BFS Capstone Project - CredX

FINAL SUBMISSION

Group Name:

1. Arpita Ghosh
2. Biplab Ghosal
3. Jagannath Sen
4. Pritam Pan

Abstract

CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss. The Leadership team believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

The primary objective is to identify the right customers using predictive models. Using history data, need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the company.

2 datasets are available for analysing the requirements of CredX. Demographic file contains applicant’s provided information during credit card application. Credit bureau file contains credit information about the applicants. Both the datasets have a Performance Tag column which signifies whether the person has defaulted after getting a credit card.

Predictive Analysis for CredX

Business Understanding	Data Understating	Data Preparation & EDA	Model Building	Model Evaluation	Application Scorecard	Financial Benefit
Using past data of the bank's applicants, need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of CredX	2 datasets are available for analysing the requirements. Demographic/ application data contains the information provided by the applicants at the time of credit card application. Credit bureau contains the information taken from the credit bureau such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.	In this phase , started with EDA processes and followed by Data preparation steps like checking for duplicate data, missing values identify and removal process, creation, outliers treatment, variable standardization etc. Applied IV and WOE methodology to identify important variables	Firstly , splitting the data set in two parts (train and test). Built models on demographic data and merged data set. Used logistics regression, Decision Tree, Random Forest and SVM algorithms for model building.	Evaluated the likelihood of default for the rejected candidates and assess whether the results meet to expectations.	Built an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.	Assessed potential financial benefit and identified important metrics which are optimized.

Data Cleaning and Preparation Steps

- Removed all records which has person’s age less than 18. Here we are assuming that minors are not eligible for credit cards.
- Removed all records having income less than 0 as it is not possible. Since percentage of these records are even less than 0.5% , hence removing them.
- Removed duplicate records from the datasets. The percentage of duplicate records is less than 1%, hence removing those.
- Merged the two datasets based on ‘*Application.ID’* variable
- Identified and separated missing values corresponding to the Performance Tag variable from the merged data set. This separate dataset can be later used for testing purpose.
- Checked for outliers and removed wherever outliers is observed.
- Removed NA values corresponding to ‘*Avgas.CC.Utilization.in.last.12.months*’ variable as it corresponds to less than 1%.
- Used **weight of evidence (WOE)** and **information value (IV)** analysis to get the variables which are strong predictors of dependent variable.
- Balanced the dataset using SMOTE Package in R, since the number of records corresponding to non default customers is very less. Balancing will help us improve the model accuracy.

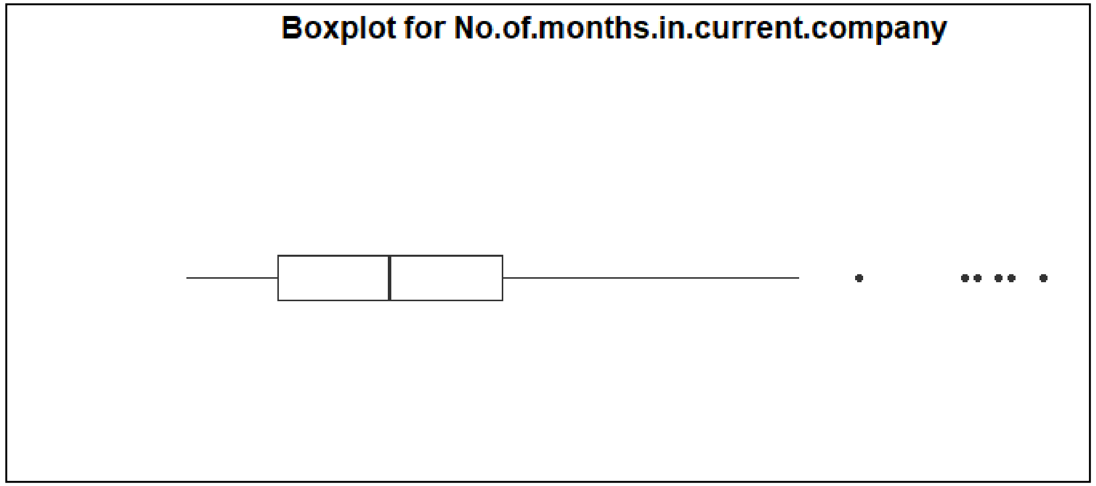
Important Predictor Variables

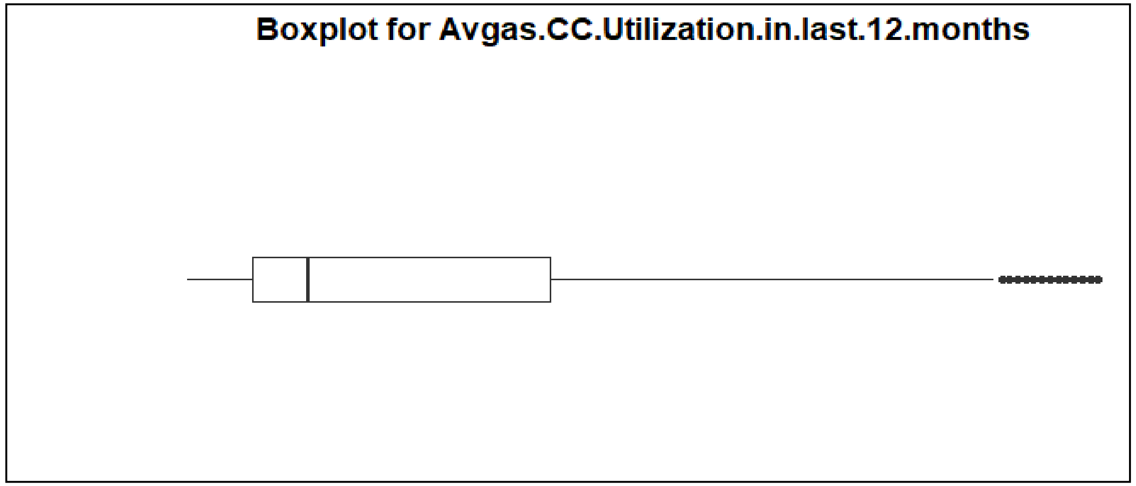
Based on Information Value of each of the variables present in the merged dataset, the following variables are considered as strong predictors of the dependent variable.

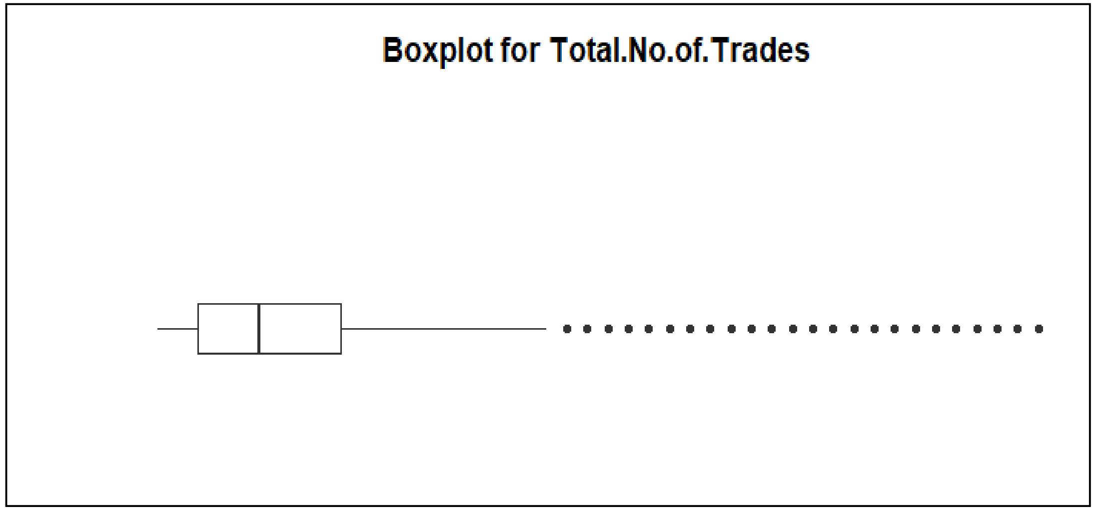
Here we are considering the cutoff as 0.20, which means variables having IV greater than 0.20 is strong predictor variable.

