No of times 30 DPD or worse in last 12 months vs Performance Tag Frequency

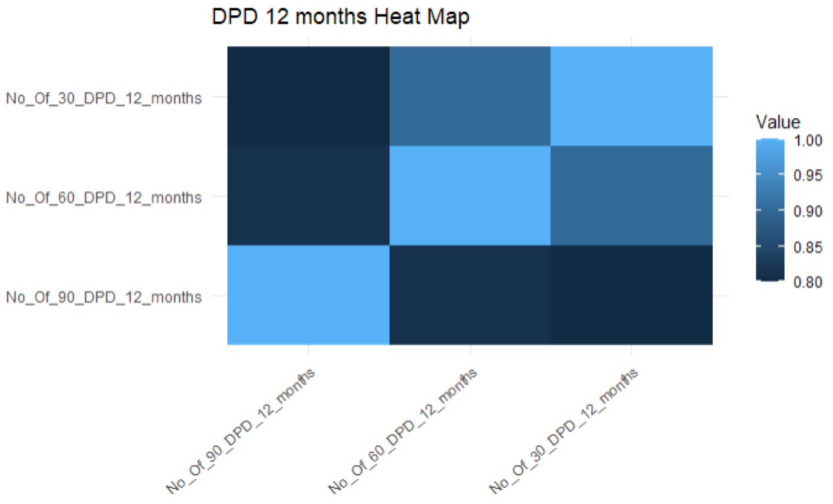

No of times 90 DPD or worse in last 12 months vs Performance Tag Frequency

EXPLORATORY DATA ANALYSIS

Correlation between DPD Variables

- The first correlation plot shows correlation between the various 6-month DPD variables. We can see that the variables have strong correlation between each other. The 30 DPD and 90 DPD have the weakest correlation.
- The second plot shows correlation between the several 12-month DPD variables. We can see that the variables have strong correlation between each other. The 30 DPD and 90 DPD have the weakest correlation followed by the 60 DPD and 90 DPD.

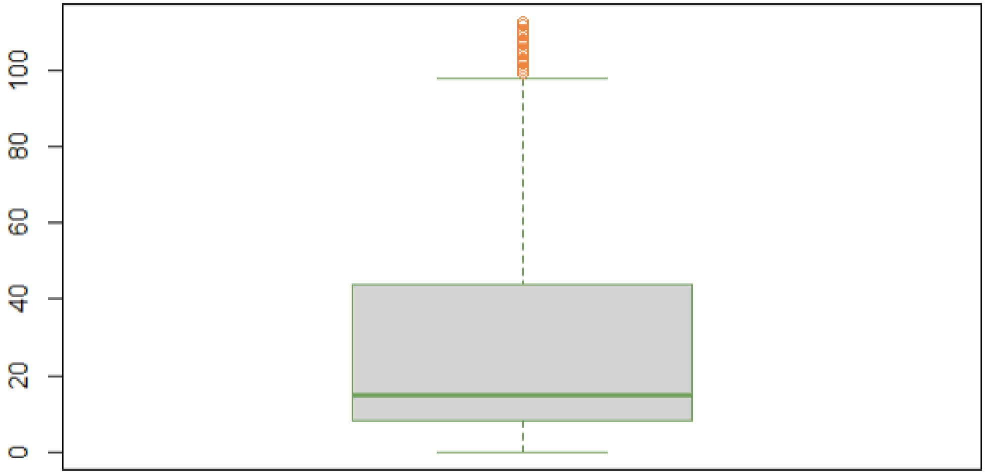




EXPLORATORY DATA ANALYSIS

Average Credit Card Utilization in Last 12 Months

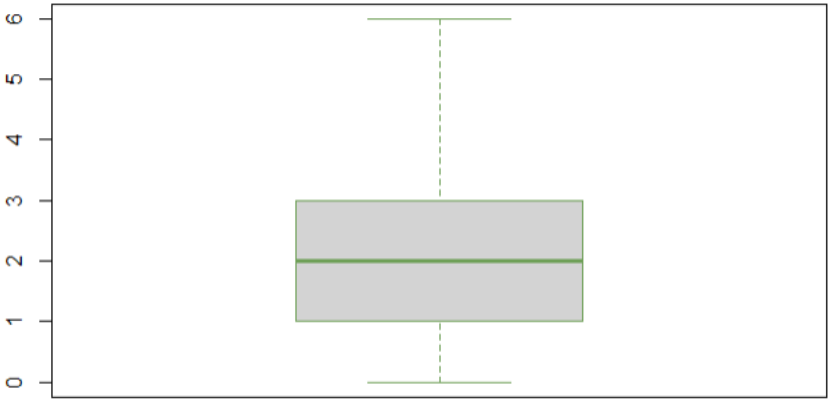
The boxplot below shows that the data has imbalance and there are multiple outliers.

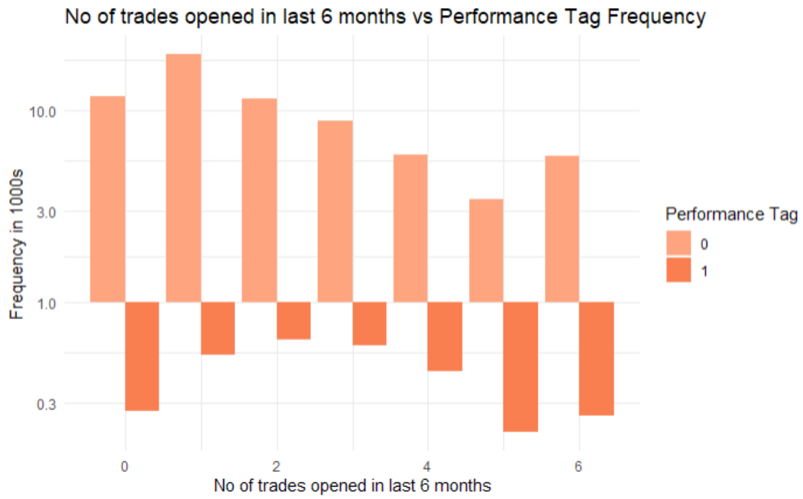


EXPLORATORY DATA ANALYSIS

Number of Trades Opened in Last 6 months

- The boxplot shows quartiles and median. There are no outliers.
- The second plot shows that the maximum number of customers have opened 1 trade in last 6 months. Most defaulters are the people who have opened 2 trades in last 6 months.

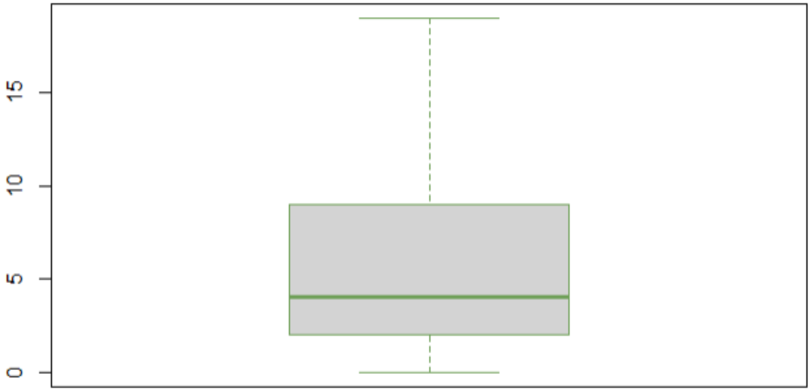


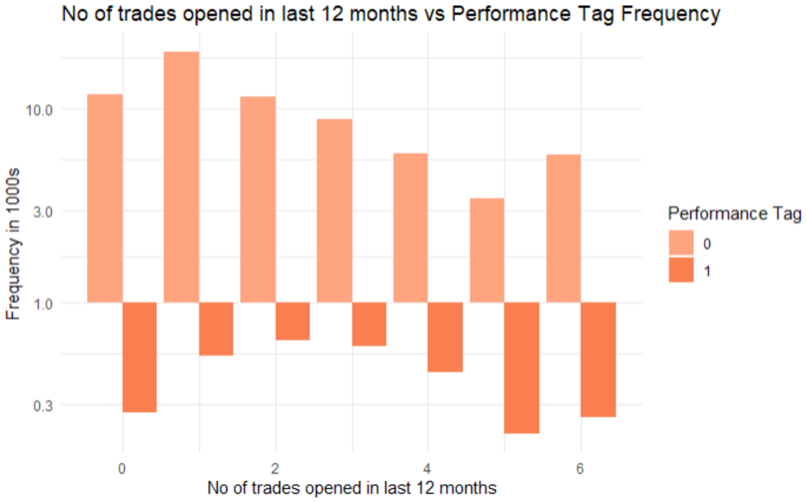


EXPLORATORY DATA ANALYSIS

No. of Trades Opened in Last 12 months

- The boxplot shows quartiles and median. There are no outliers. The data is skewed to the left.
- The second plot shows that the maximum number of customers have opened 1 trade in last 12 months. Most defaulters are the people who have opened 2 trades in last 12 months.

