CredX Capstone Project

A comprehensive Data Driven Solution to credit risk
PG Diploma (Data Science), IIIT Bangalore &
UpGrad ‘ 2019

Anindya Chakraborty
Jahnavi Chintakunta
Aradhana Tripathi
Goushalya Kothai Nachiar

Business Objectives

CredX, a leading credit card provider, wants to mitigate credit risk by acquiring right customers

This advanced 'Data Driven' solution tries to answer the most critical business problems of Credit Loss by analysing bank's past applicant's Data Models driven by Machine Learning.
This study also determines the factors affecting credit risk and suggests a Application Scorecard model which is financially viable by the bank.

Data Understanding

- There are two datasets available for the analysis. One Demographic data for the Bank's customers and second is Credit Bureau Data
- Demographic Data is sources at the time of customer applying for the credit card.The dataset contains customer level information such as Age, Gender, Marital Status, Income
- Credit Bureau data contains customer's credit history such as defaults in the past, non payments of dues, Home Loan or Auto Loan & Performance Tag. The performance tag is given to a customer which have defaulted (Performance Tag = 1)
- Records where there were no Performance Tag present are the applicant's whose application was rejected by the bank
- Both the datasets are highly skewed (only 4% of defaults)

Data Sources, Cleaning and Outlier Treatment

- Credit Bureau and Customer's Demographic Data is used to built the predictive models.
- Age column has negative age entries and less than 18 years were replaced by 18 as minimum age to apply for credit card is 18
- Duplicate Application IDs were removed. NA records upto 3% were removed
- Rejected Application ID is stored is a separate dataset.
- Outlier Treatment was done with capping done at appropriate percentile value for continuous variables
- Column names were changed for certain functions in R to work

Analysis Flow

Data Preparation
Merge Demographic and Credit Bureau Data on Application ID

Remove duplicate rows

Check NA/NAN and outlier treatment

Information Value & WOE Analysis
Computation of Information Value

Certain columns are removed from further analysis based on low IV Value (Cutoff = 0.02)

Imputation of dataset with the WOE values

Model Building
Built the model on training dataset

SMOTE sampling technique is used for balancing the dataset

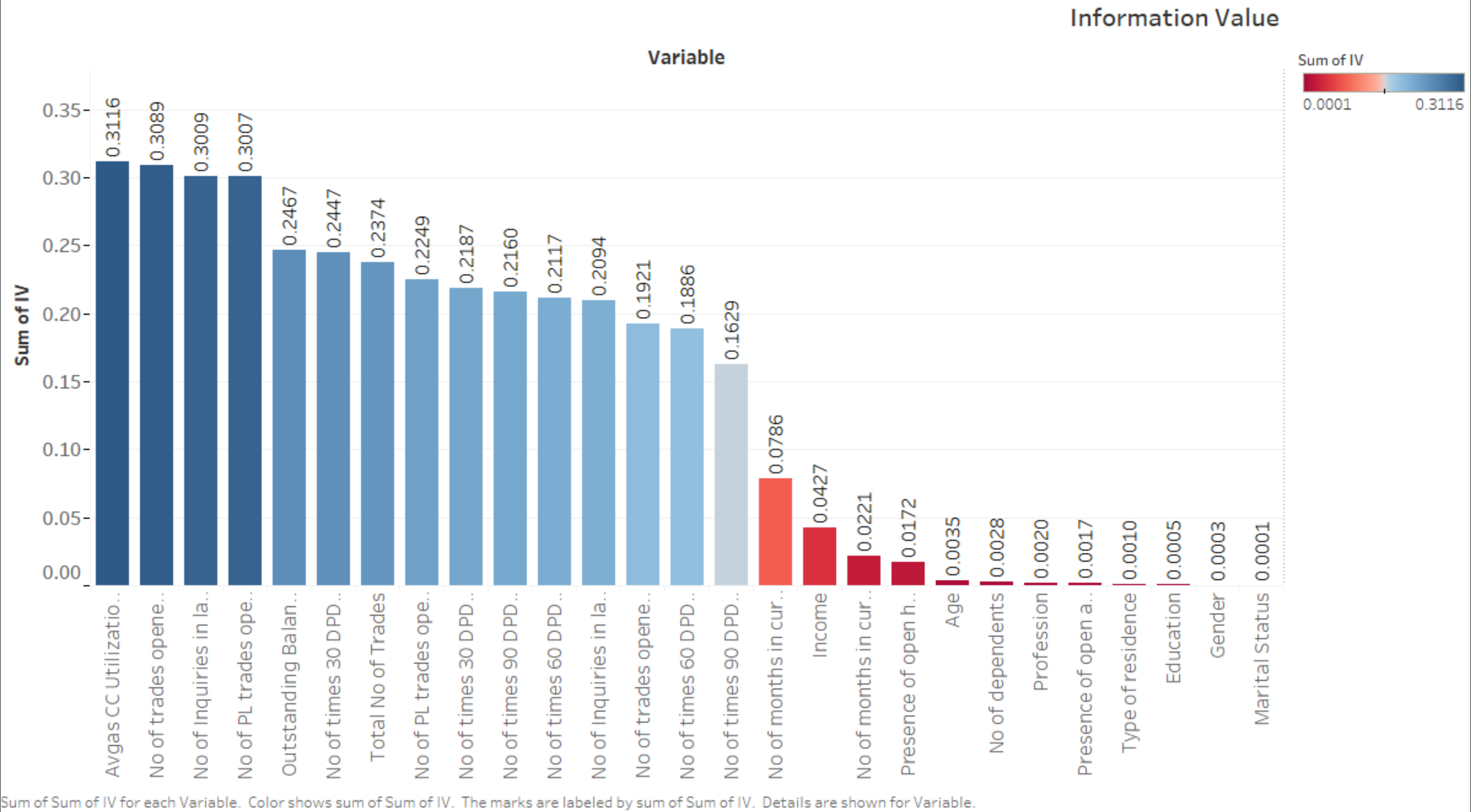
Logistic Regression, Decision Tree & Random Forest Models are built