Variable Name	IV
Avgas.CC.Utilization.in.last.12.months	0.31
No.of.trades.opened.in.last.12.months	0.30
No.of.PL.trades.opened.in.last.12.months	0.30
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans	0.30
Outstanding.Balance	0.25
No.of.times.30.DPD.or.worse.in.last.6.months	0.24
Total.No.of.Trades	0.24
No.of.PL.trades.opened.in.last.6.months	0.22
No.of.times.90.DPD.or.worse.in.last.12.months	0.21
No.of.times.60.DPD.or.worse.in.last.6.months	0.21
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.20

Outliers identified using EDA



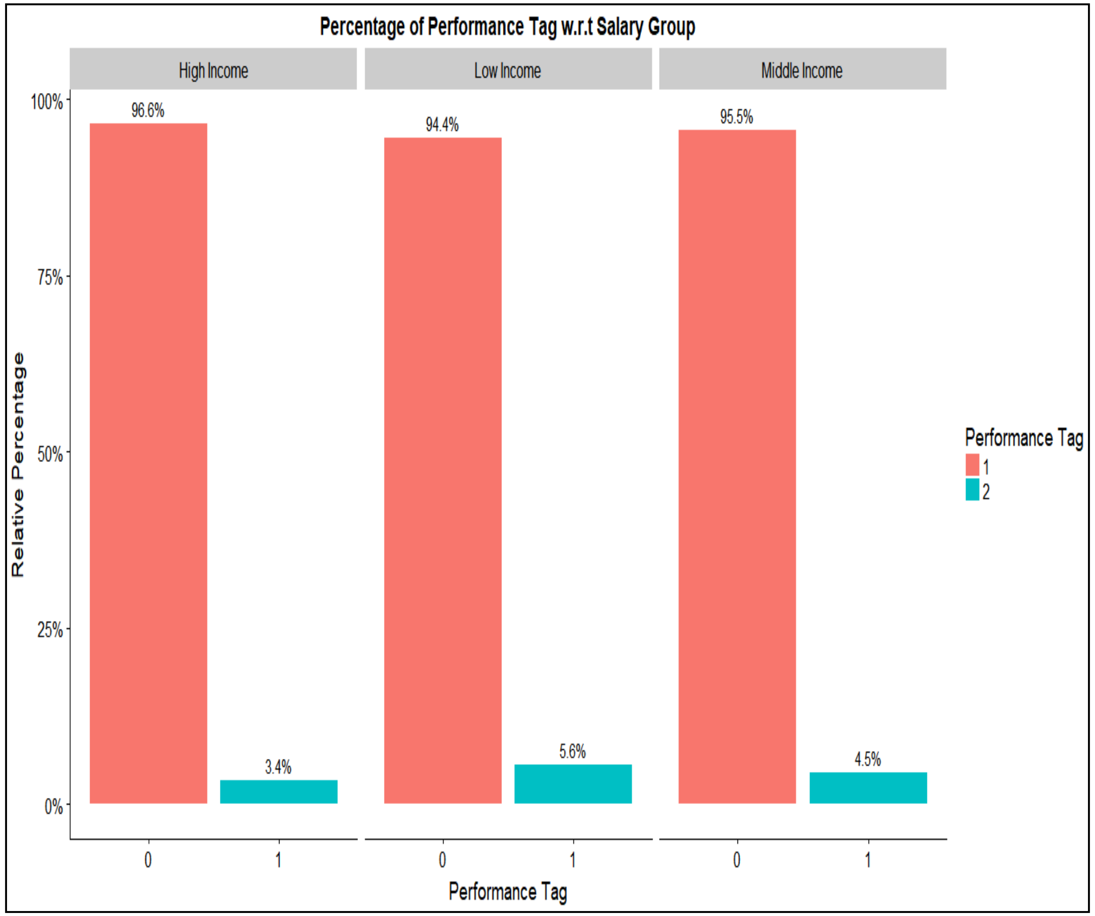


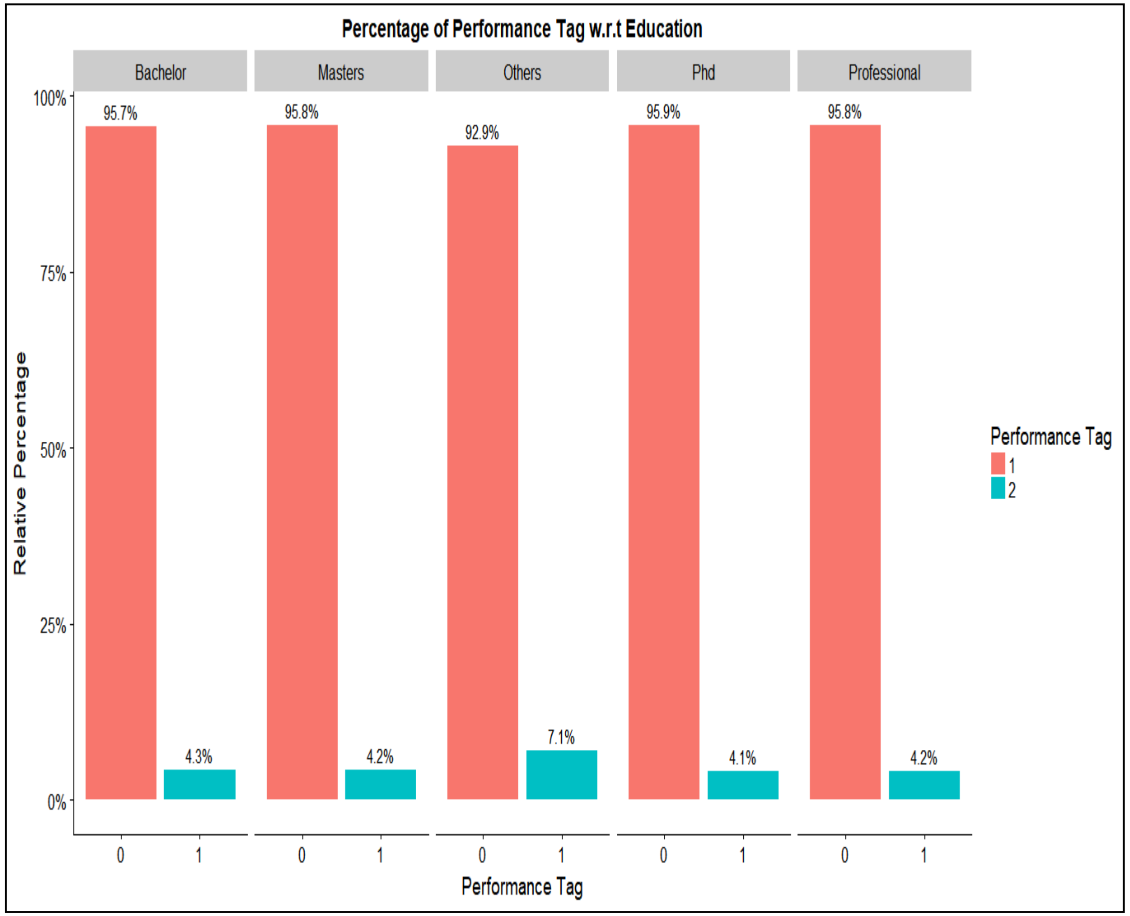


Using EDA, outliers were observed in 3 variables:

- *No.of.months.in.current.company*
- *Avgas.CC.Utilization.in.last.12.months*
- *Total.No.of.Trades*