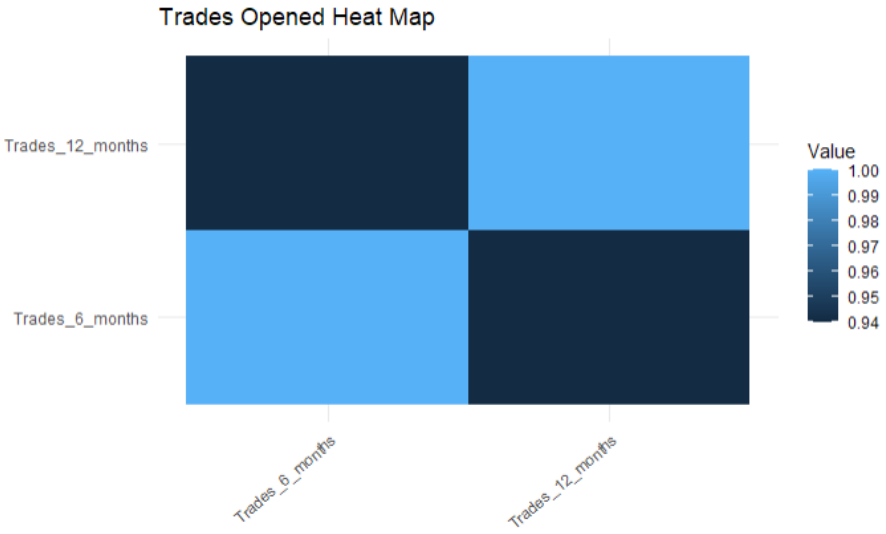




EXPLORATORY DATA ANALYSIS

Correlation between Trades Opened Variables in Last 6/12 Months

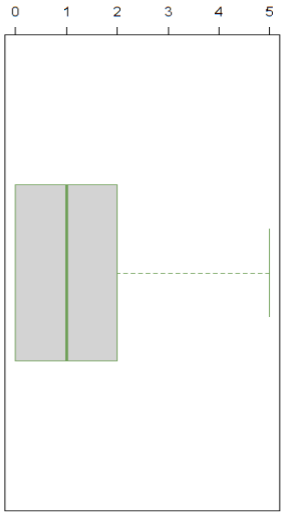
There is a correlation of 0.94 between Trades Opened in last 6 and 12 months.

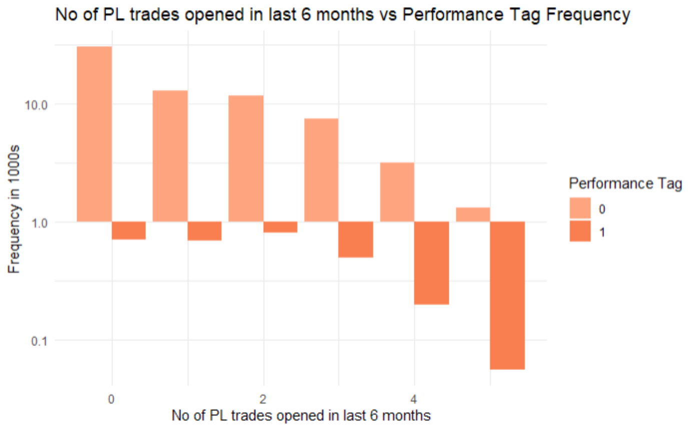


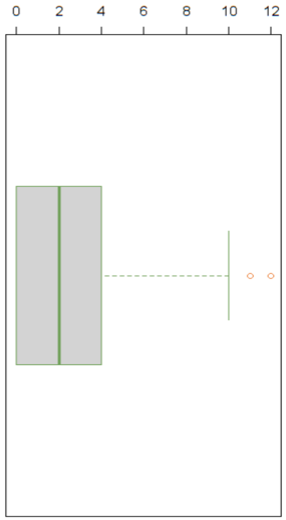
EXPLORATORY DATA ANALYSIS

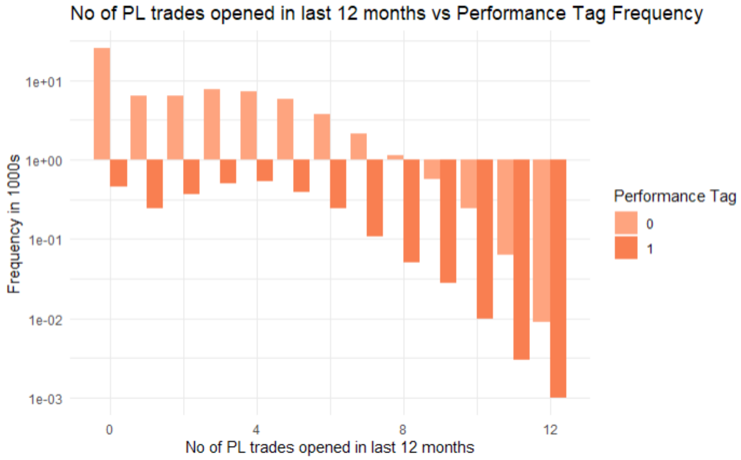
Number of PL Trades Opened in Last 6/12 months

- The boxplot shows quartiles and median. There are outliers for number of PL trades opened in last 12 months.
- The second plot shows that most people have defaulted when the number of PL trades opened in last 6 and 12 months is 2 and 4 respectively.





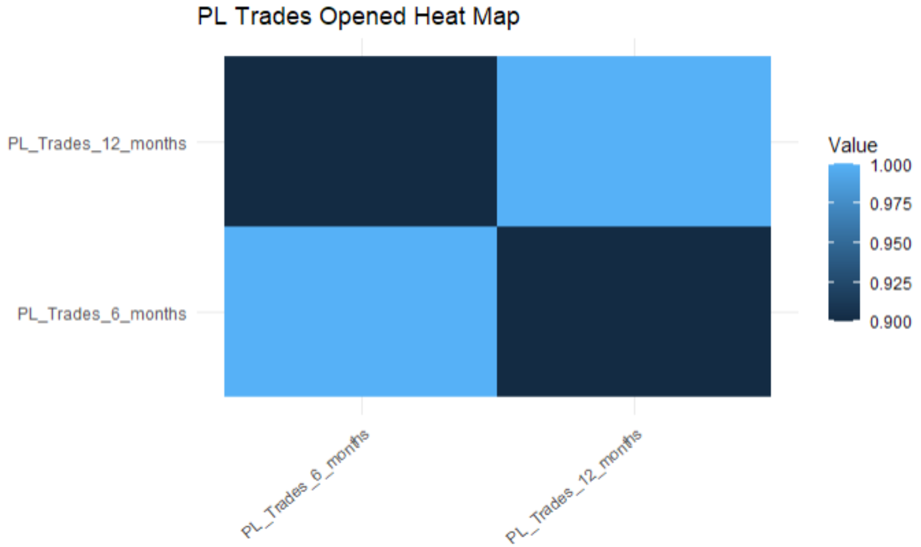




EXPLORATORY DATA ANALYSIS

Correlation between PL Trades Opened Variables in Last 6/12 Months

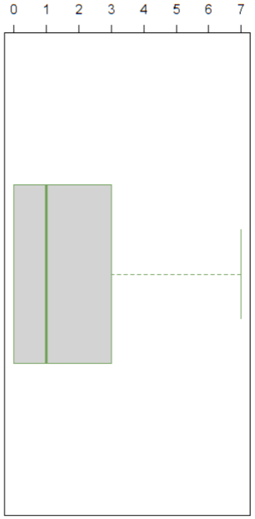
There is a correlation of 0.90 between PL Trades Opened in last 6 and 12 months.

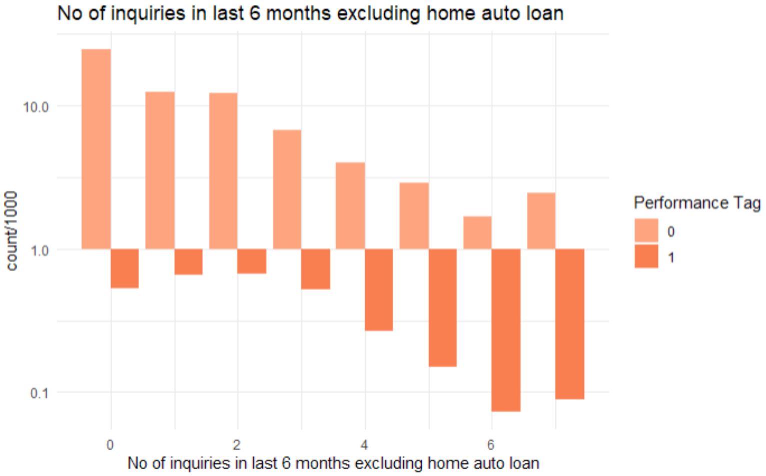


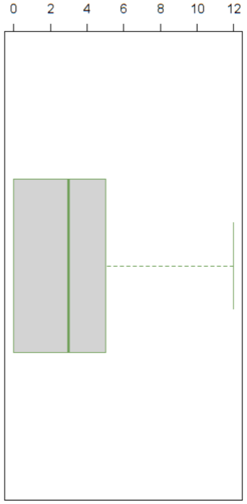
EXPLORATORY DATA ANALYSIS

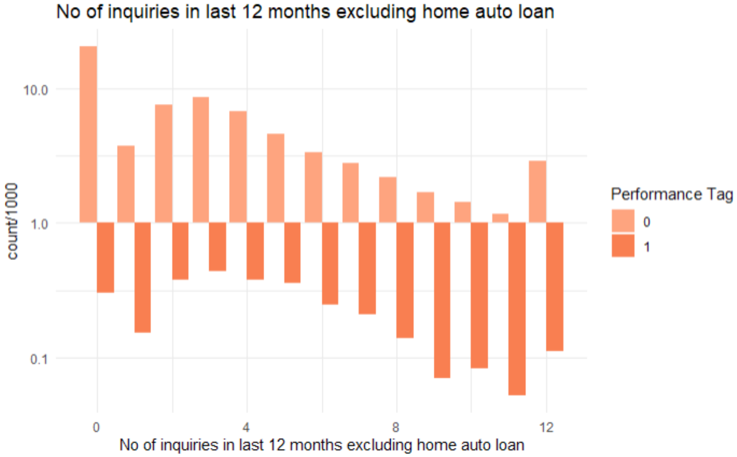
Number of Inquiries in Last 6/12 months

- The boxplots shows quartiles and median. There are no outliers. The data is skewed.
- The plots on the right show that most people default when they have 2 and 3 inquiries in last 6 and 12 months respectively.