Model Evaluation
Model is evaluated on test data

Confusion Matrix Evaluation

Application Score card is built from Random Forest model

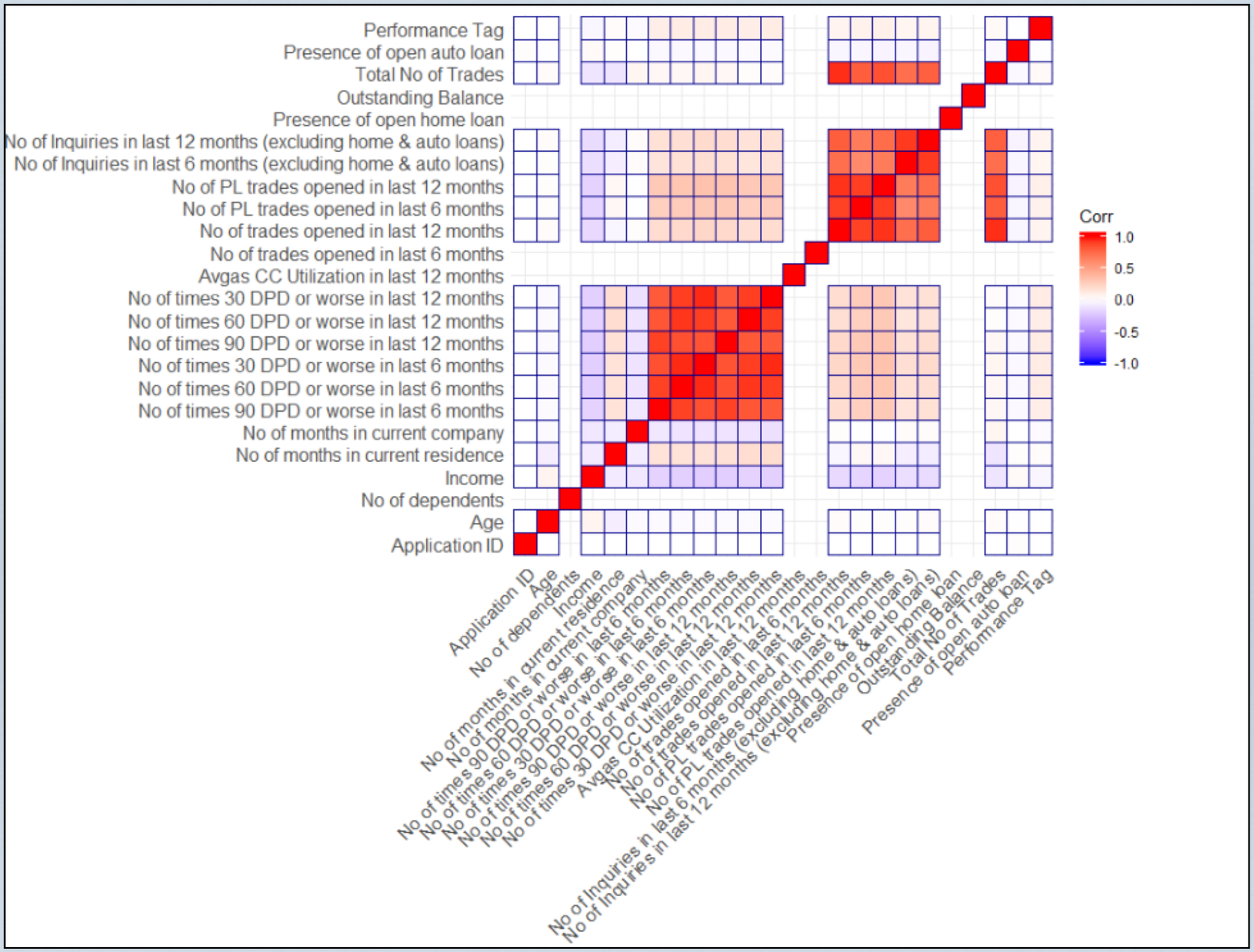
Information Value



Sum of Sum of IV for each Variable. Color shows sum of Sum of IV. The marks are labeled by sum of Sum of IV. Details are shown for Variable.

Tableau Plot

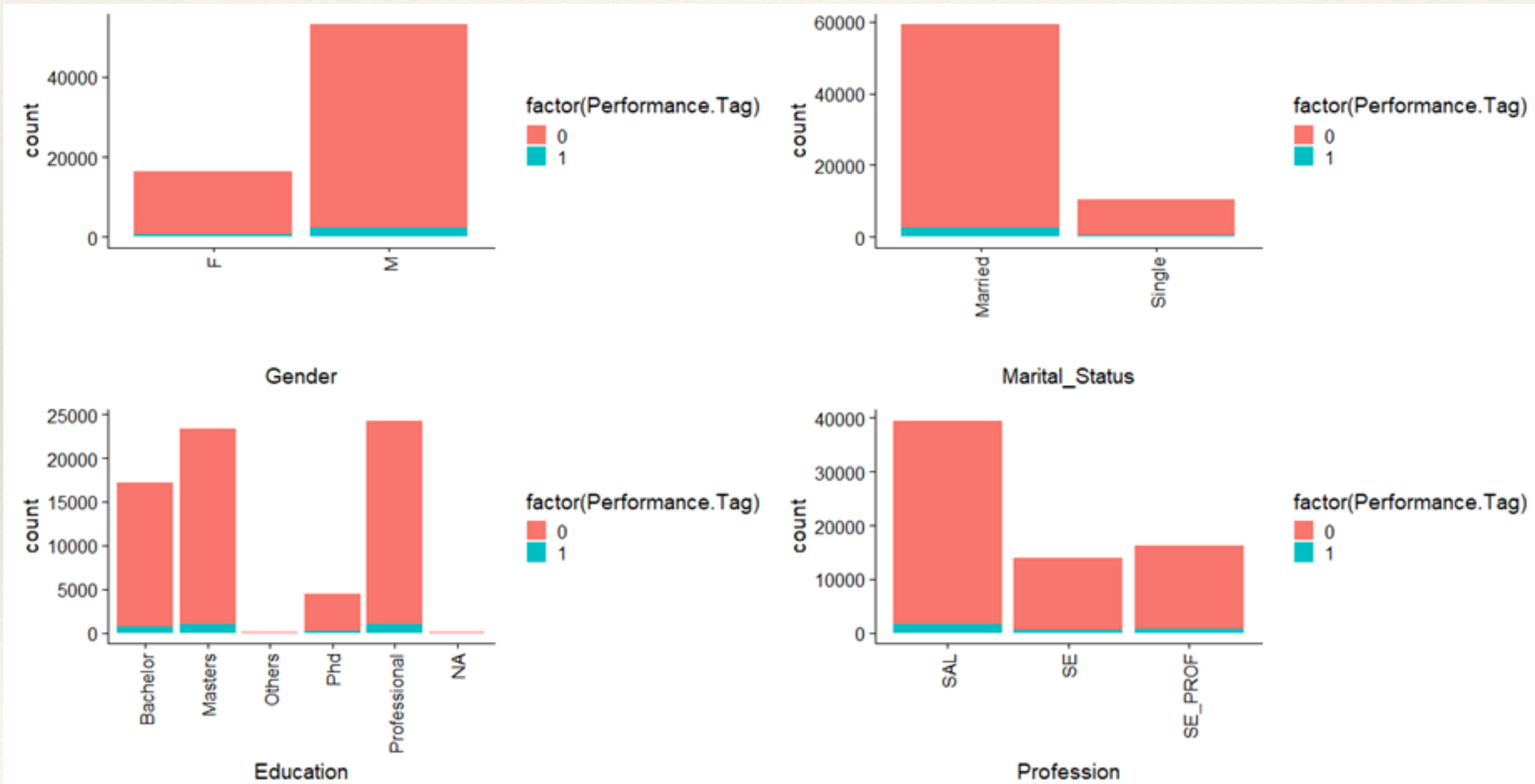
Correlation Plot



- The correlation plot suggests the correlation between the variables of the dataset. The highly correlated variables are dealt by Variance Inflation Factor Analysis

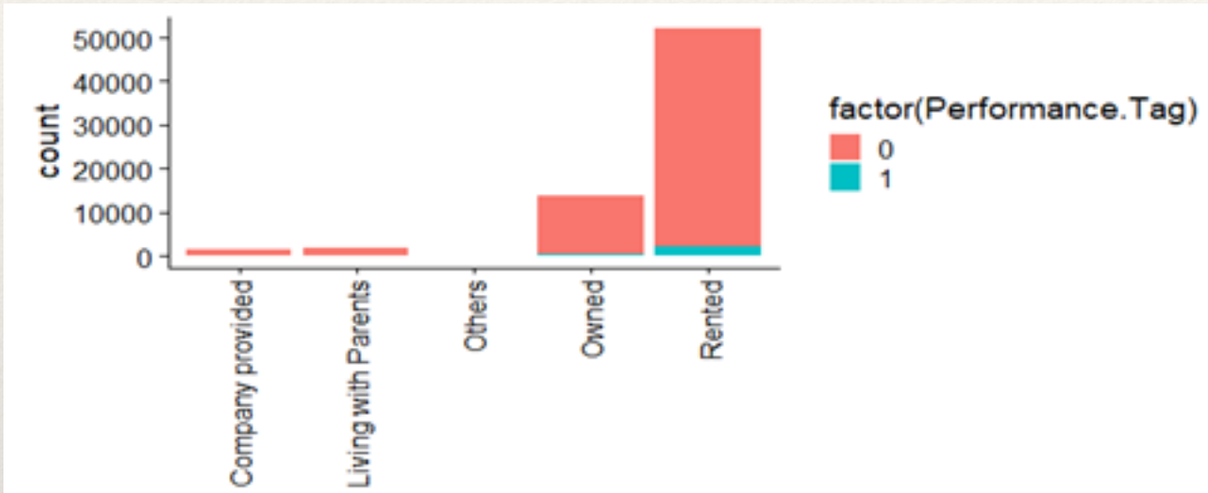
Data Insights from Exploratory Data Analysis