Some important insights from EDA…

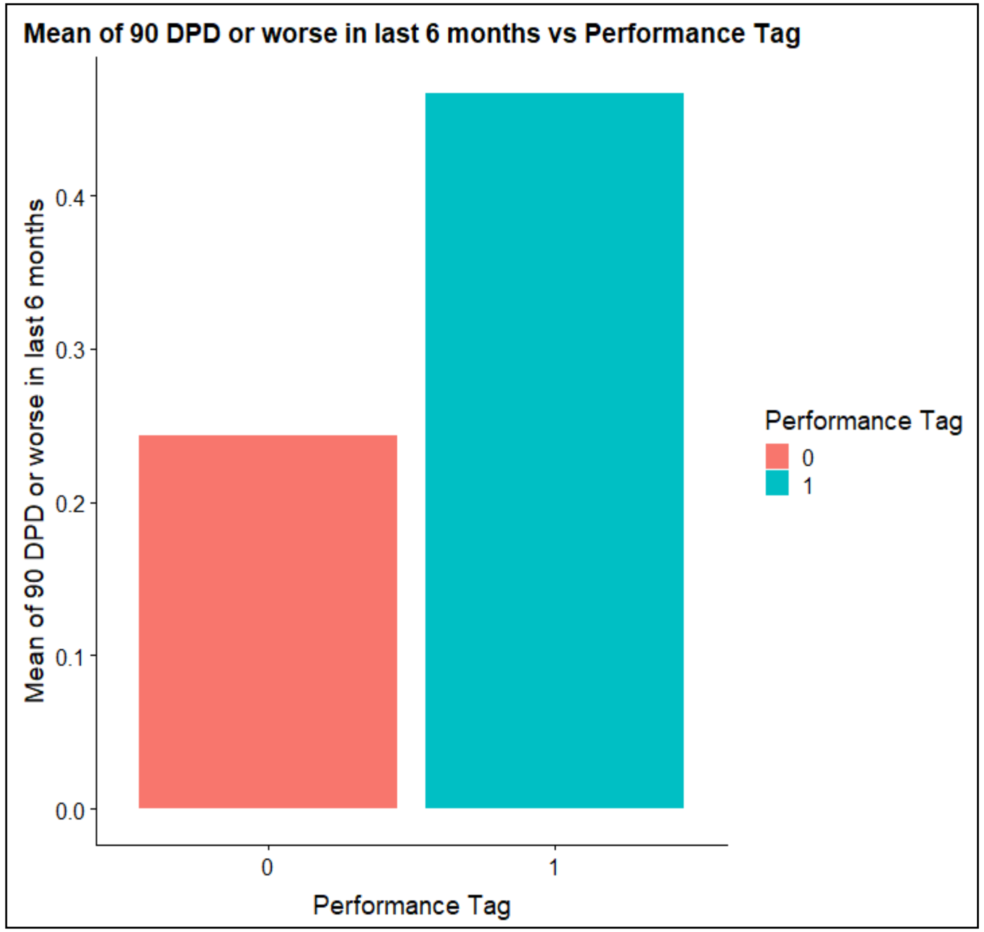


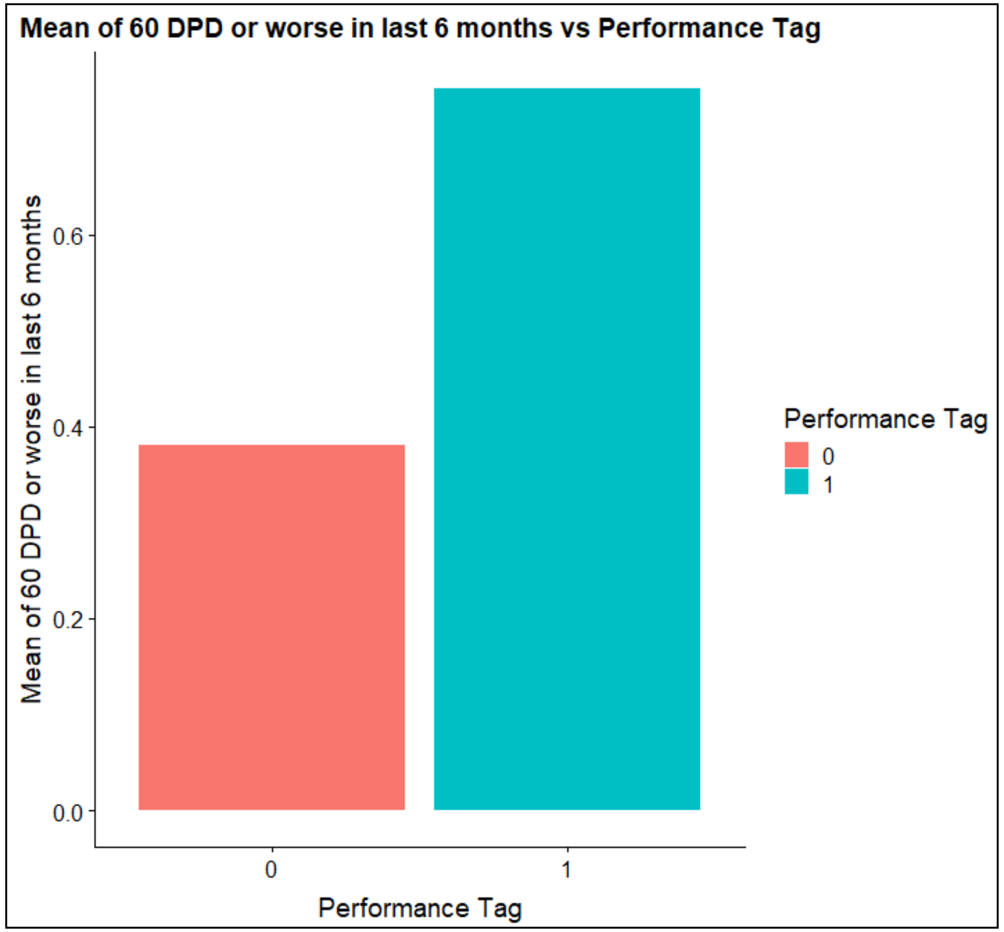


Low Income group people tends to default slightly more in credit cards compared to high or middle income group

The people whose Education details are not available or chose not to disclose education details default credit cards more compared to other group