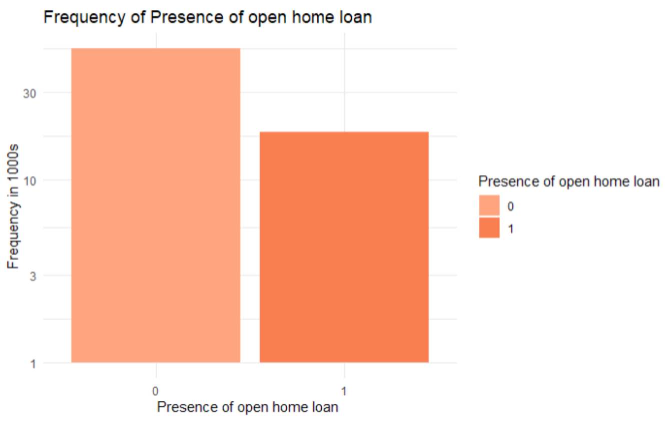




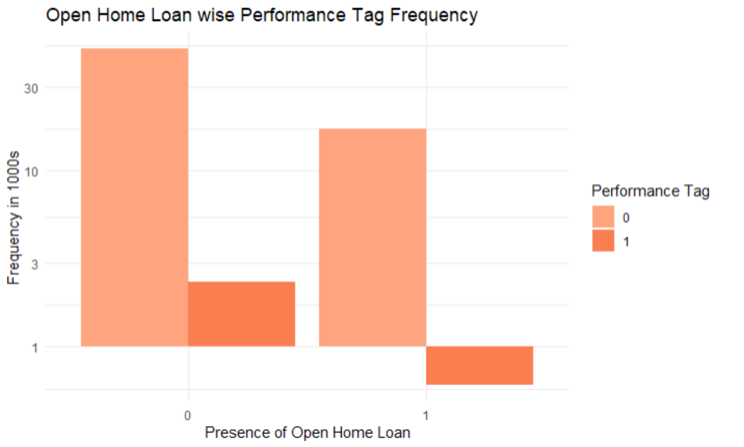
EXPLORATORY DATA ANALYSIS

Presence of Home Loan

- Most customers do not have an open home loan.
- Largest number of defaulters are among people who have an open home loan.



Frequency of Presence of open home loan

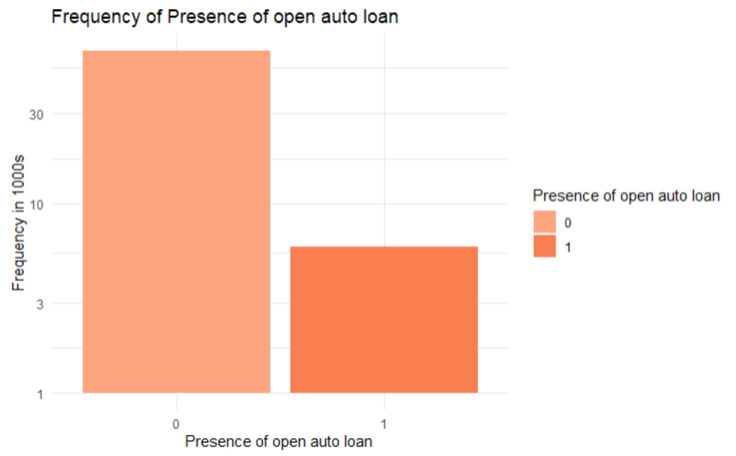


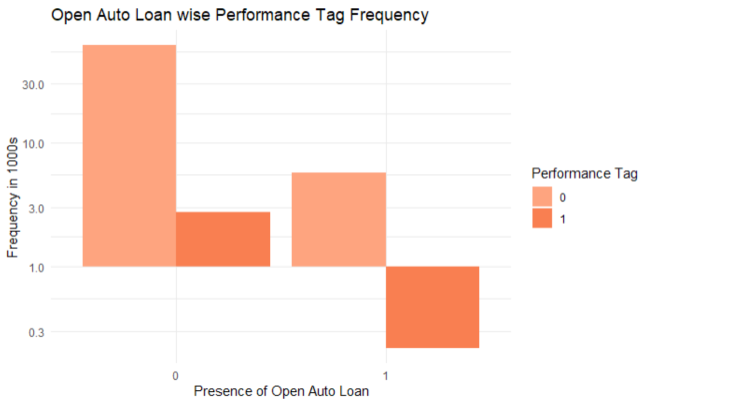
Open Home Loan wise Performance Tag Frequency

EXPLORATORY DATA ANALYSIS

Presence of Open Auto Loan

- Most customers are people who do not have an open auto loan.
- However, the highest number of defaulters are people who have an open auto loan.

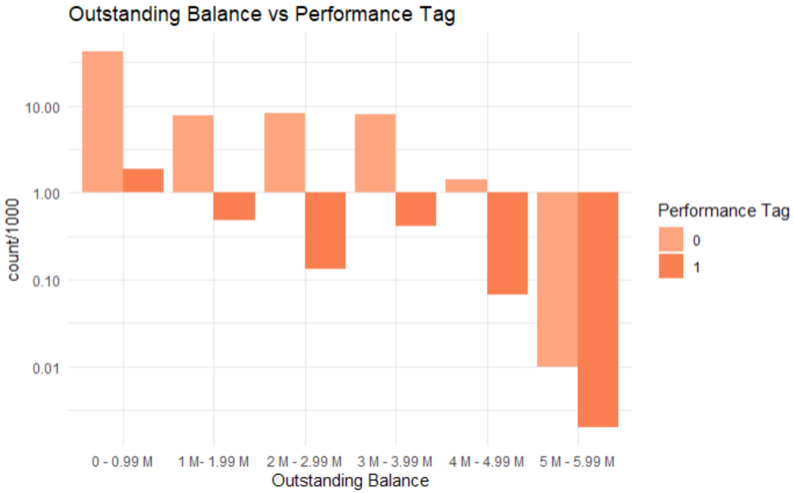

Frequency of Presence of open auto loan


Open Auto Loan wise Performance Tag Frequency

EXPLORATORY DATA ANALYSIS

Outstanding Balance

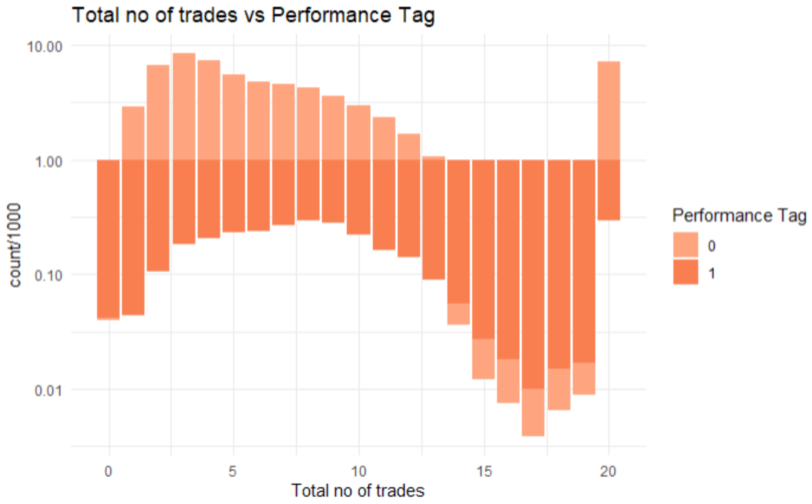
- Most people have an outstanding balance below 1 M.
- The highest number of defaulters are among people who have an outstanding balance below 1 M.



EXPLORATORY DATA ANALYSIS

Total No. of Trades

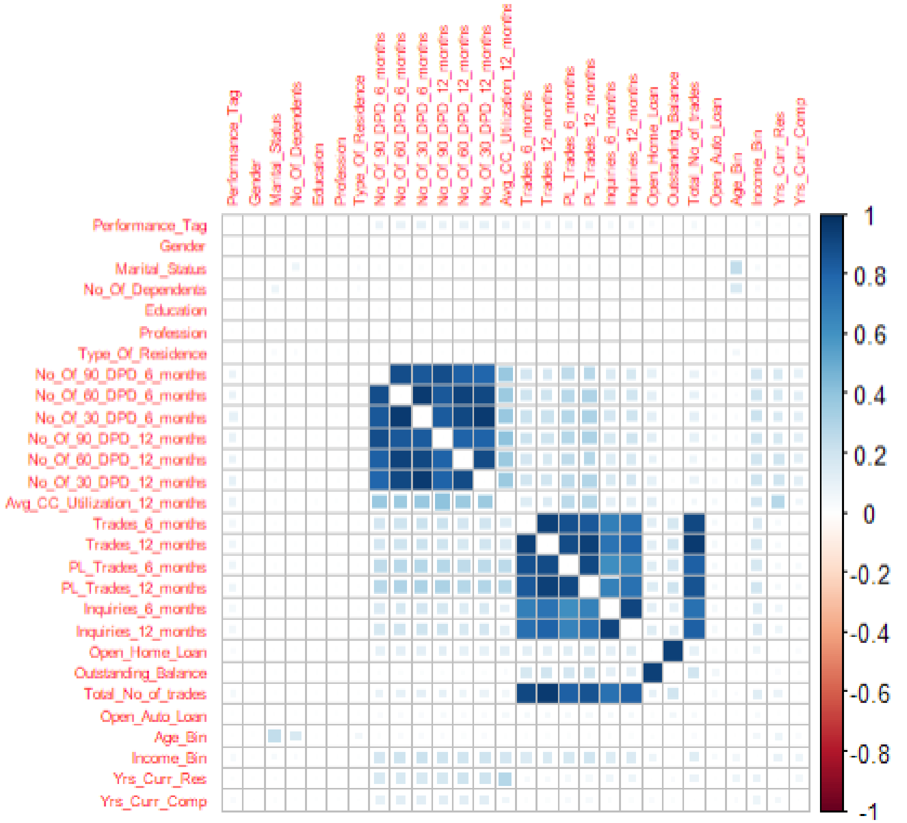
- Most customers have 3 trades.
- Segment with the most defaulters are customers who have 8 trades.



EXPLORATORY DATA ANALYSIS

Correlation among all variables

- We can see that all the DPD variables are strongly correlated to each other.
- Similarly, all the Trades and PL Trades account are strongly correlated to each other.



DATA TRANSFORMATION

Model Building Approach

- **Outlier Treatment:** Outlier detection is performed using boxplot on continuous variables and quantile functions. The variables with outliers have been corrected by capping the outliers to the nearest non-outlier variable at the boundary.
- **Data Split:** The final data set is split into a training set and a test set in 70:30 ratio for model building. All models are tested on test data sets that were kept separate from the training set.
- **Data Sampling:** The given data is highly imbalanced. We have used the technique of SMOTE to balance the dataset.