Grid Plot of Categorical Variables with Performance Tag

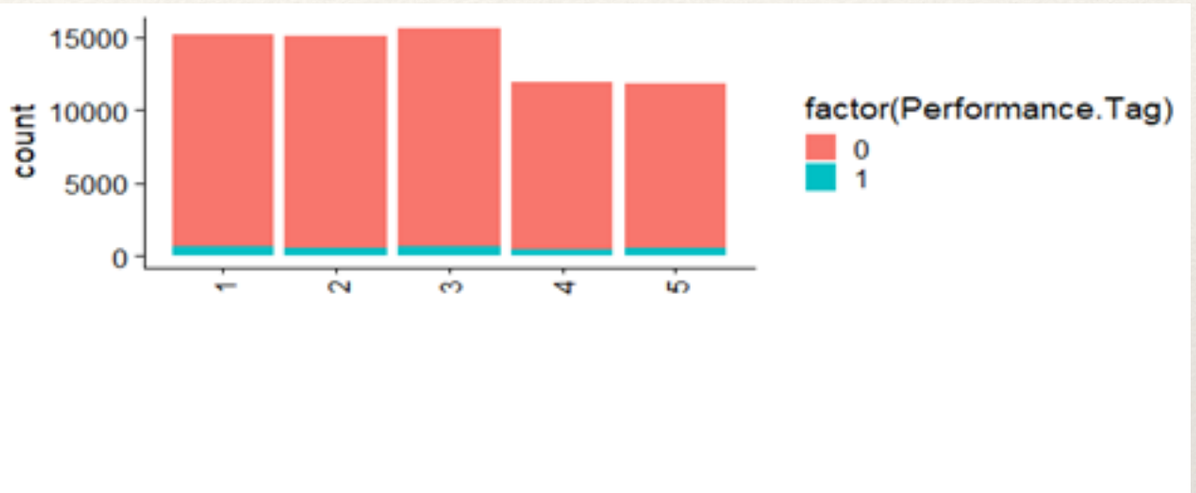


EDA

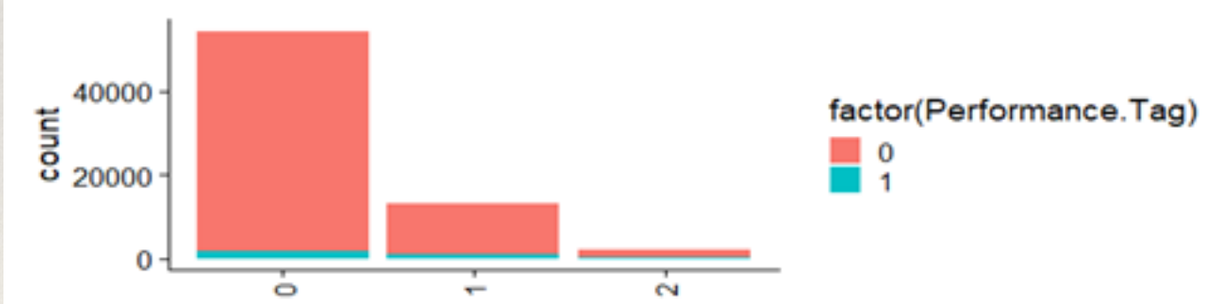
Type of Residence, No of Dependents vs Performance Tag



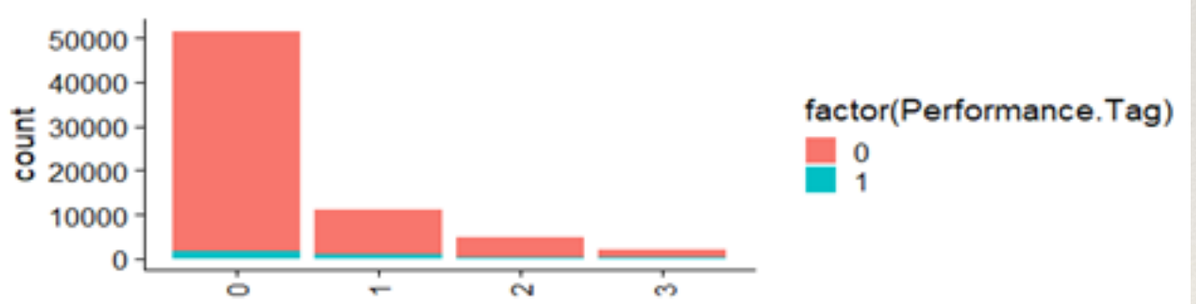
`Type of residence`



`No of dependents`



`No of times 90 DPD or worse in last 6 months`



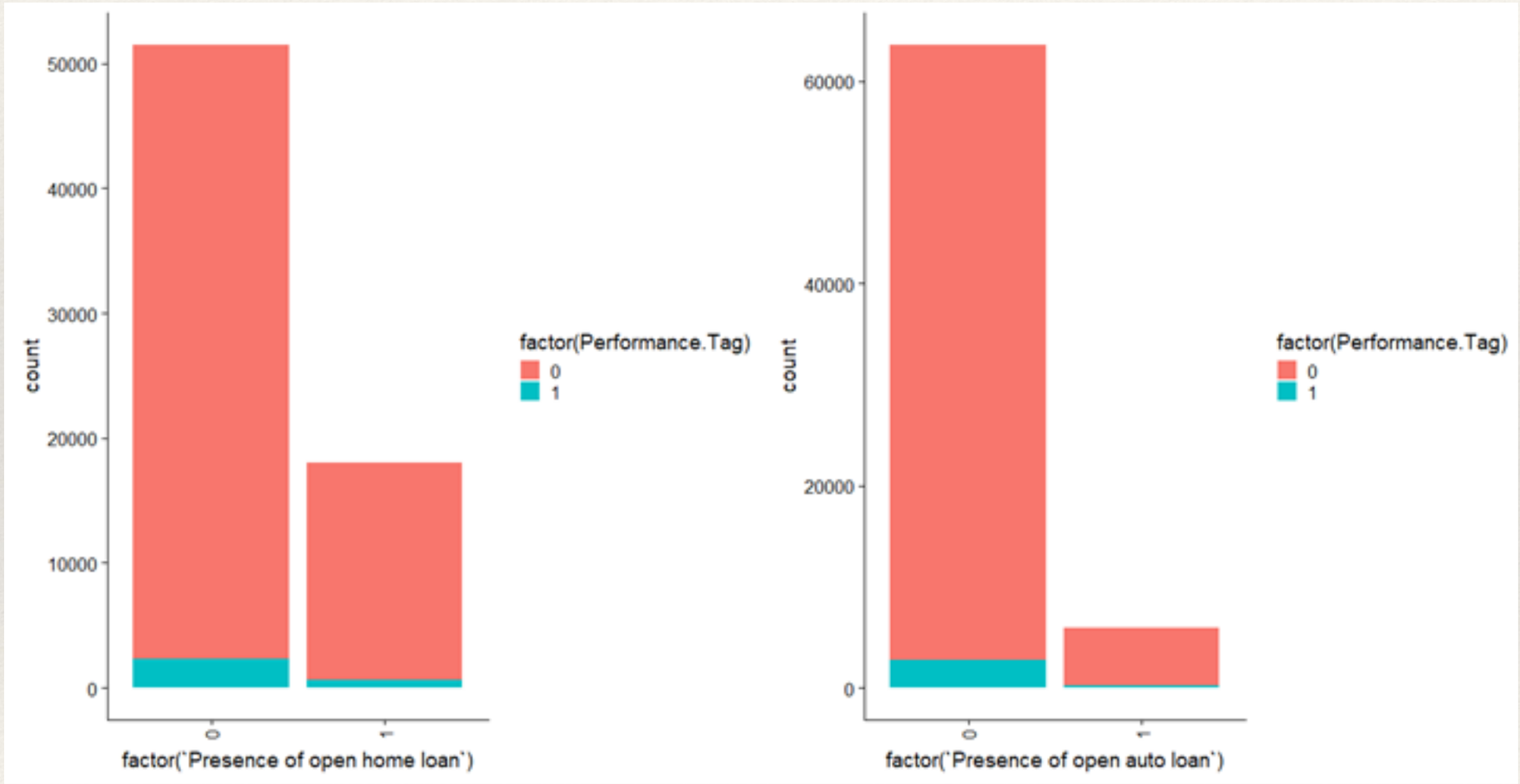
`No of times 60 DPD or worse in last 6 months`