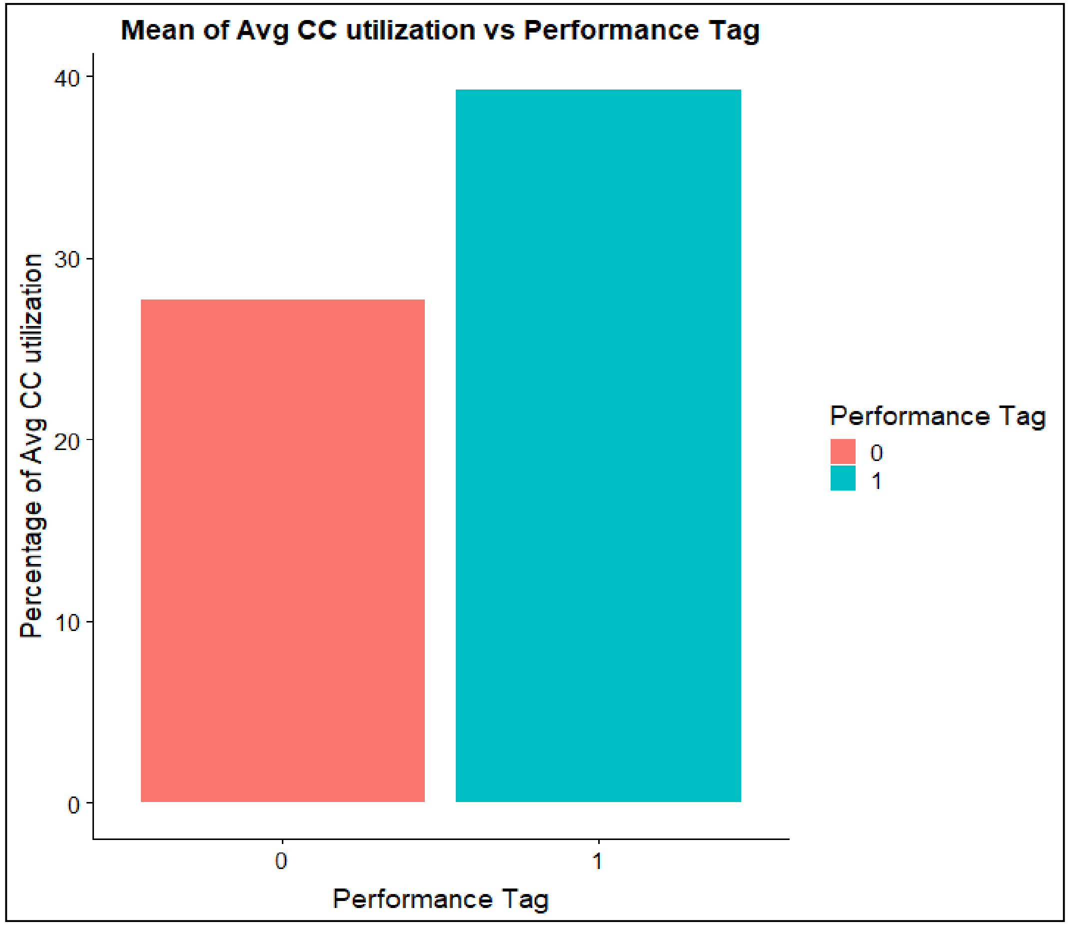
Some important insights from EDA…

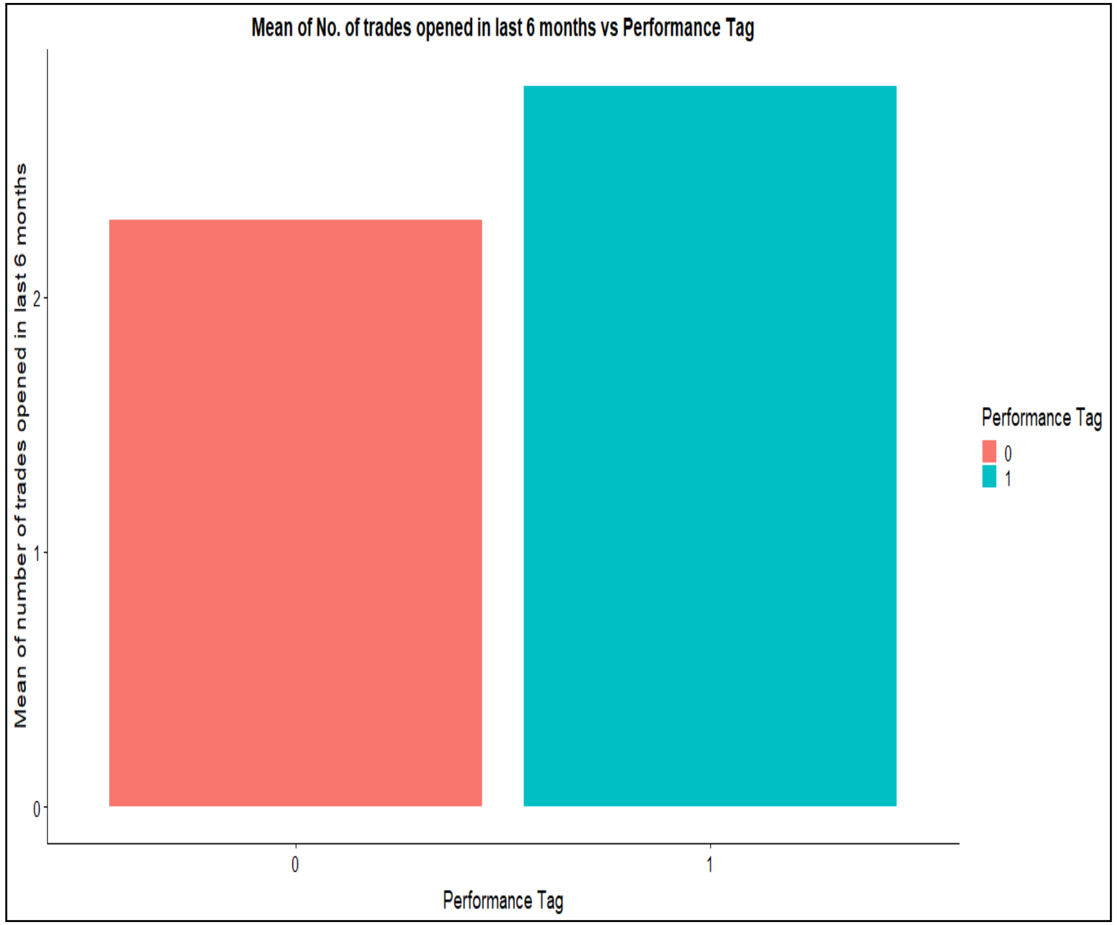




On an average, customers who haven't paid dues since 90 days in last 6 months or customers who haven't paid dues since 60 days in last 6 months tends to default more in their credit card bills.