MODEL BUILDING – SUPPORT VECTOR MACHINES

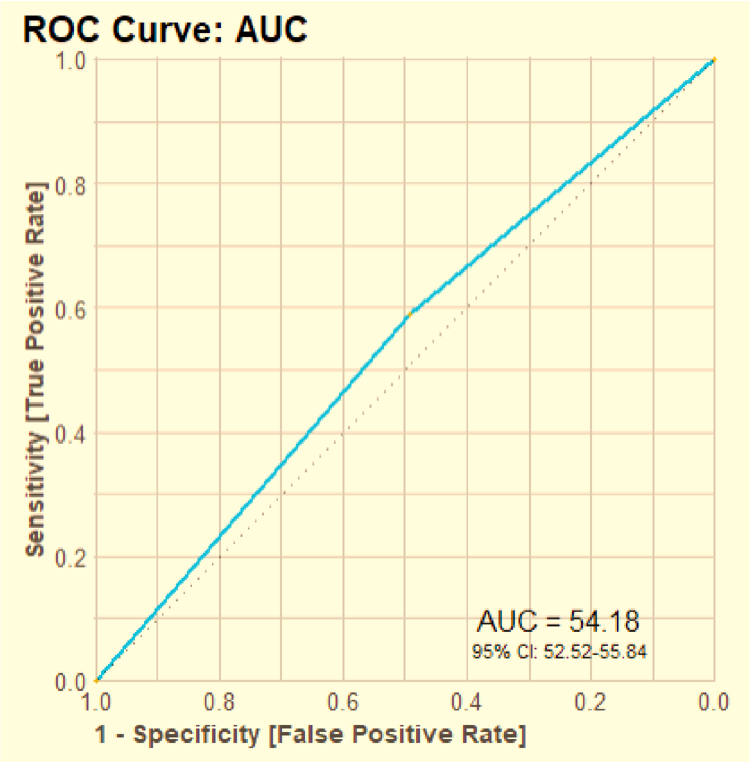
Model Variables

- **(Demographic) Independent Variables:** Age, Gender, Marital Status, Number of Dependents, Income, Education, Profession, Type of Residence, Number of Months in Current Residence, Number of Months in Current Company.
- **Dependent Variable:** Performance Tag

MODEL BUILDING – SUPPORT VECTOR MACHINES

ROC Curve

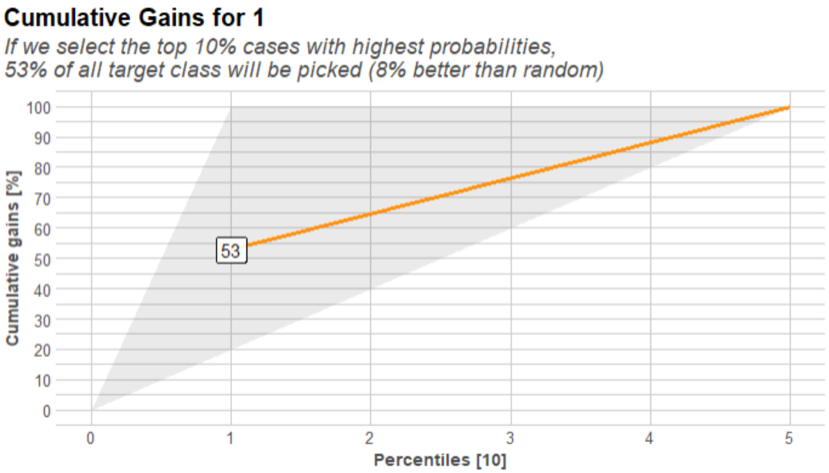
- Here are the important statistics for our model.
- The ROC curve shows that this model is almost as good as a random model only.
- A large number of non-defaulters are classified as defaulters, and therefore it is a poor model based only on the demographic data.



MODEL BUILDING – SUPPORT VECTOR MACHINES

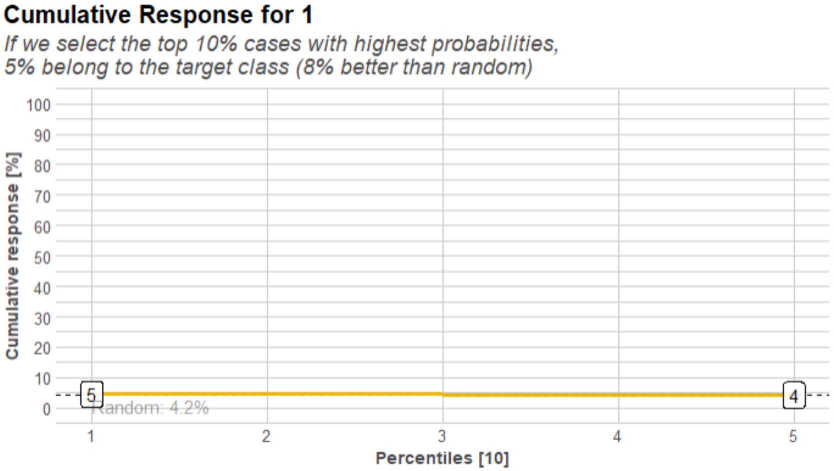
Gain and Response

- From the first plot on the right we can see that there is a gain of only 8 percent for the defaulter class. It means that if we select the top 10 percent cases with highest probabilities, 53 percent of all target class will be picked (1 here).
- The second plot on the right shows that if we select the top 10 percent cases with highest probabilities, 5 percent belong to the target class (1 here).



Cumulative Gains for 1

If we select the top 10% cases with highest probabilities, 53% of all target class will be picked (8% better than random)



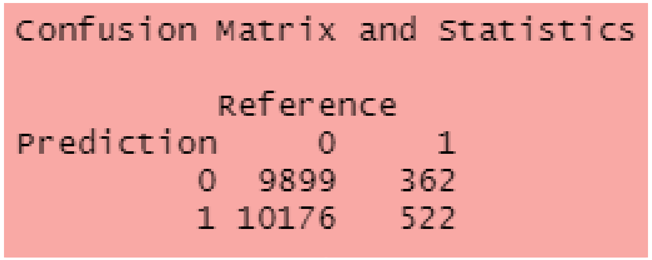
Cumulative Response for 1

If we select the top 10% cases with highest probabilities, 5% belong to the target class (8% better than random)

MODEL BUILDING – SUPPORT VECTOR MACHINES

Confusion Matrix

- The confusion matrix on the right shows that a large number of non-defaulters are classified as defaulters, and therefore it is a poor model based only on the demographic data.
- The accuracy is too low and performs worse than a random model under certain conditions.



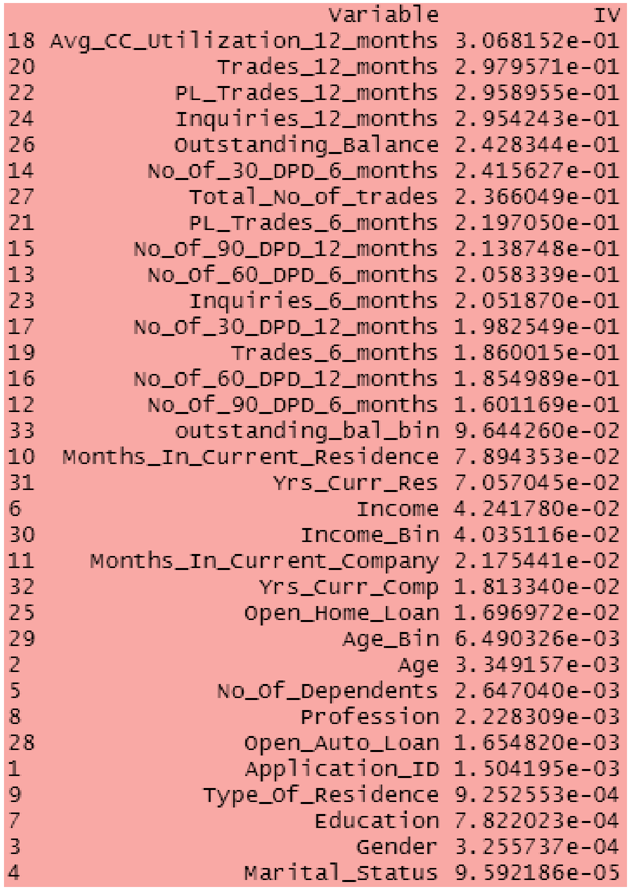
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  9899   362
         1 10176   522
```

Statistic	Value
Accuracy	49.72%
Sensitivity	49.31%
Specificity	59.05%

MODEL BUILDING – LOGISTIC REGRESSION

WOE and IV Analysis

- We have converted the variables into their WOE bins and calculated their Information Values. This is done to understand the relevance of each of the variables in determining the target variable.
- On the right is a list of variables in decreasing order of their IV values. There are no variables with a very strong information content.
- In our logistic regression model, we have used only the first 12 variables on the basis of their IV value

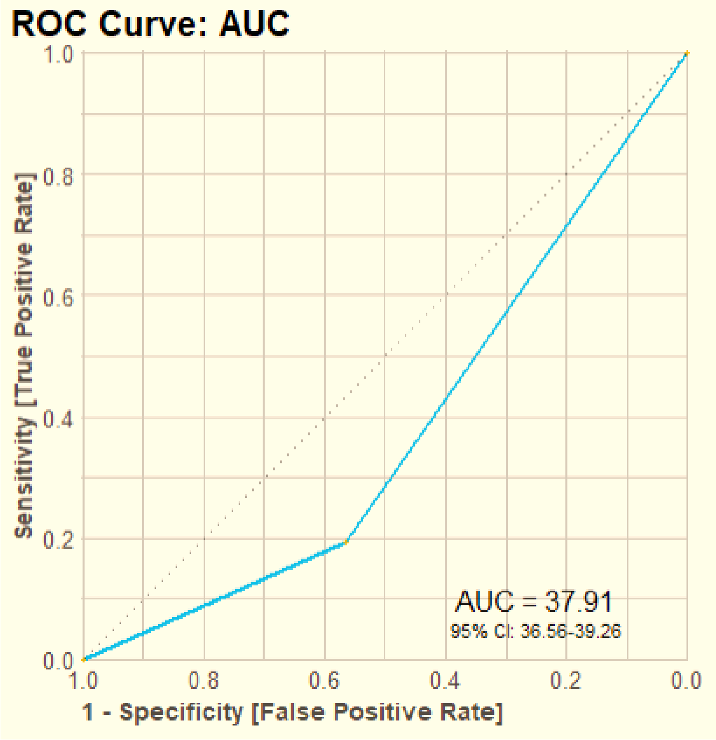