EDA

Days Past Due vs Performance Tag

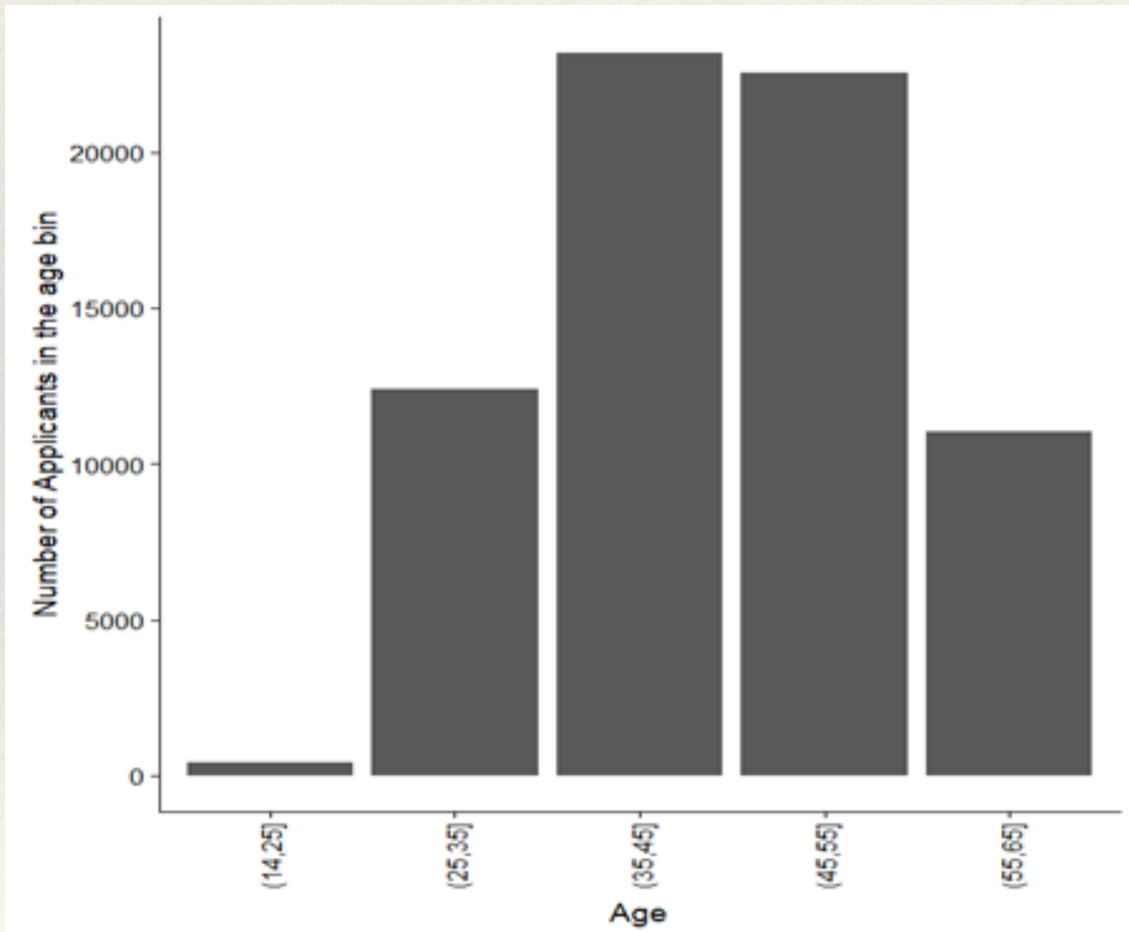
EDA

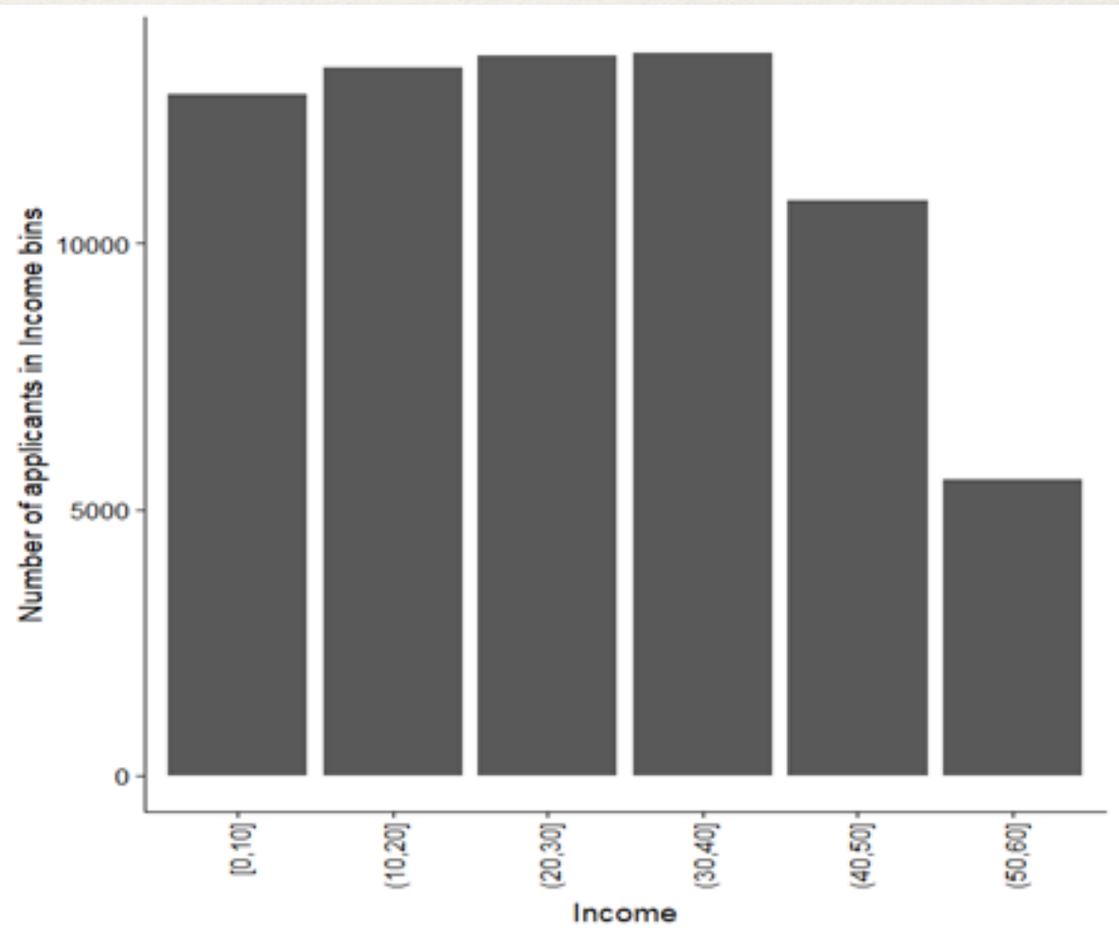
Open Home Loan. Auto Loan vs Performance Tag