Some important insights from EDA…

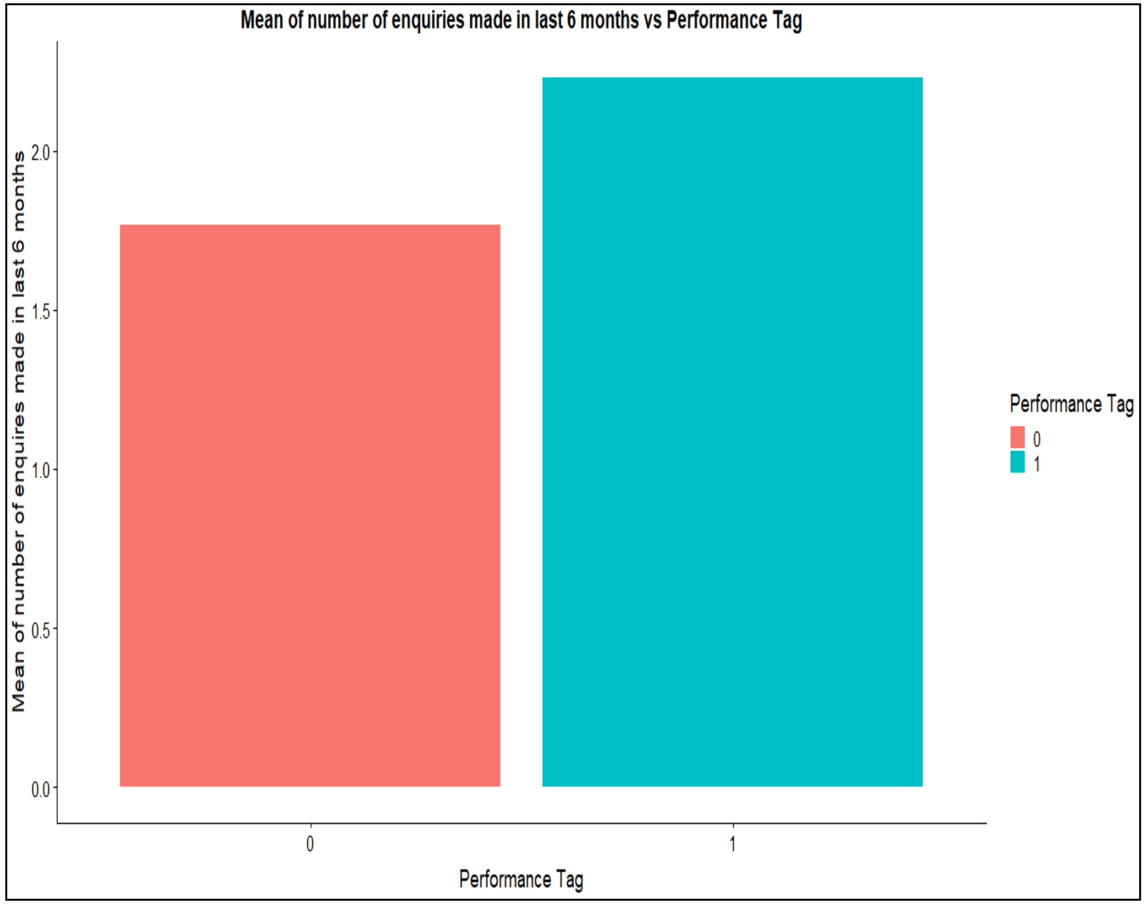


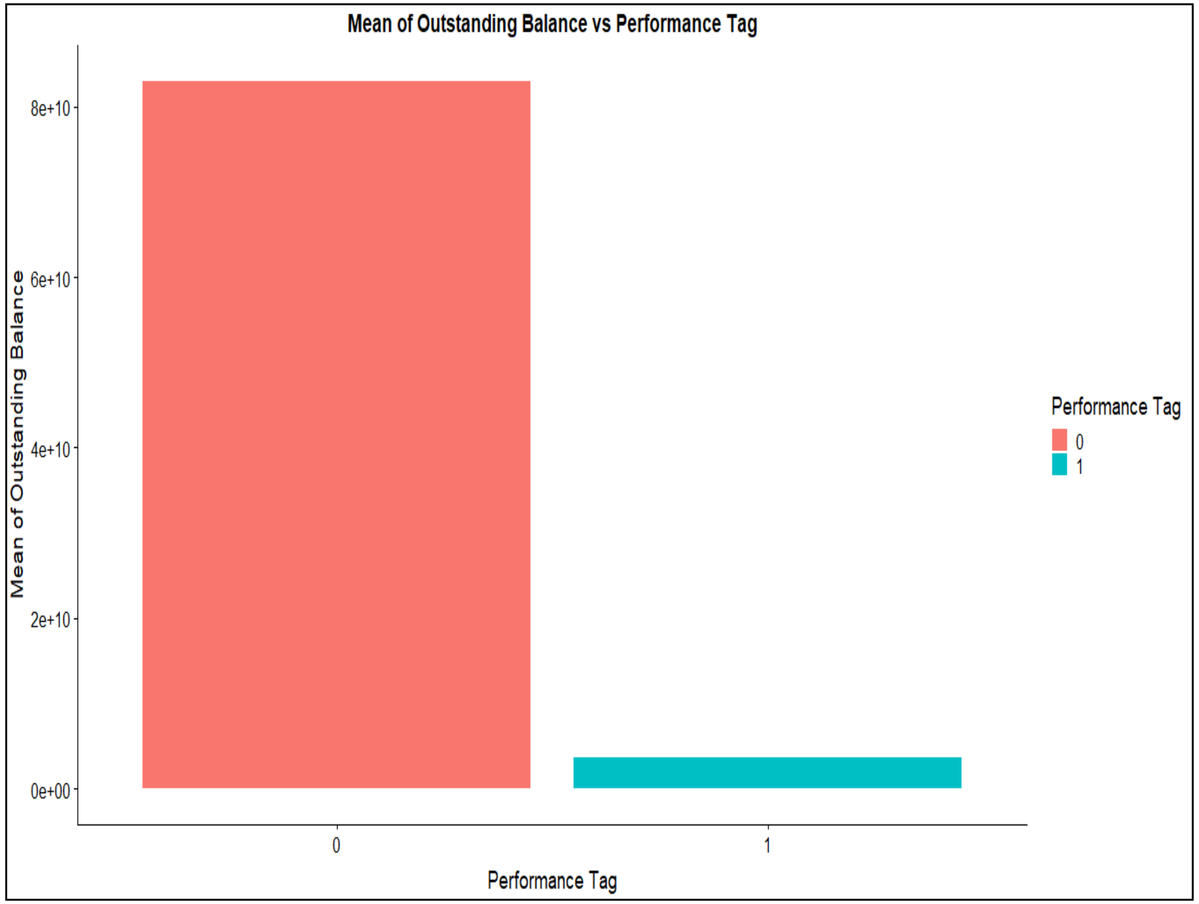


Here we can observe that, on an average defaulted customers tends to use more of credit utilization compared to not defaulted customer

Defaulted customers have opened more number of trades compared to non-defaulted customers

Some important insights from EDA…

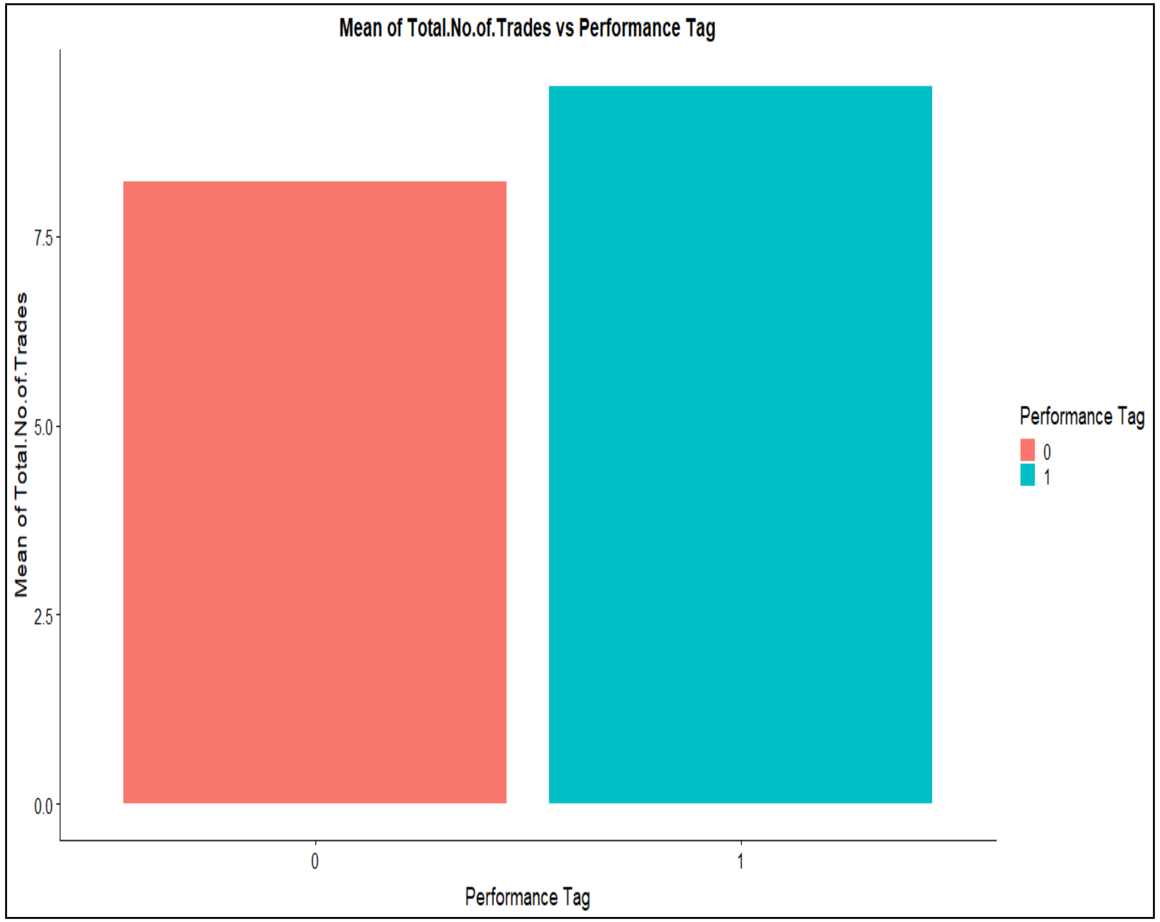




Here we observe that on average number of enquiries made in last 6 months is more for defaulted customers than for non-defaulted ones.

Non defaulted customers have much more average outstanding balance compared to defaulted customers

Some important insights from EDA…



On average, defaulted customers do more number of trades compared to non-defaulted customers.

Based on extensive EDA, we have observed some of the variables which are identified as strong predictor of dependent variable.

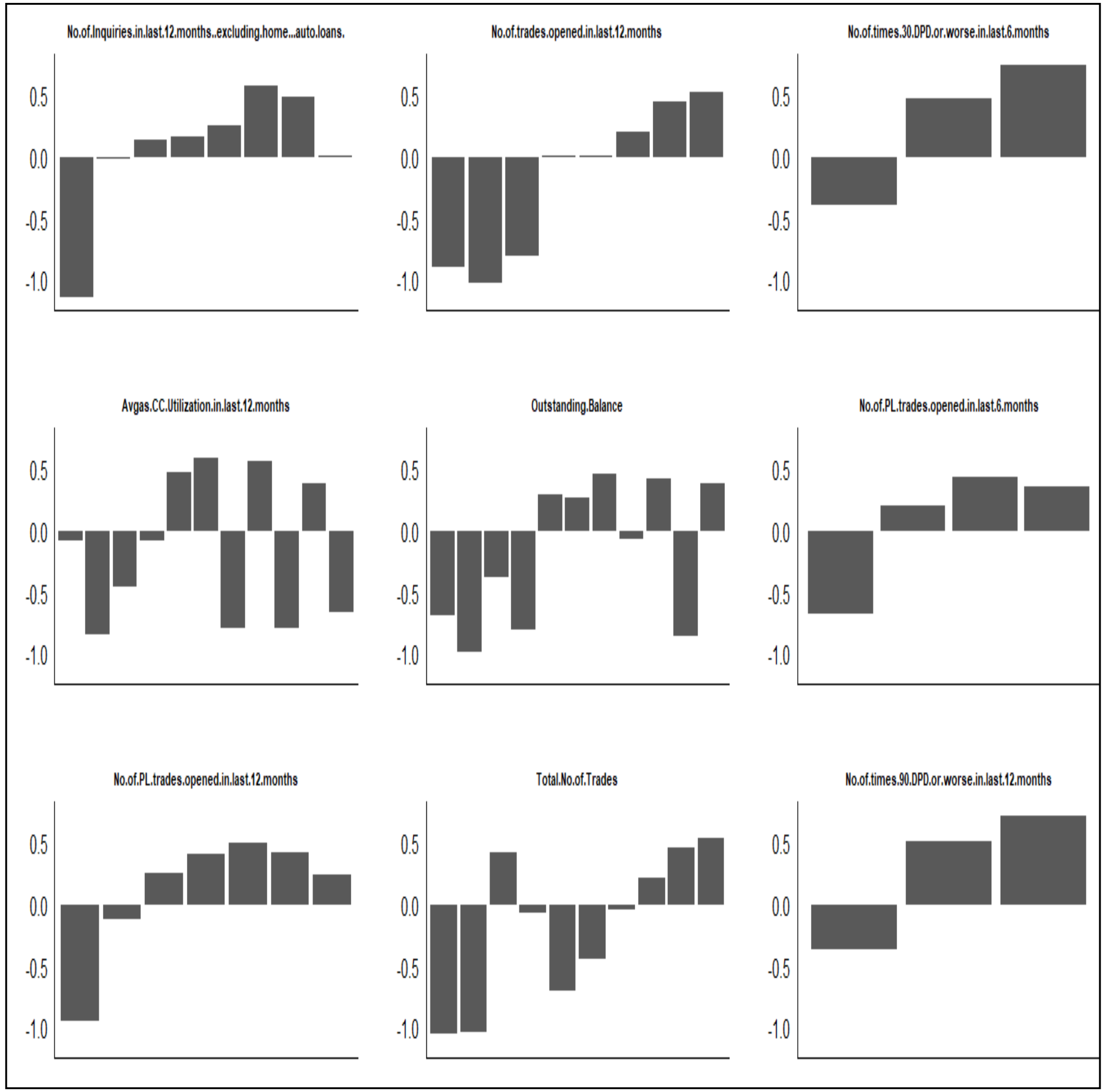
- *Gender*
- *Marital Status*
- *No. of Dependents*
- *Age Group*
- *Salary Group*
- *Education*
- *Profession*
- *Type of Residence*
- *No of times 90 DPD or worse in last 6 months*
- *No of times 60 DPD or worse in last 6 months*
- *No. of Trades*
- *No. of Enquires*
- *Outstanding Balances*