```
                       Variable           IV
18 Avg_CC_Utilization_12_months 3.068152e-01
20             Trades_12_months 2.979571e-01
22          PL_Trades_12_months 2.958955e-01
24          Inquiries_12_months 2.954243e-01
26          Outstanding_Balance 2.428344e-01
14        No_Of_30_DPD_6_months 2.415627e-01
27           Total_No_of_trades 2.366049e-01
21           PL_Trades_6_months 2.197050e-01
15       No_Of_90_DPD_12_months 2.138748e-01
13        No_Of_60_DPD_6_months 2.058339e-01
23           Inquiries_6_months 2.051870e-01
17       No_Of_30_DPD_12_months 1.982549e-01
19              Trades_6_months 1.860015e-01
16       No_Of_60_DPD_12_months 1.854989e-01
12        No_Of_90_DPD_6_months 1.601169e-01
33          outstanding_bal_bin 9.644260e-02
10  Months_In_Current_Residence 7.894353e-02
31                 Yrs_Curr_Res 7.057045e-02
6                        Income 4.241780e-02
30                   Income_Bin 4.035116e-02
11    Months_In_Current_Company 2.175441e-02
32                Yrs_Curr_Comp 1.813340e-02
25               Open_Home_Loan 1.696972e-02
29                      Age_Bin 6.490326e-03
2                           Age 3.349157e-03
5              No_Of_Dependents 2.647040e-03
8                    Profession 2.228309e-03
28               Open_Auto_Loan 1.654820e-03
1                Application_ID 1.504195e-03
9             Type_Of_Residence 9.252553e-04
7                     Education 7.822023e-04
3                        Gender 3.255737e-04
4                Marital_Status 9.592186e-05
```

MODEL BUILDING – LOGISTIC REGRESSION

ROC Curve

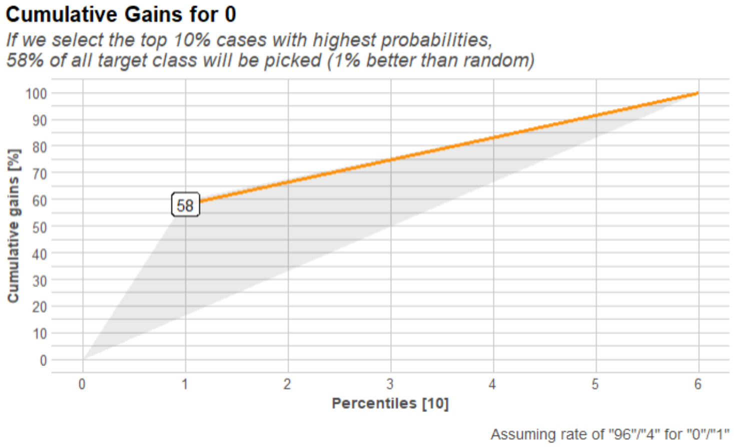
- Here are the important statistics for our model.
- The ROC curve shows that this model is not a better classifier than our previous SVM model. The AUC is smaller here.



MODEL BUILDING – LOGISTIC REGRESSION

Gain and Response

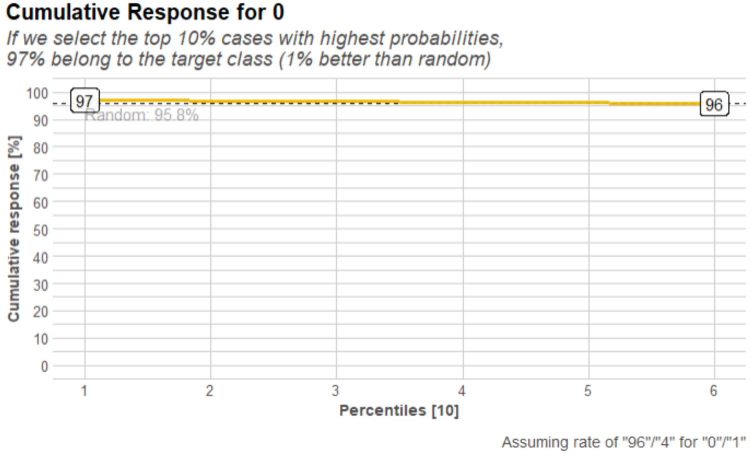
- From the first plot on the right we can see that there is a gain of only 1 percent for the non-defaulter class. It means that if we select the top 10 percent cases with highest probabilities, 58 percent of all target class will be picked (0 here).
- The second plot on the right shows that if we select the top 10 percent cases with highest probabilities, 97 percent belong to the target class (0 here). This is expected as variables and data are more inclined to the non-defaulter class.



Cumulative Gains for 0

If we select the top 10% cases with highest probabilities, 58% of all target class will be picked (1% better than random)

Assuming rate of "96"/"4" for "0"/"1"



Cumulative Response for 0

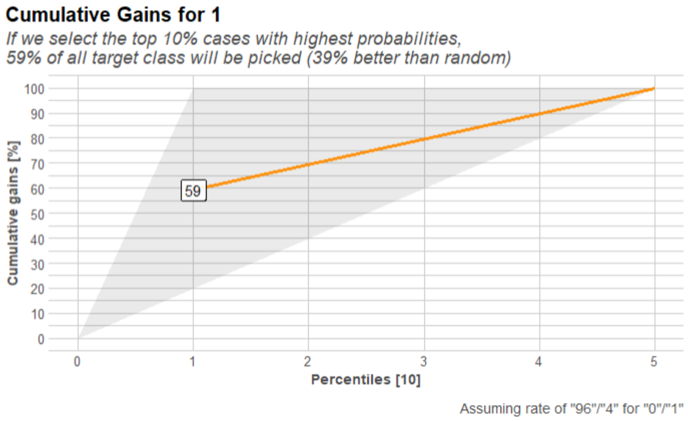
If we select the top 10% cases with highest probabilities, 97% belong to the target class (1% better than random)

Assuming rate of "96"/"4" for "0"/"1"

MODEL BUILDING – LOGISTIC REGRESSION

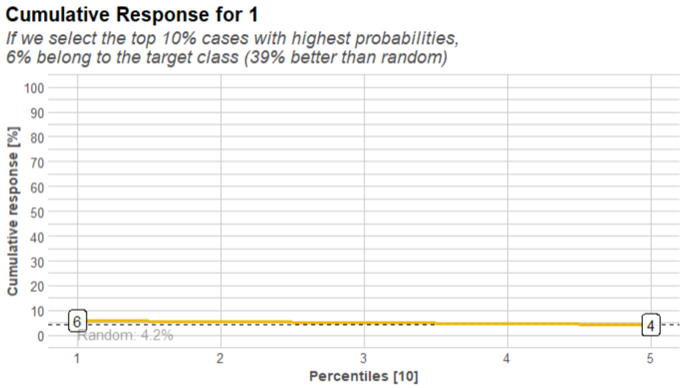
Gain and Response

- From the first plot on the right we can see that there is a gain of only 8 percent for the defaulter class. It means that if we select the top 10 percent cases with highest probabilities, 59 percent of all target class will be picked (1 here).
- The second plot on the right shows that if we select the top 10 percent cases with highest probabilities, 6 percent belong to the target class (1 here). This is 39 percent better than the random data.


Cumulative Gains for 1

If we select the top 10% cases with highest probabilities, 59% of all target class will be picked (39% better than random)

Assuming rate of "96"/"4" for "0"/"1"

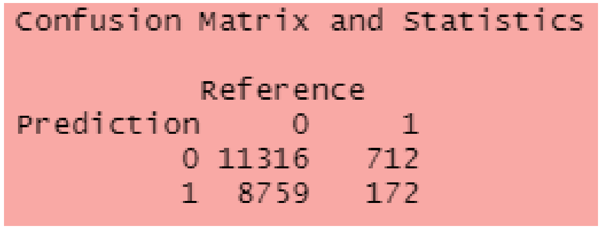

Cumulative Response for 1

If we select the top 10% cases with highest probabilities, 6% belong to the target class (39% better than random)

MODEL BUILDING – LOGISTIC REGRESSION

Confusion Matrix

- The confusion matrix on the right shows that a large number of defaulters are classified as non-defaulters, and therefore it is a poor model based on the combined demographic and credit data.
- The accuracy is higher than our previous SVM model.



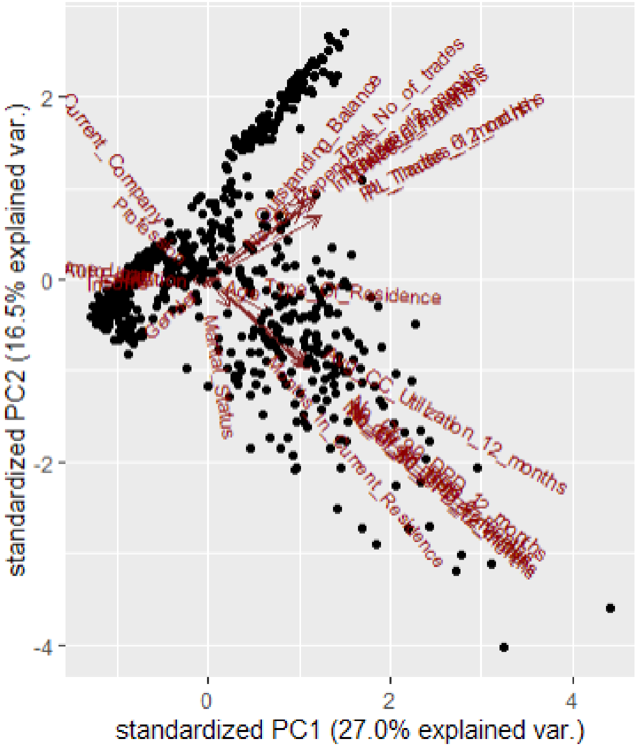
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11316   712
         1  8759   172
```