EDA

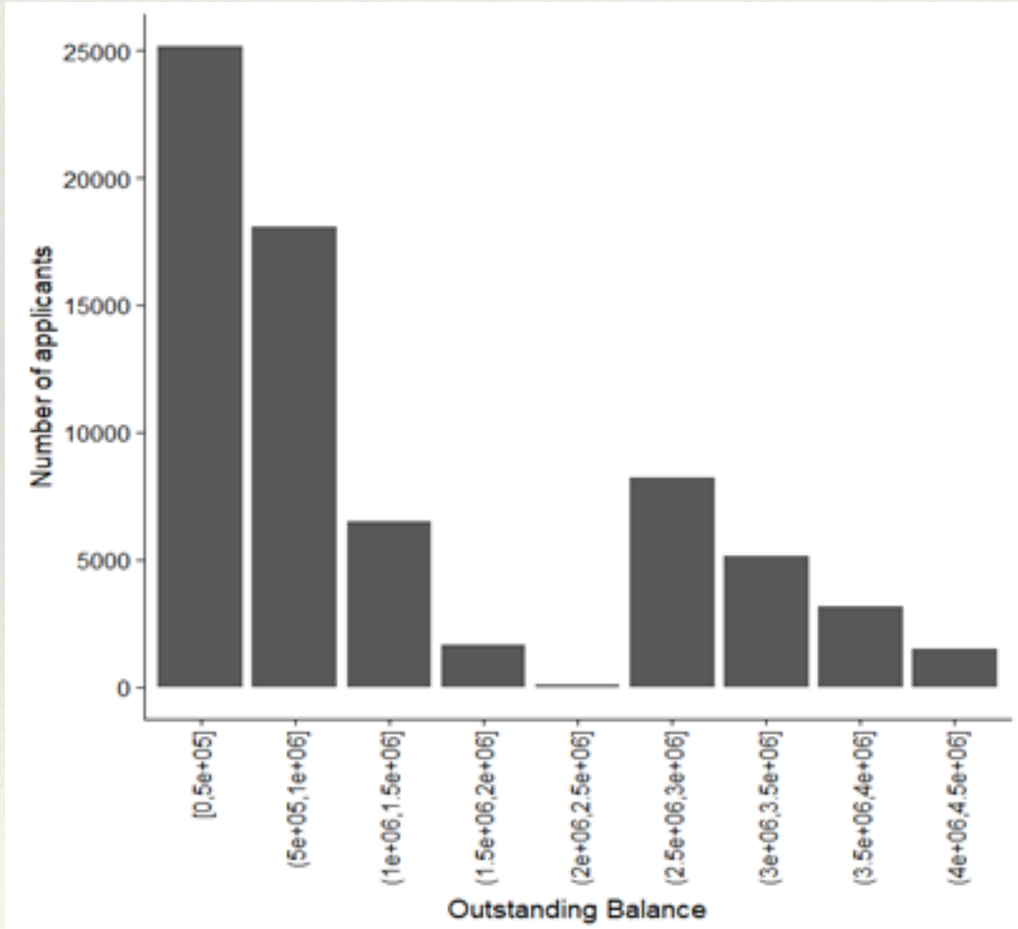
Age Distribution shows that most of applicants are in age group of 35 - 55

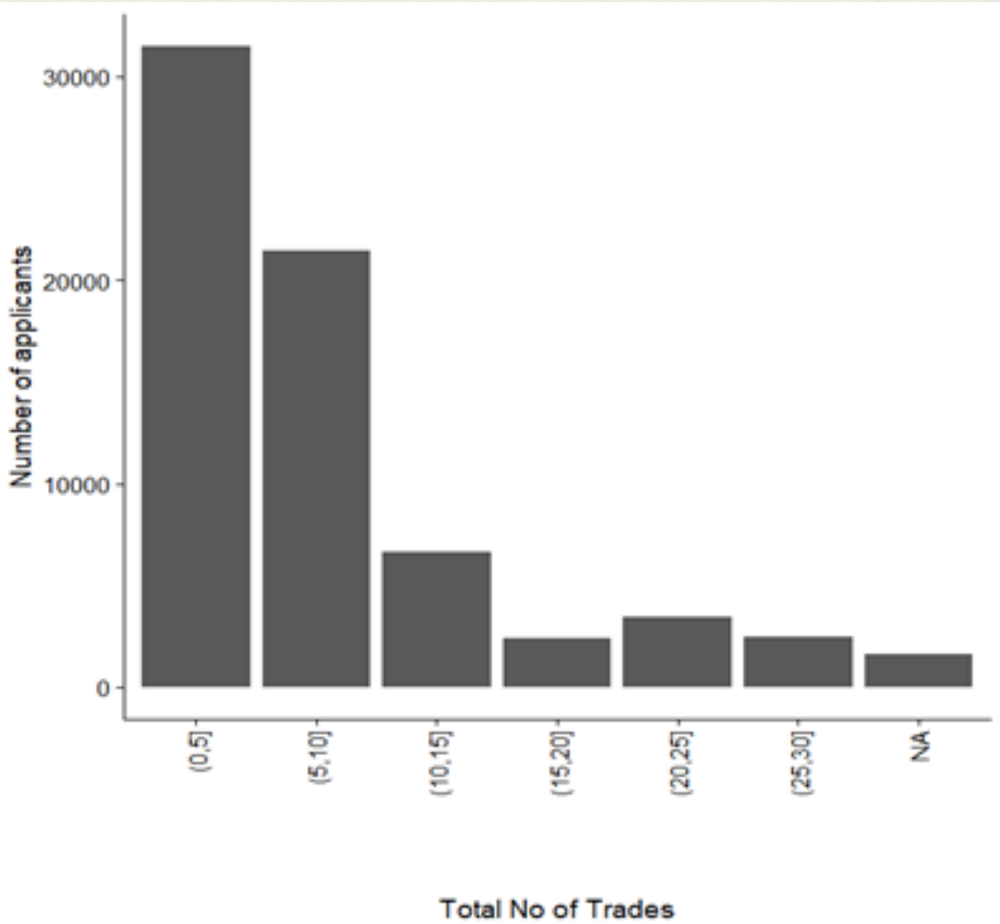




EDA

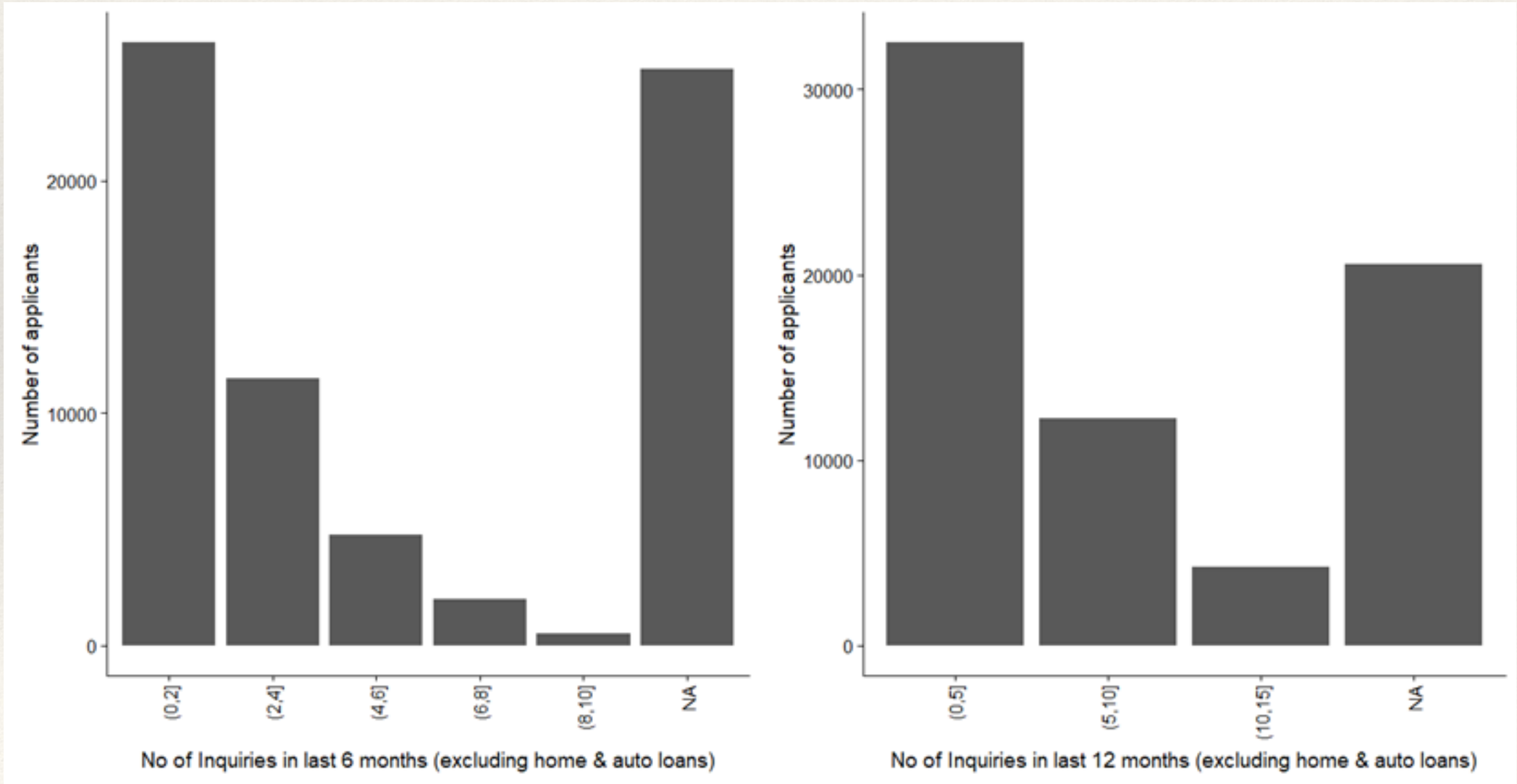
Outstanding Balance & Total Number of Trades