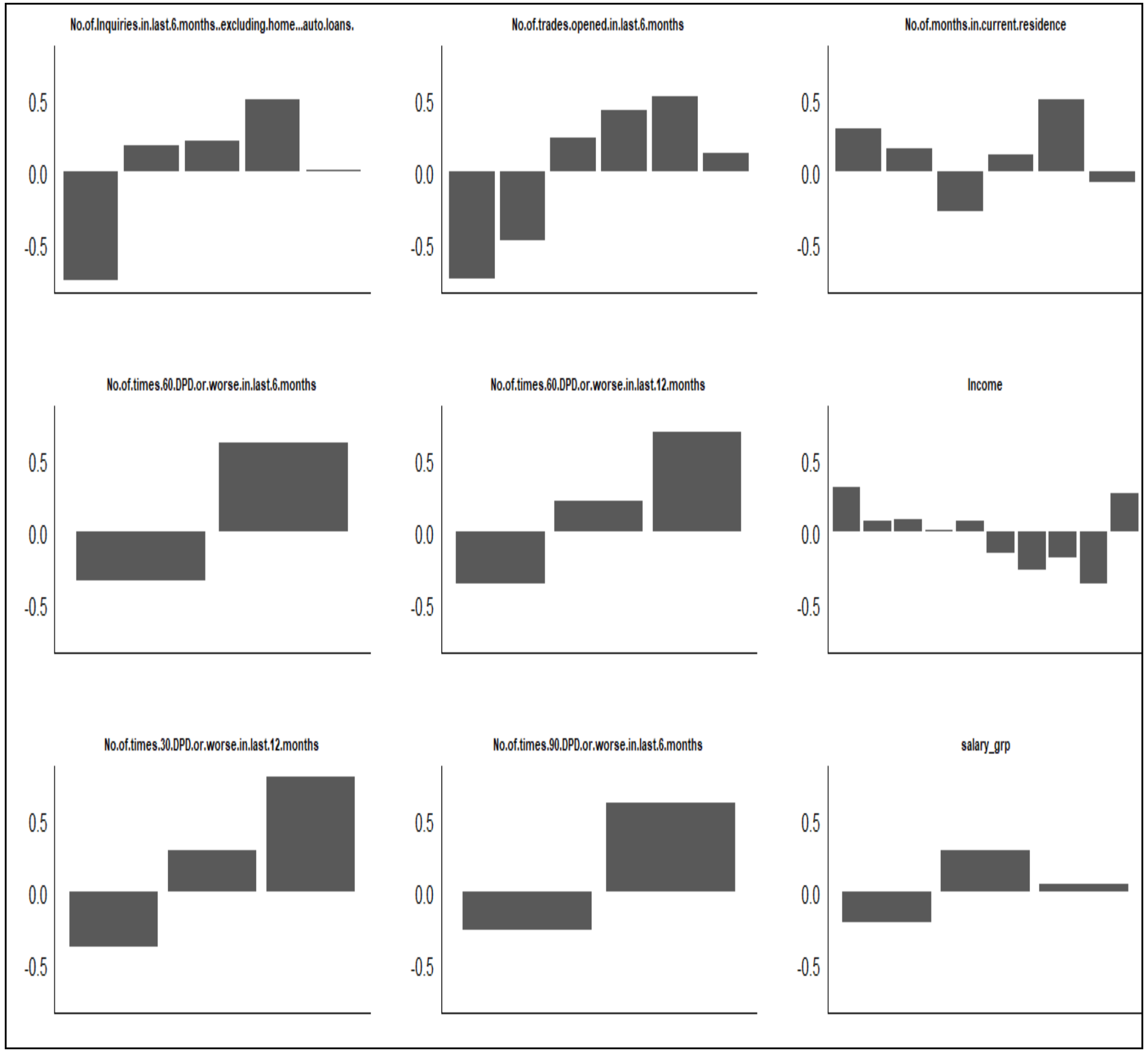
Some important insights from EDA…

Some of the these variables were also identified as strong predictor of the dependent variable based on the IV values.

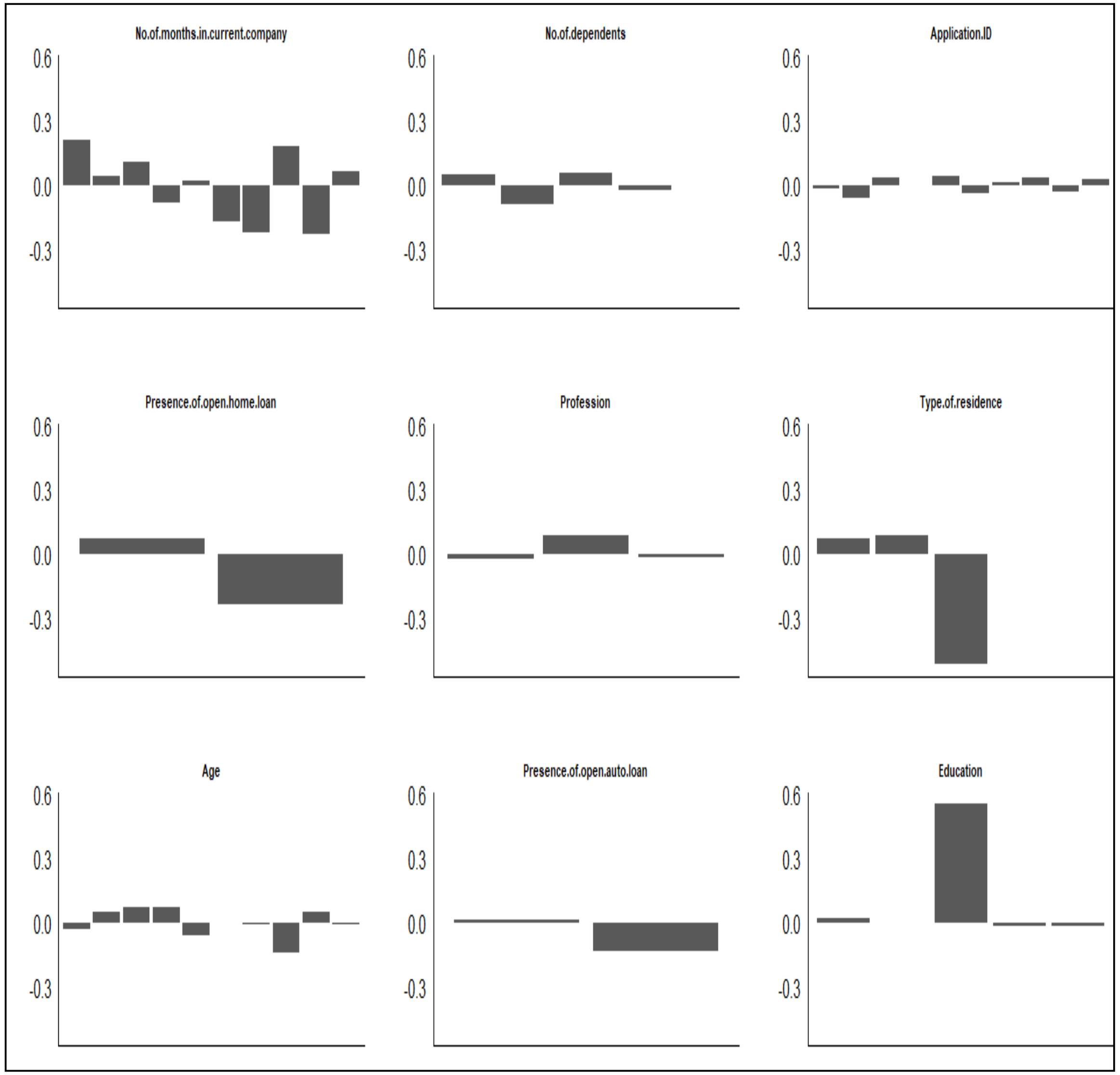
- *Salary Group*
- *Education*
- *No of times 90 DPD or worse in last 6 months*
- *No of times 60 DPD or worse in last 6 months*
- *Avgas CC Utilization in last 12 months*
- *No of trades opened in last 6 months*
- *No of Inquiries in last 6 months*
- *Outstanding Balance*
- *Total No of Trades*