Statistic	Value
Accuracy	54.81%
Sensitivity	56.37%
Specificity	59.05%

MODEL BUILDING – RANDOM FOREST

Principal Component Analysis

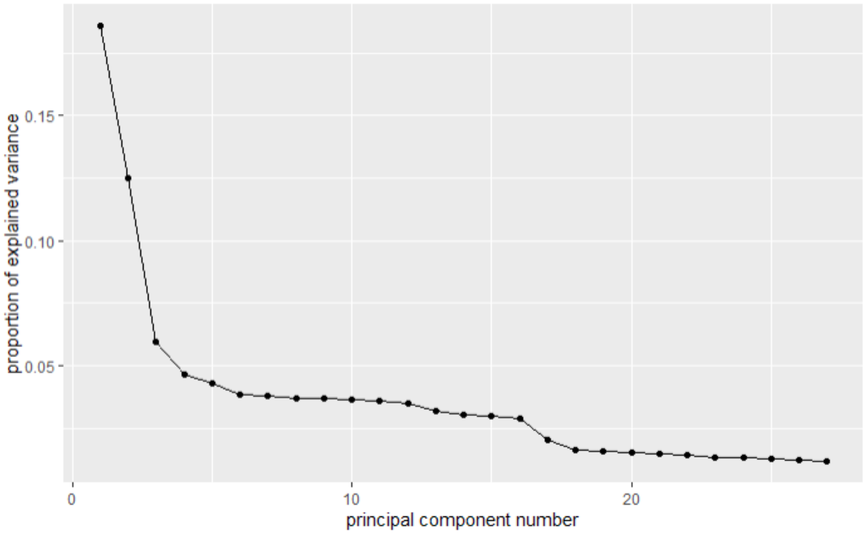
- We performed PCA on our dataset to reduce the number of dimensions
- As seen in the plot, the axes were rotated to obtain the principal components.
- PC1 explains 27 percent of the variance while PC2 explains 16.5 percent of the total variance in our data.



MODEL BUILDING – RANDOM FOREST

Principal Component Analysis

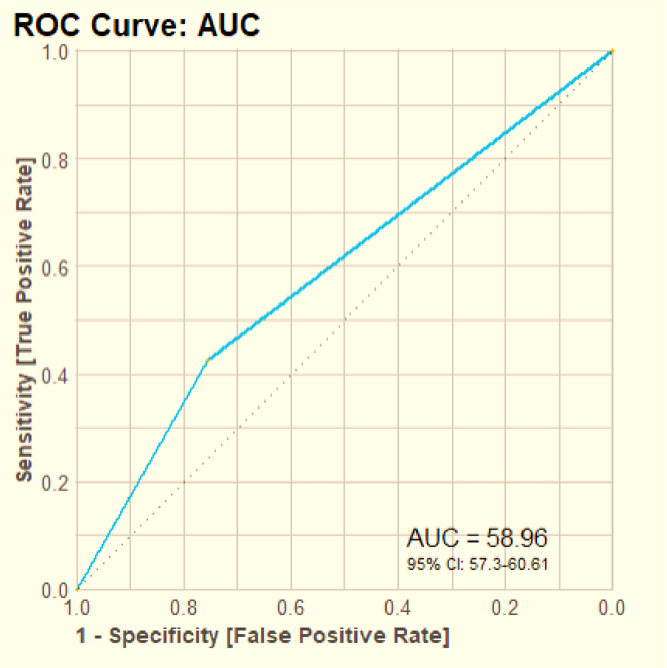
- The plot on the right shows the overall percentage of variance explained by the several principal components.
- Essentially, 97.5 percent variance can be explained through 20 principal components
- We have used only the first 20 principal components in our modeling.



MODEL BUILDING – RANDOM FOREST

ROC Curve

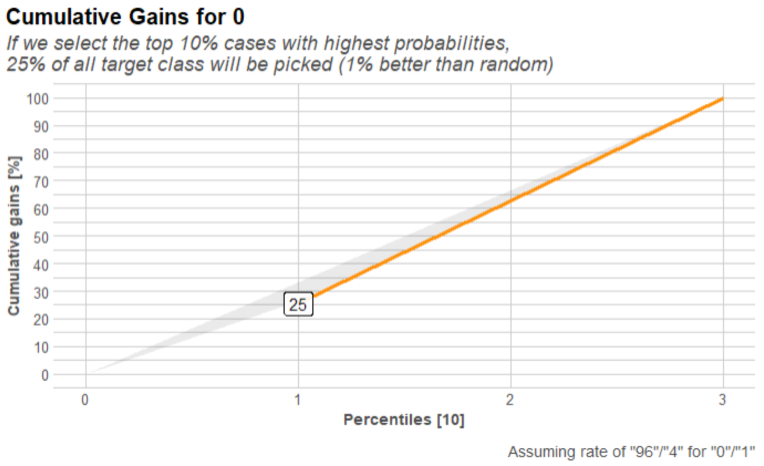
- Here are the important statistics for our model.
- The ROC curve shows that this model is a better classifier than our previous classification models. The AUC is the largest among all the three models, thus indicating that this model is most efficient at classifying the class variable.



MODEL BUILDING – RANDOM FOREST

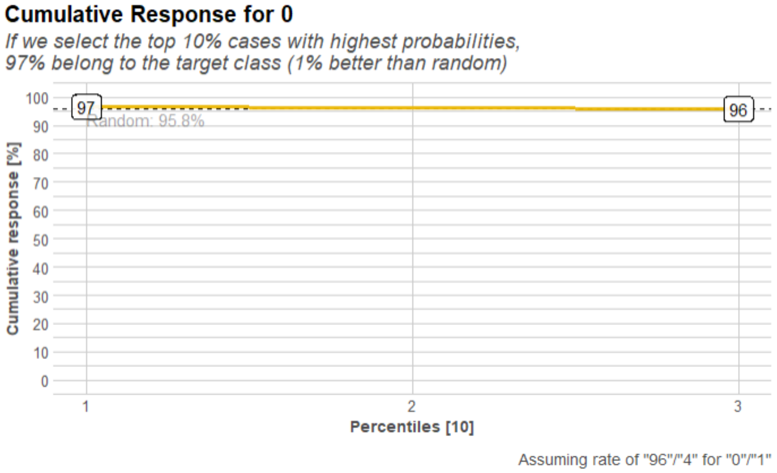
Gain and Response

- From the first plot on the right we can see that there is a gain of 1 percent for the non-defaulter class. It means that if we select the top 10 percent cases with highest probabilities, 25 percent of all target class will be picked (0 here).
- The second plot on the right shows that if we select the top 10 percent cases with highest probabilities, 97 percent belong to the target class (0 here). This is expected as our variables and data are more inclined to the non-defaulter class.


Cumulative Gains for 0

If we select the top 10% cases with highest probabilities, 25% of all target class will be picked (1% better than random)

Assuming rate of "96"/"4" for "0"/"1"


Cumulative Response for 0

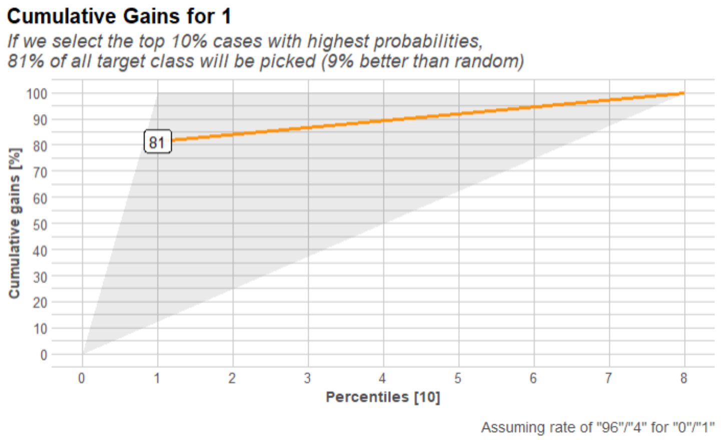
If we select the top 10% cases with highest probabilities, 97% belong to the target class (1% better than random)

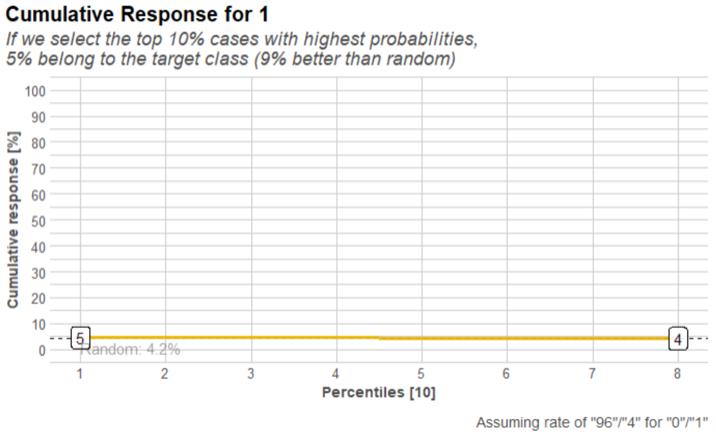
Assuming rate of "96"/"4" for "0"/"1"

MODEL BUILDING – RANDOM FOREST

Gain and Response

- From the first plot on the right we can see that there is a gain of only 8 percent for the defaulter class. It means that if we select the top 10 percent cases with highest probabilities, 81 percent of all target class will be picked (1 here).
- The second plot on the right shows that if we select the top 10 percent cases with highest probabilities, 5 percent belong to the target class (1 here). This is 9 percent better than the random data.