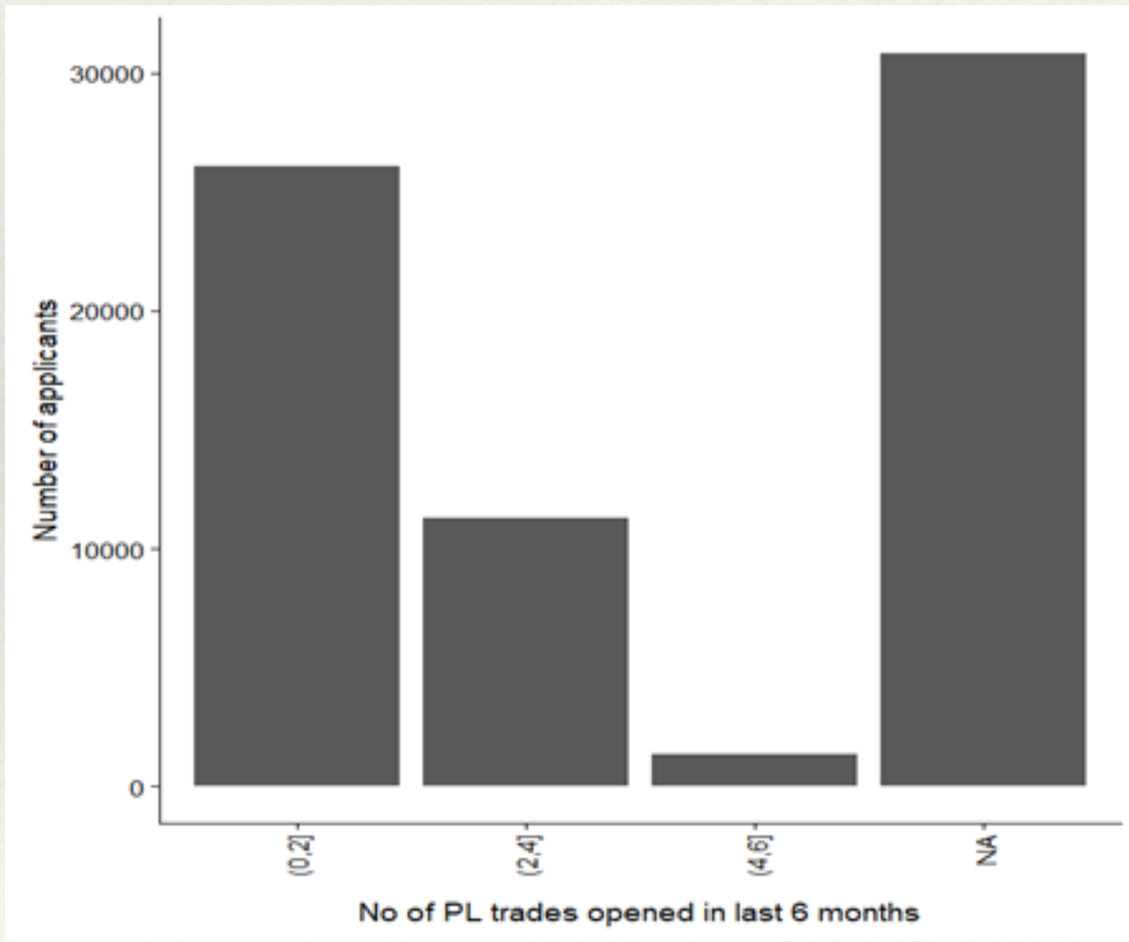
EDA

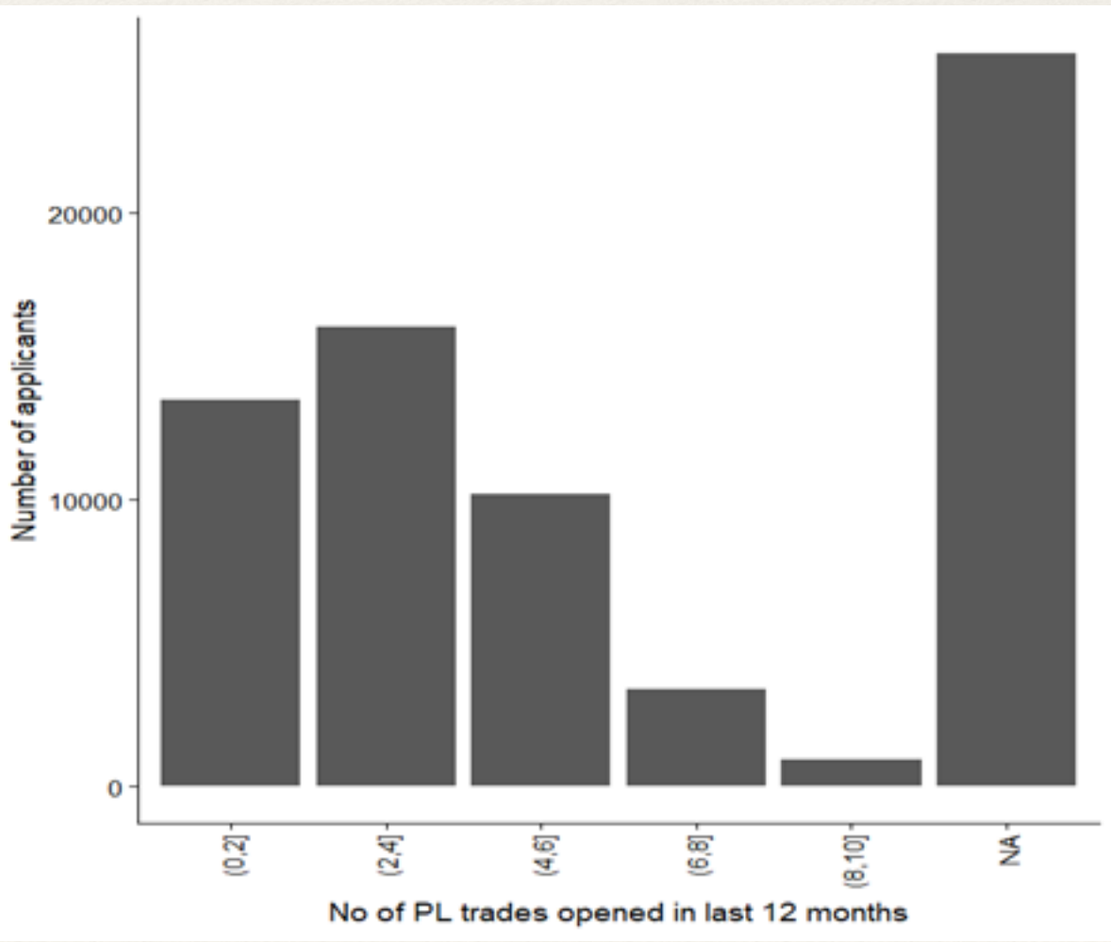
Number of Inquiries



EDA

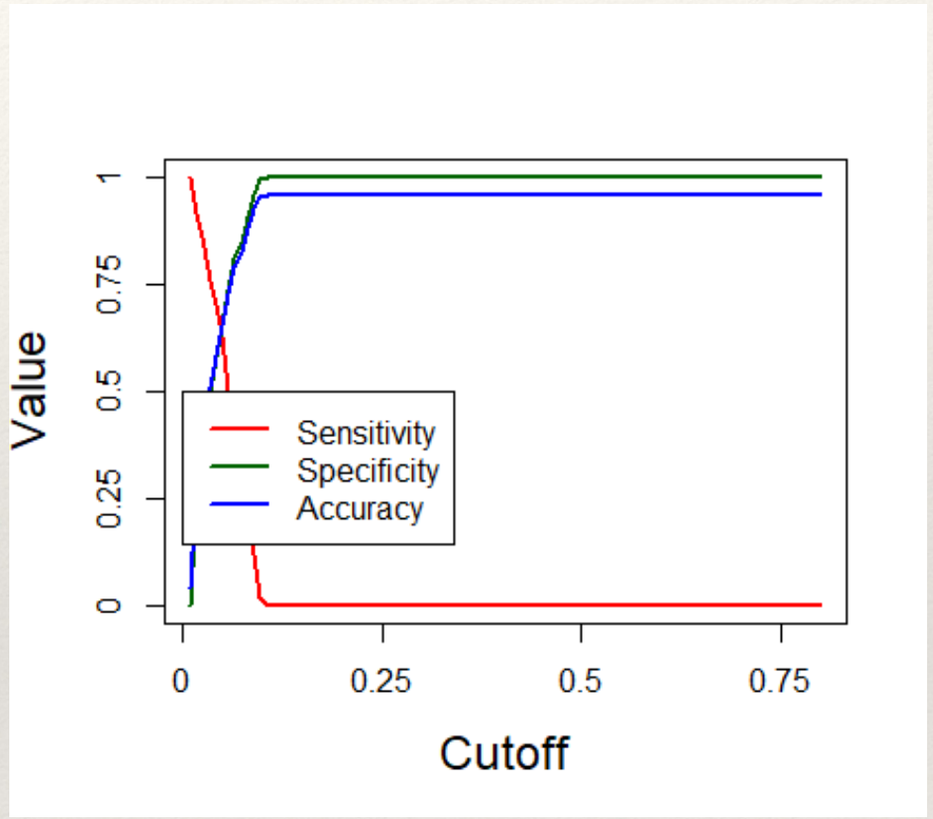
Profit/Loss Trades





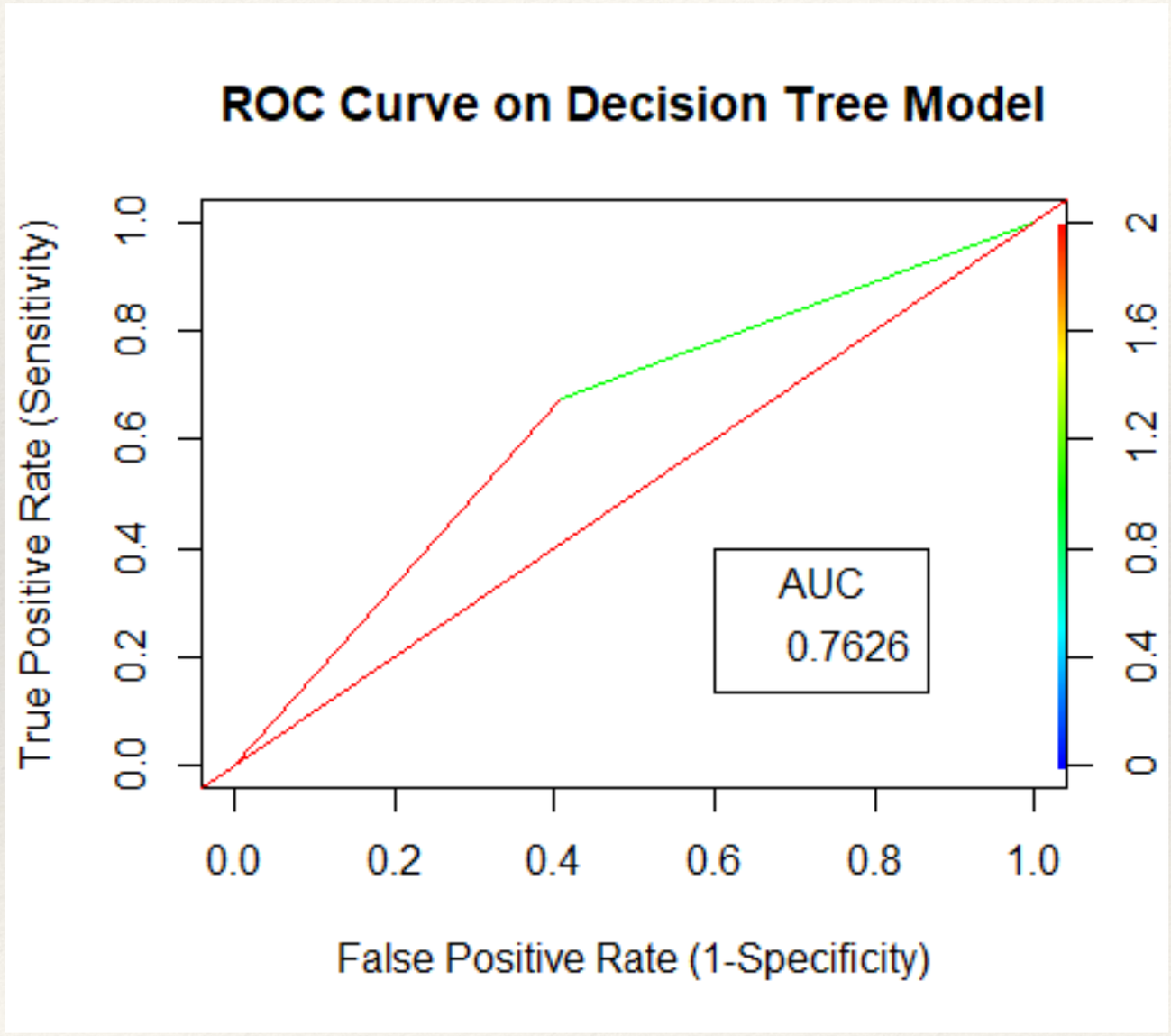
Logistic Regression ROC & AUC

Logistic Regression ROC
Class : Binomial Classification



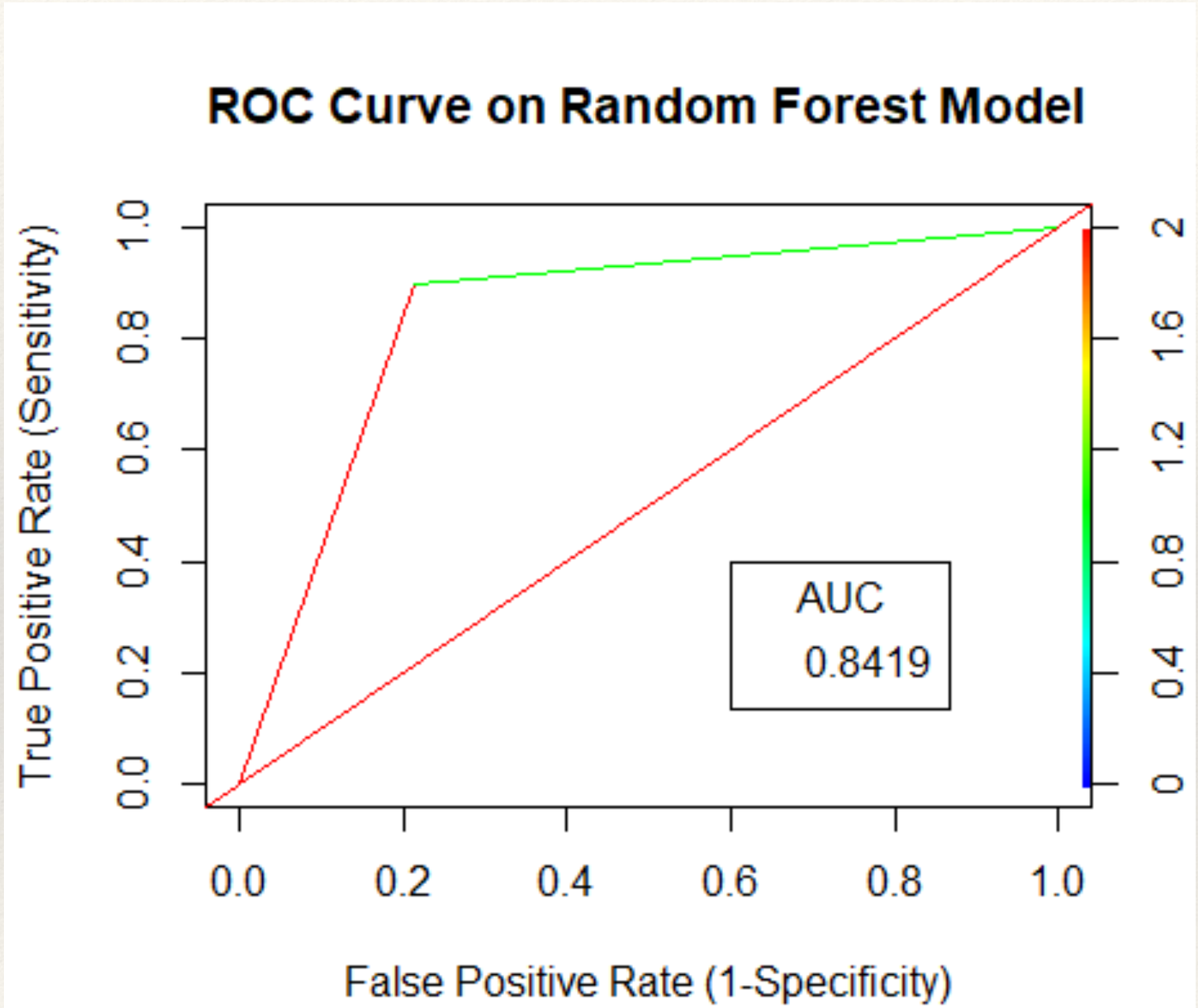
Area Under Curve 63%
Accuracy 62%
Sensitivity 62%
Specificity 63%
KS Statistics 25%

Decision Tree ROC & AUC



Area Under Curve 76%
Accuracy 62%
Sensitivity 60%
Specificity 62%
KS Statistics 53%

Random Forest ROC & AUC



Area Under Curve 84%

Accuracy 79%

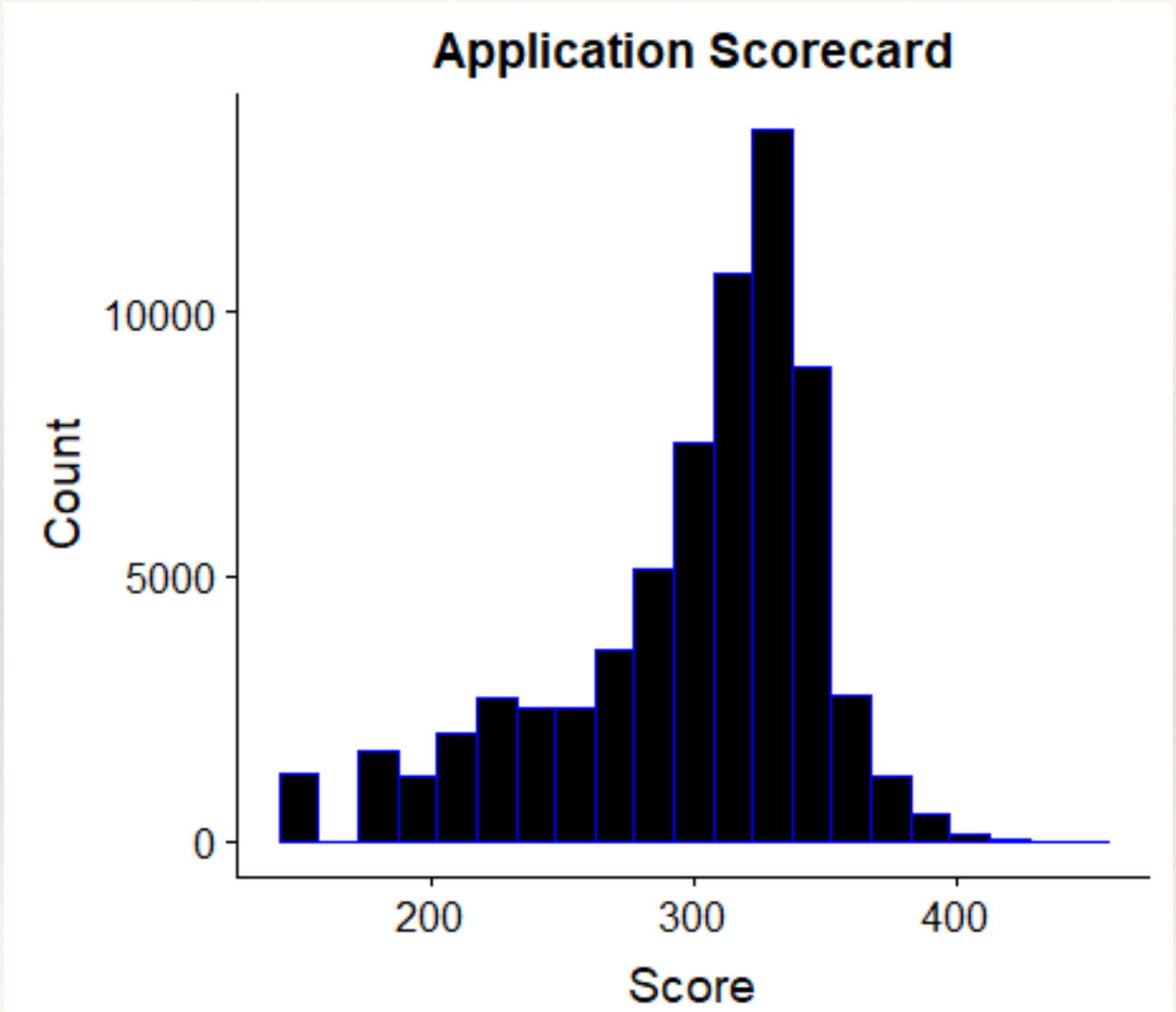
Sensitivity 89%

Test 79%

KS Statistics 52%