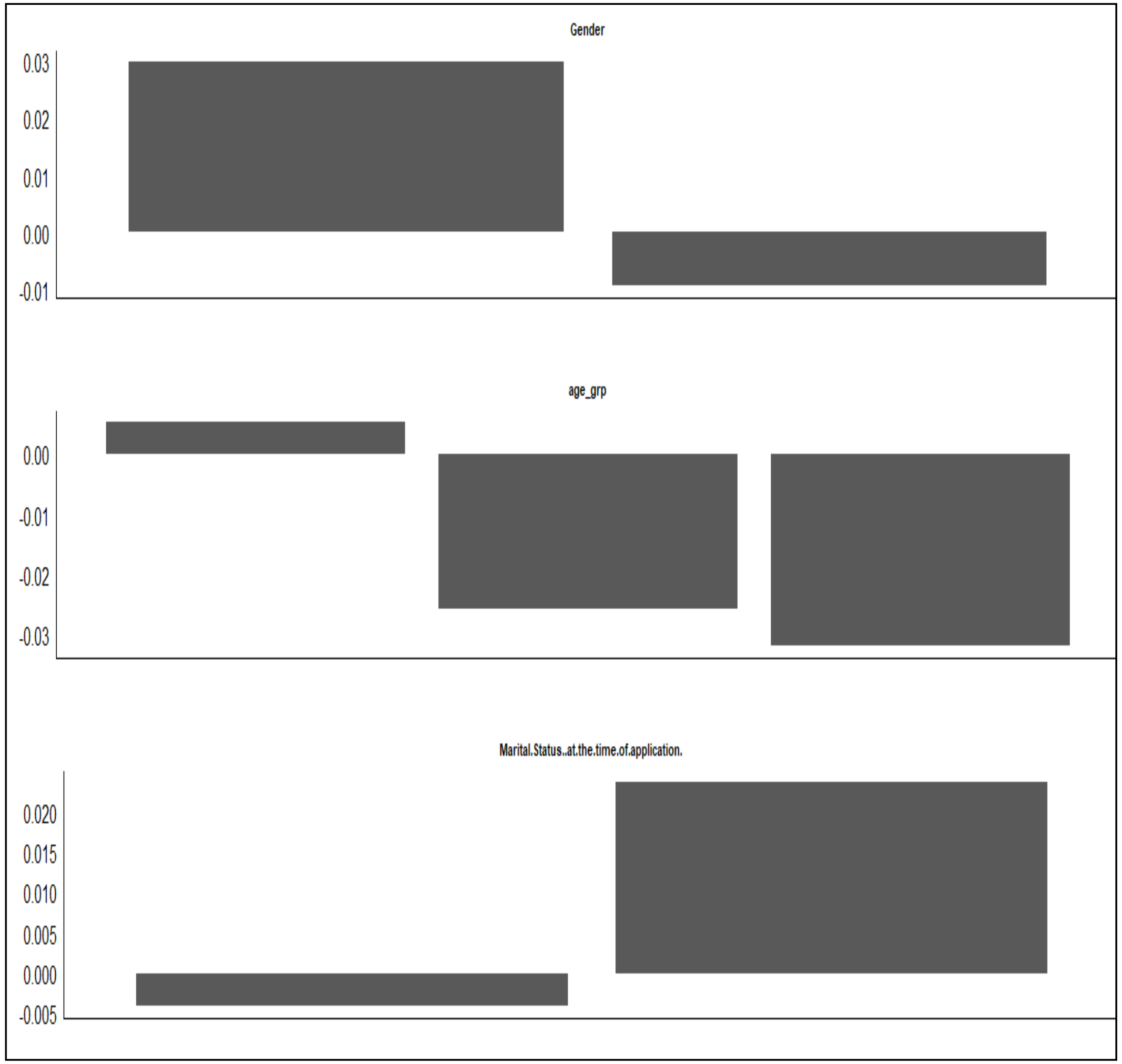
WOE Plots


No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
No.of.trades.opened.in.last.12.months
No.of.times.30.DPD.or.worse.in.last.6.months
Avgas.CC.Utilization.in.last.12.months
Outstanding.Balance
No.of.PL.trades.opened.in.last.6.months
No.of.PL.trades.opened.in.last.12.months
Total.No.of.Trades
No.of.times.90.DPD.or.worse.in.last.12.months
0.5
0.0
-0.5
-1.0


No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
No.of.trades.opened.in.last.6.months
No.of.months.in.current.residence
No.of.times.60.DPD.or.worse.in.last.6.months
No.of.times.60.DPD.or.worse.in.last.12.months
Income
No.of.times.30.DPD.or.worse.in.last.12.months
No.of.times.90.DPD.or.worse.in.last.6.months
salary_grp
0.5
0.0
-0.5

WOE Plots

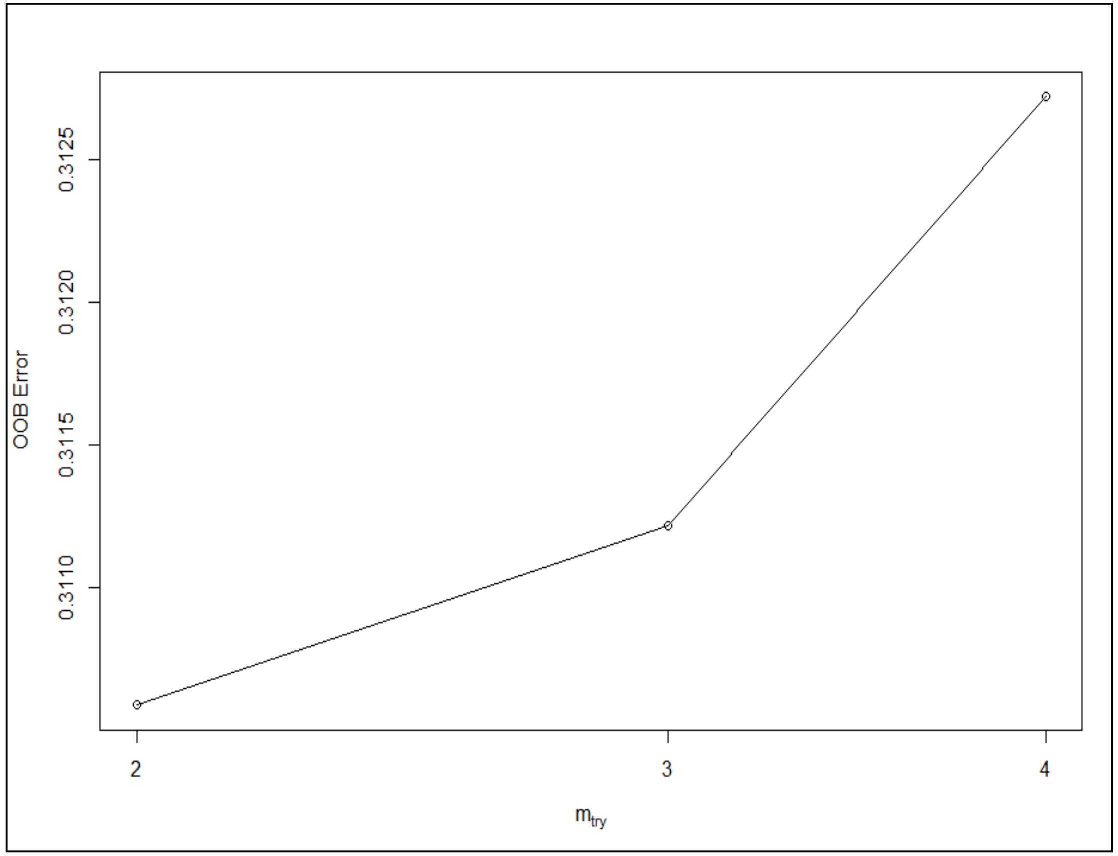

No.of.months.in.current.company
No.of.dependents
Application.ID
Presence.of.open.home.loan
Profession
Type.of.residence
Age
Presence.of.open.auto.loan
Education
0.6
0.3
0.0
-0.3


Gender
0.03
0.02
0.01
0.00
-0.01
age_grp
0.00
-0.01
-0.02
-0.03
Marital.Status..at.the.time.of.application.
0.020
0.015
0.010
0.005
0.000
-0.005

Model Building Procedures

- In the model building process , we have used combined/merged data set and demographic data set.
- Applied SMOTE function to balanced the data sets.
- Used 3 different algorithms (Logistics Regression, Decision Tree and Random Forest) to find out the highest accuracy.
- Built models on balanced data set as well as on unbalanced data set.
- In the next slides we are going to provide the details about final model on two data sets.