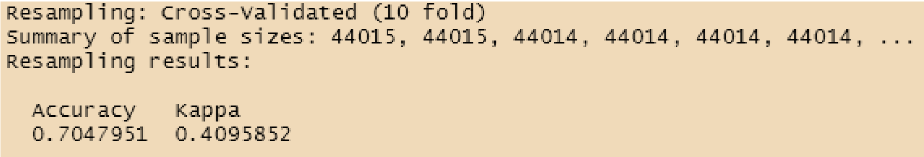




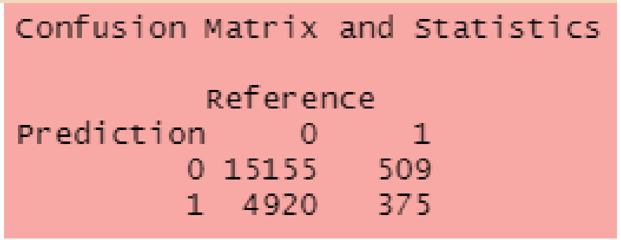
MODEL BUILDING – RANDOM FOREST

Cross Validation & Confusion Matrix

- 10-fold cross-validation was performed on the Random Forest model, and the accuracy turned out to be around 70.48 percent.
- The confusion matrix on the right shows a good classification ratio. We will assess the cost performance of the model using this matrix in the next section.
- The accuracy and sensitivity is higher than our previous classifier models.



```
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 44015, 44015, 44014, 44014, 44014, 44014, ...
Resampling results:

  Accuracy   Kappa
  0.7047951  0.4095852
```



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15155   509
         1  4920   375
```

Statistic	Value
Accuracy	74.10%
Sensitivity	75.49%
Specificity	42.42%

MODEL BUILDING – CHOOSING THE BEST MODEL

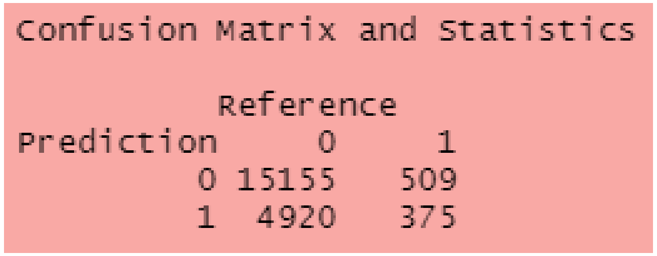
- From our previous analysis, we have observed that:
  - The Random Forest model has the highest AUC, therefore it has the best classification ability among all the three models.
  - It has the highest accuracy and sensitivity.

Therefore, we choose the RF model as the best classifier and will perform the Cost Benefit Analysis on this model.

COST BENEFIT ANALYSIS

Data Description

- In our CBA dataset we have 20,959 rows of data.
- 20,075 of those are non-defaulters and 884 are defaulters per the original data.
- If our model is used, according to the matrix on the right, there will be 15,664 (12,155 + 509)customers whose credit card application will be accepted. The remaining 5,295 (4,920 + 375) applications will be rejected as they have been predicted as defaulters.
- Out of these 15,664 accepted customers, 509 will default since they are false positives.



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15155   509
         1  4920   375
```

COST BENEFIT ANALYSIS

Calculating Net Gain without and with our Model

- Assumptions
  - Average profit from each non-defaulter: CAD 5,000
  - Average capital loss due to each defaulter: CAD 100,000
- Net Profit without using our Model:
  - Total Profit from all non-defaulters: 20,075*5000 = CAD 100,375,000
  - Total Loss due to all the defaulters: 884*10,0000 = CAD 88,400,000
  - Net Profit = CAD 11,975,000
- Net Profit when using our Model:
  - Total Profit from all the true positives: 5,155 * 5,000 = CAD 75,775,000
  - Total loss due to false positives = CAD 50,900,000
  - Net Profit = CAD 24,875,000

Financial Advantage: (24,875,000 – 11,975,000) = CAD 12,900,000

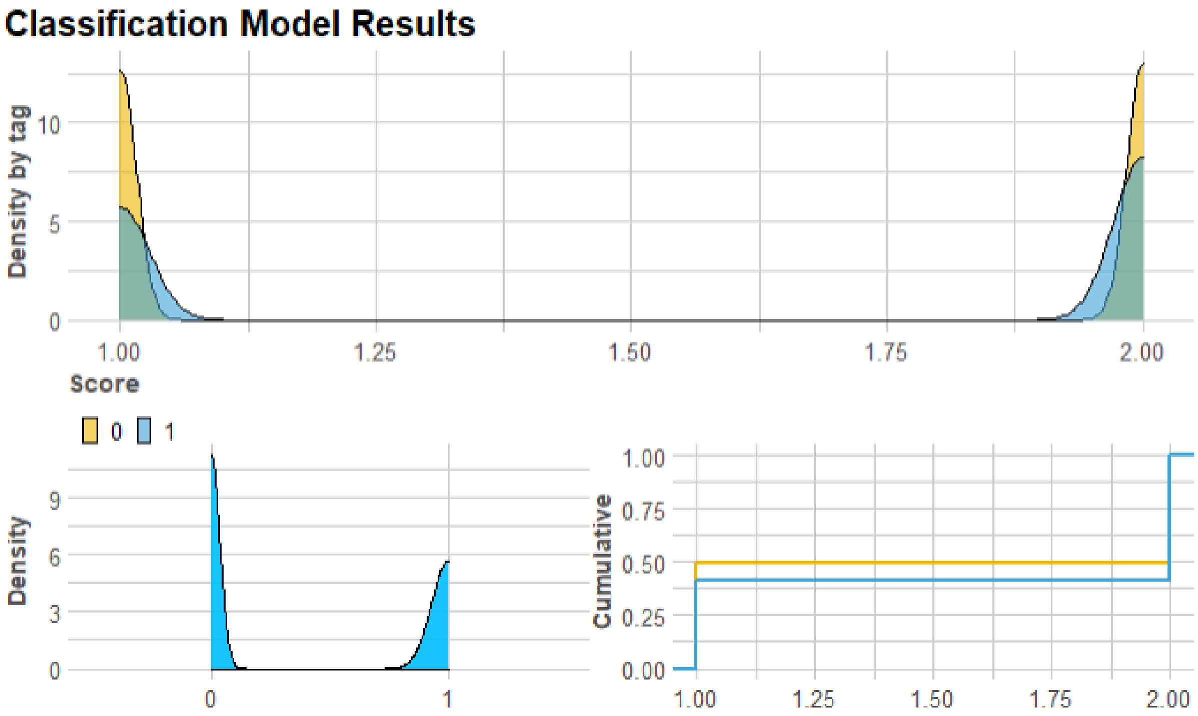
CONCLUSION

- Random Forest model is chosen as the final model with 74.10 percent accuracy.
- We found that our credit loss percentage was decreased when we used this model.
- 31.40 percent of the non-defaulting candidates were rejected by our model, which resulted in a revenue loss. However, overall our model has provided a net financial gain/advantage of approximately **CAD 13 Million.**

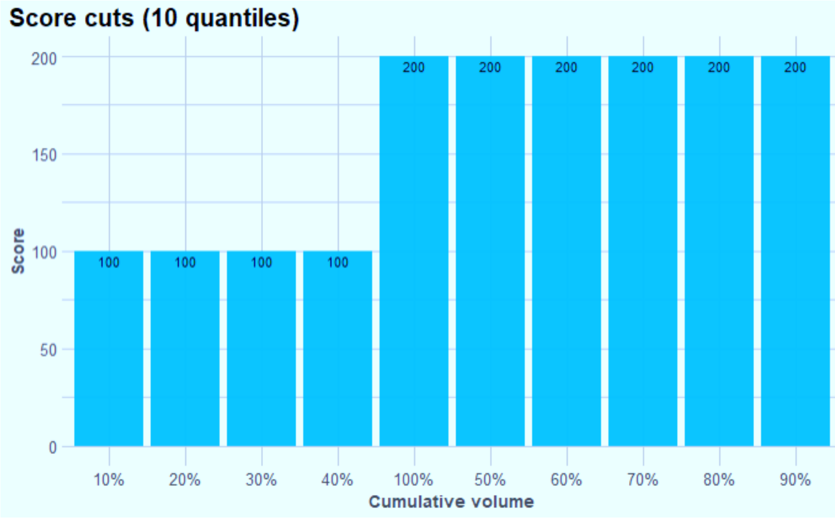
APPENDIX

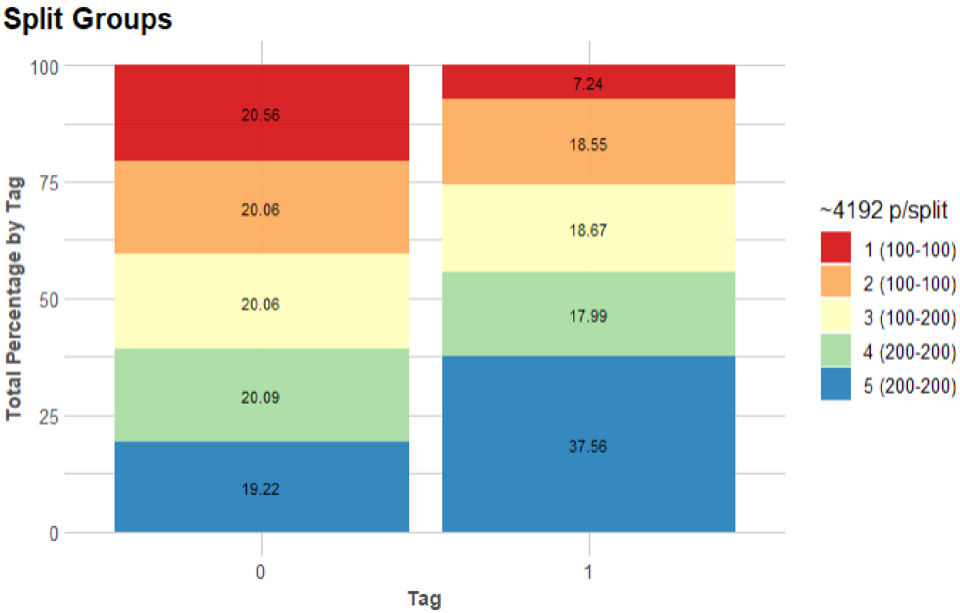
This section contains some of the plots that were not used in the main analysis. These have been attached here without interpretational comments.