Based on evaluation metrics, this model is chosen for further derivation of application score card. The model is further tested on balanced data using Sampling Technique -- SMOTE

Application Scorecard Distribution



This application scorecard is built on Random Forest Model

Max Application Score : 542

Min Application Score : 242

Application Cutoff Score : 334

Financial Benefit Analysis

	Confusion Matrix						
	Reference						
Prediction	***Good Cust(0)***	***Bad Cust(1)***					
Good Cust(0)	52626	775					
Bad Cust(1)	13998	2164					
Accuracy	78.76						
Sensitivity	78.99						
Specificity	73.63						
Percentage of Good Customers predicted as Bad customer by the model	21.01	Good Customers are Revenue generated for the bank, Bad Custoers are loss for the bank	Revenue Lost vs Loss Saved	**13998 * Rev < 2164 * Loss**	*Solving this equation*	**Loss > 6.46 Rev**	*Since Loss and Rev values are not known, the Financial Benefit Analysis is concluded here*

Good Customers are the revenue generated for the credit card company, while Bad Customers are Loss for the company

- Our model has predicted 13998 Good customers as Bad customers. These are Revenue Lost for the bank
- Our model has classified 2164 customers as Bad customers which is Loss Saved for the bank
- In this case, model will be Financially Benefitting when Loss Saved is greater than Revenue Lost for the bank
- Or 13998 * Rev < 2164 * Loss
- On Solving this equation, Loss > 6.46 Rev
- Since Loss and Revenue Data is not available, the Financial Benefit analysis is concluded here.