Final Model: Random Forest on Demographic Data

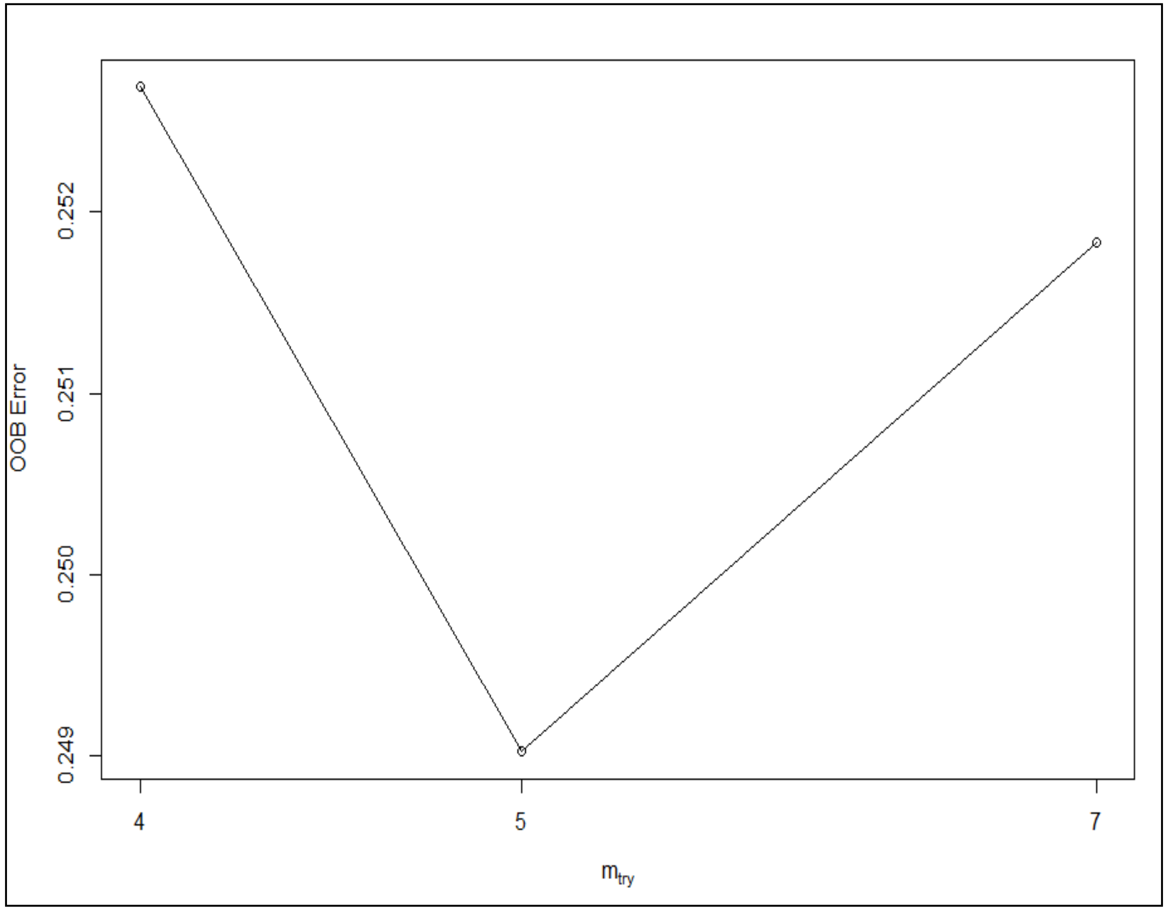


- After applying different algorithms on balanced and unbalanced data sets, we have identified comparatively optimized model evaluation statistics using Random Forest on balanced demographic data.
- Using 200 “ntree” and 2 “mtry” model has been derived and predicted the below reference data.

Prediction	Reference 0	Reference 1
0	14794	582
1	4869	316

- Accuracy : 73.49%
- Sensitivity: 75.23%
- Specificity: 35.18%

Final Model: Random Forest on Combined/Merged Data

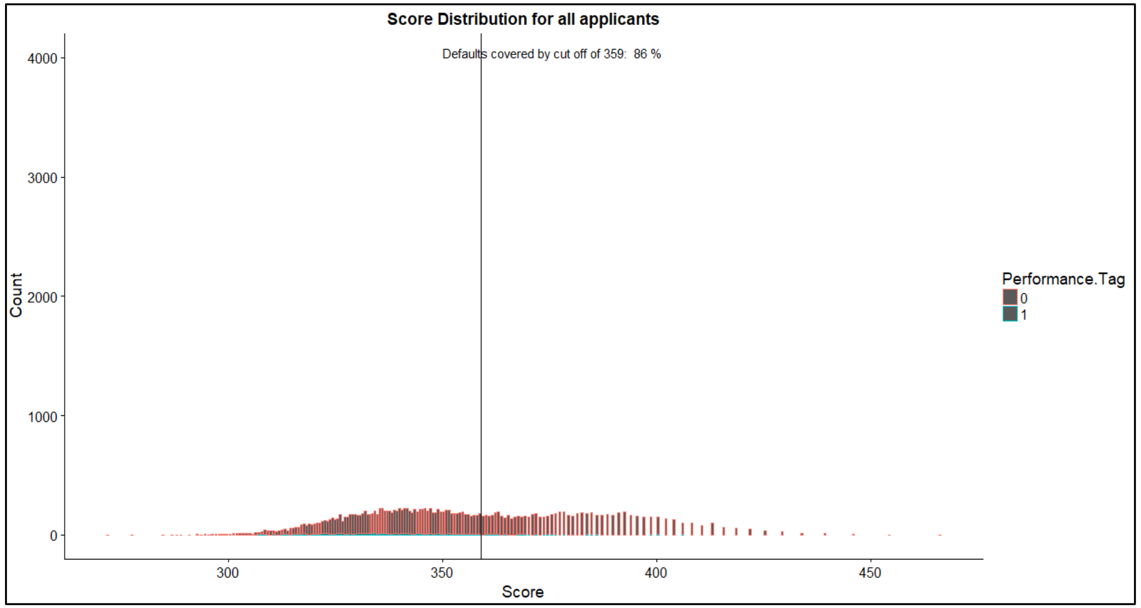


- After applying different algorithms on balanced and unbalanced data sets, we have identified comparatively optimized model evaluation statistics using Random Forest on balanced combined/merged data.
- Using 200 "ntree" and 5 "mtry" model has been derived and predicted the below reference data.

Prediction	Reference	0	1
	0	15419	524
	1	4297	321

- Accuracy : 76.55%
- Sensitivity: 78.20%
- Specificity: 37.98%

Application Scorecard



Percentage	Score
0	201
20	308
40	327
60	340
80	359
100	466

- From the application scorecard analysis , it is identified that the scorecard cut off value is set to 359.
- The above cut off score value needs to be used for financial analysis to identify the approval rate and net credit loss.
- The application scorecard is used to find that 3% of originally rejected applicants could be provided with credit card.

Financial Analysis

- As per Profit and Loss perspective, the objective is to minimize the net credit loss.
- Application scorecard is used for determining desired trade off between risk level and approval rate.
- The Balanced cut-off score (359) needs to be used as strict benchmark to determine the eligibility of any application whether it will be accepted or rejected.
- Using the mentioned models and cut-off score, management can re-evaluate the applications.
- With suggested cut-off score of 359 , more than 80% of applicants would be approved.

Conclusions

Based on the previous analysis , we can conclude the following things.

1. If Higher management will follow this final model , then will receive the optimised results.
2. Provided a clear picture of potential financial benefit from P&L perspective.
3. During creation of model the key points has been kept in mind
   - Implications of using this model for auto approvals and rejections process.
   - How to avoid Potential credit loss using this model.
   - List of assumptions on which model has been built.