SVM DENSITY PLOTS

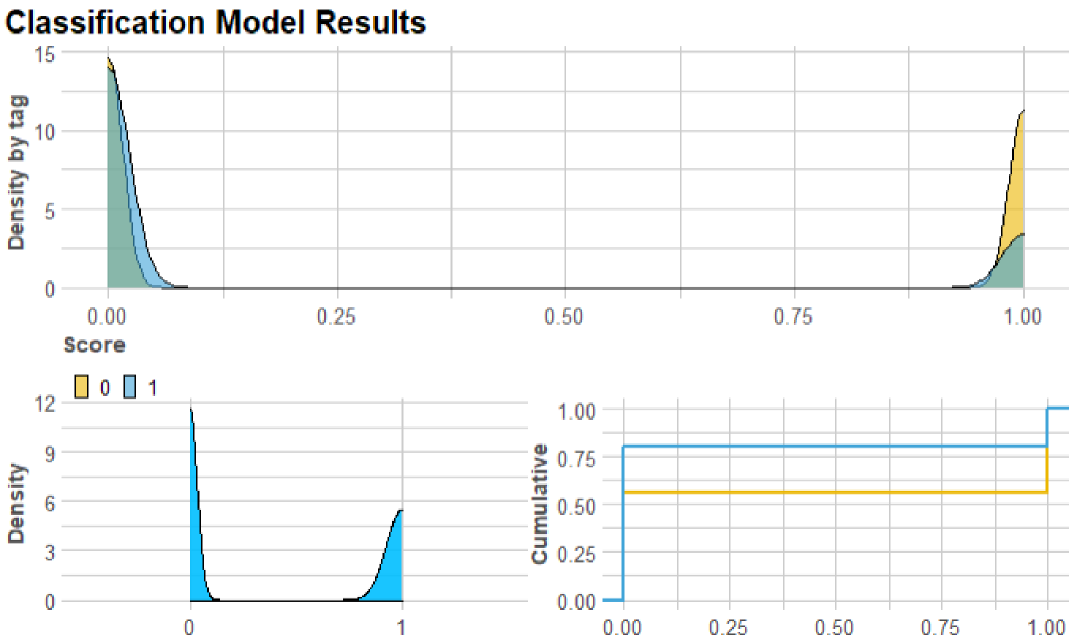


SVM CUTS AND SPLIT PLOTS



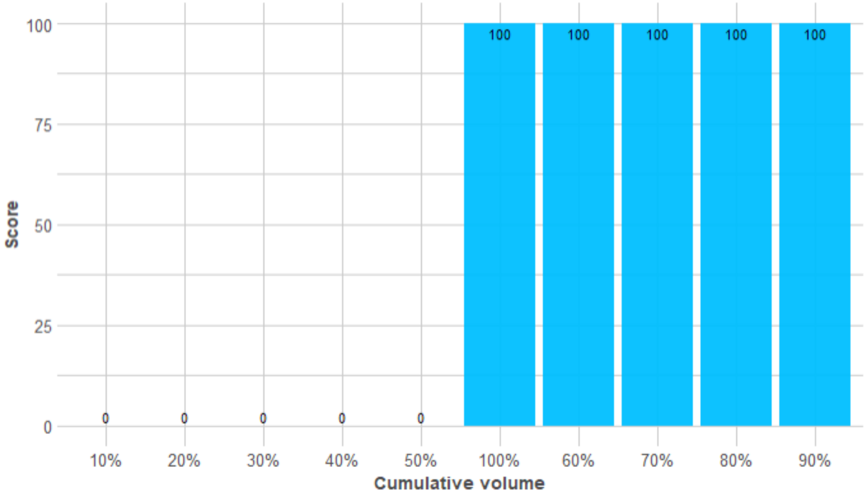


LOGISTIC REGRESSION DENSITY PLOTS

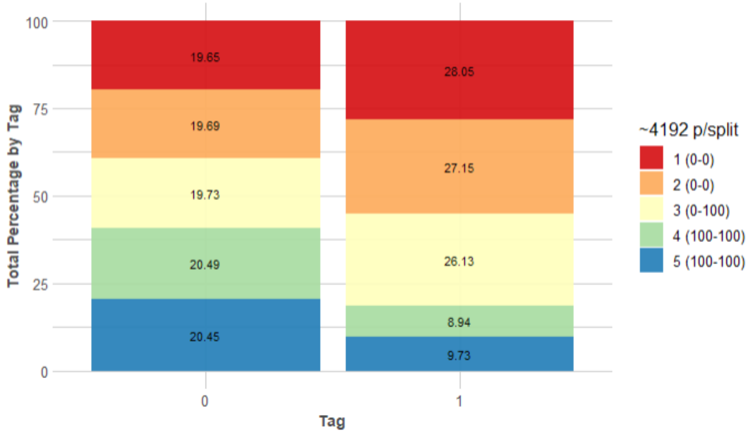


LOGISTIC REGRESSION CUTS AND SPLIT PLOTS

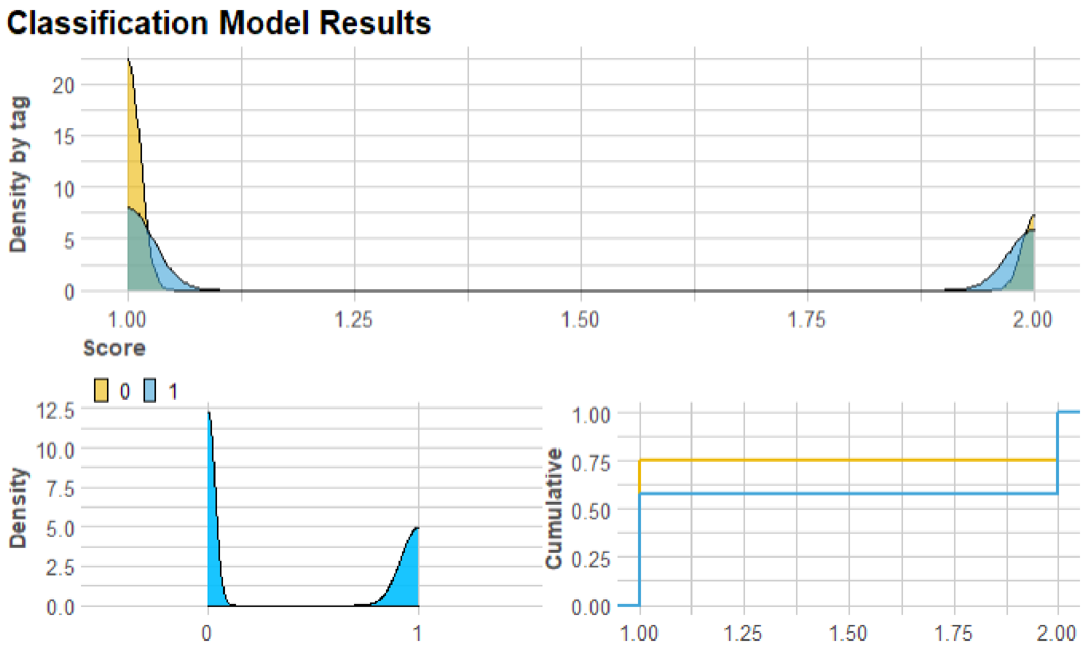
Score cuts (10 quantiles)



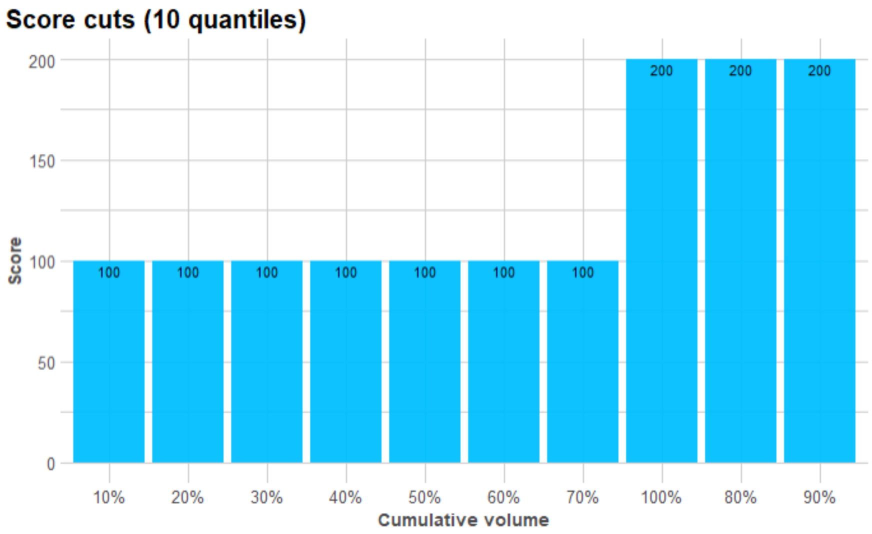
Split Groups

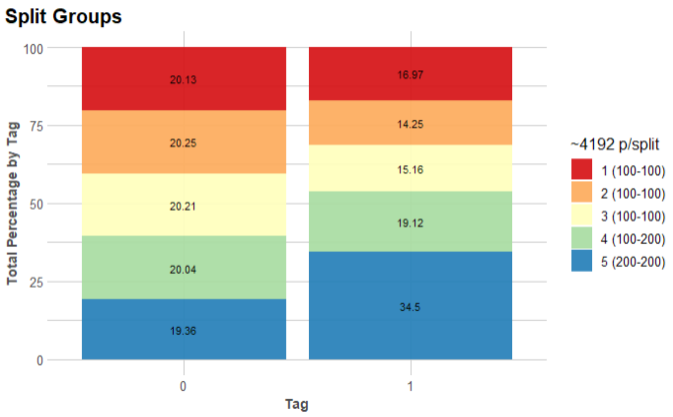


RANDOM FOREST DENSITY PLOTS



RANDOM FOREST CUTS AND SPLIT PLOTS


Score cuts (10 quantiles)


Split Groups