BFS Capstone Project
CredX – Credit Card Application Management

Final Submission

Group:

1. Shankadeep Ghosal
2. Ujwal Kiran Bhargava
3. Ayon Sarkar
4. Shruti GS

Introduction, Business Understanding and Objective

- CredX is a leading credit card provider that gets **thousands of credit card applicants every year**.
- However in the past few years, it has experienced an **increase in credit loss**.
- Hence ,the CEO believes that the best strategy to mitigate credit risk is to '**acquire the right customers**'.
- Using the past data of the bank's applicants, we need to

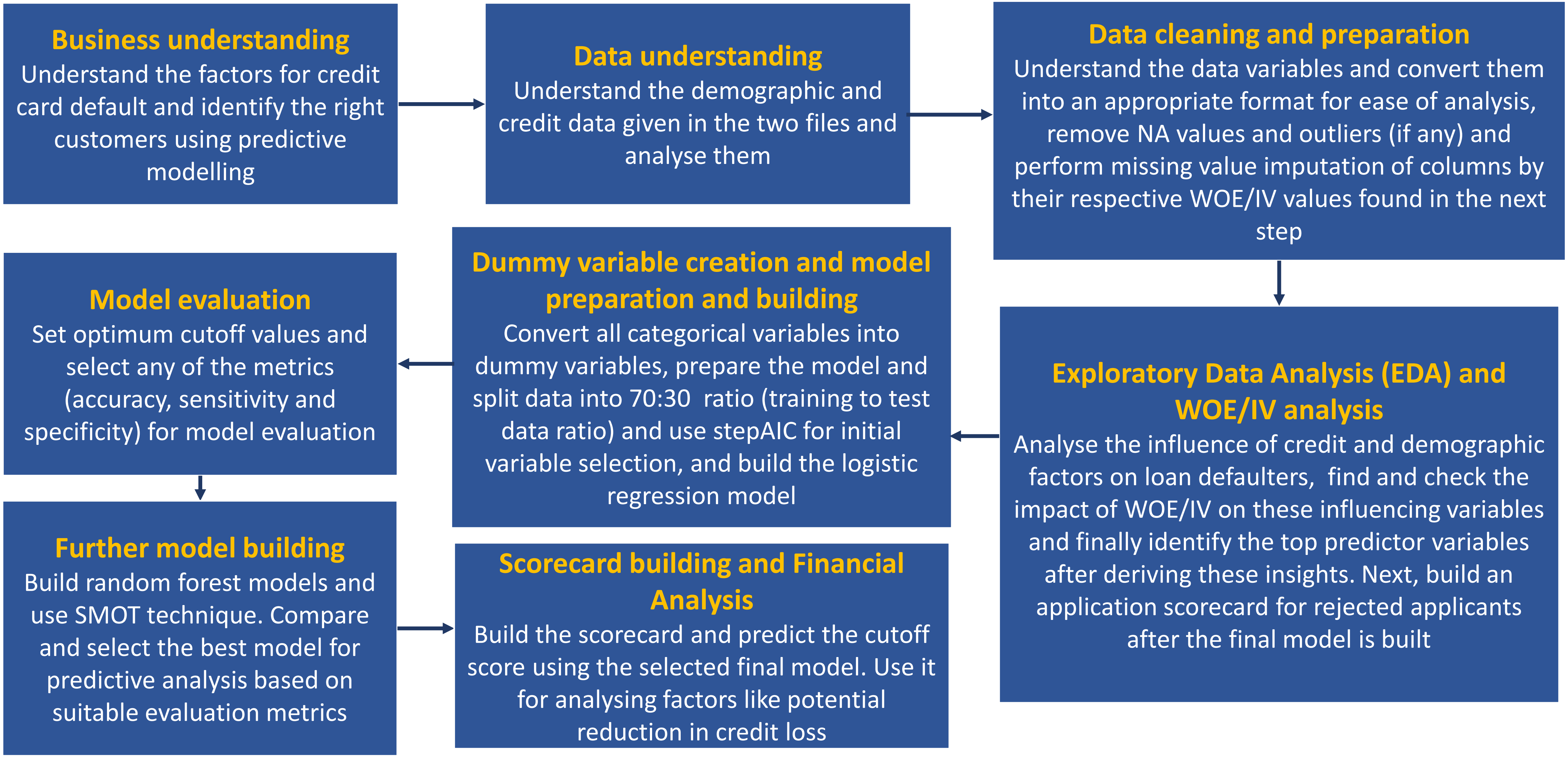
▪ **Determine the factors affecting credit risk**

▪ **Create strategies to mitigate the acquisition risk**

▪ **Assess the financial benefit of the project**.

- Thus, the main objective and intention of this capstone project is to help CredX **identify the right customers using predictive modelling** and **subsequent analysis of the models built**.

Problem Solving Methodology



Business understanding

Understand the factors for credit card default and identify the right customers using predictive modelling

Data understanding

Understand the demographic and credit data given in the two files and analyse them

Data cleaning and preparation

Understand the data variables and convert them into an appropriate format for ease of analysis, remove NA values and outliers (if any) and perform missing value imputation of columns by their respective WOE/IV values found in the next step

Exploratory Data Analysis (EDA) and WOE/IV analysis

Analyse the influence of credit and demographic factors on loan defaulters, find and check the impact of WOE/IV on these influencing variables and finally identify the top predictor variables after deriving these insights. Next, build an application scorecard for rejected applicants after the final model is built

Dummy variable creation and model preparation and building

Convert all categorical variables into dummy variables, prepare the model and split data into 70:30 ratio (training to test data ratio) and use stepAIC for initial variable selection, and build the logistic regression model

Model evaluation

Set optimum cutoff values and select any of the metrics (accuracy, sensitivity and specificity) for model evaluation

Further model building

Build random forest models and use SMOT technique. Compare and select the best model for predictive analysis based on suitable evaluation metrics

Scorecard building and Financial Analysis

Build the scorecard and predict the cutoff score using the selected final model. Use it for analysing factors like potential reduction in credit loss

Data Understanding

- There are two data sets in this project — **demographic** and **credit bureau** data.
- **Demographic/application data**: This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- **Credit bureau**: This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.
- Both files contain a performance tag which represents whether the applicant has gone 90 days past due or worse in the past 12-months (i.e. defaulted) after getting a credit card.

Data Cleaning and Preparation

- Checking all the columns for **missing values** and imputing them with the respective WOE values obtained after analysis.
- Checking all the columns for **NA values** and removing them since they are very few in number as compared to original data.
- Checking the necessary columns for **duplicate values** and removing the same.
- **Detecting outliers using the quartiles and boxplots** and removing the same.
- Finding **WOE/IV values of driver variables** for EDA and further analysis.
- Creating a **master file with all the relevant variables**.
- Here, **Application ID is the primary key** and hence, it is used for merging the demographic and credit data files.
- Thus**, 29 variables** are present in the final dataset after merging (11 in the original demographic data set and 19 in the original credit data set).
- Creating **dummy variables** before model building.

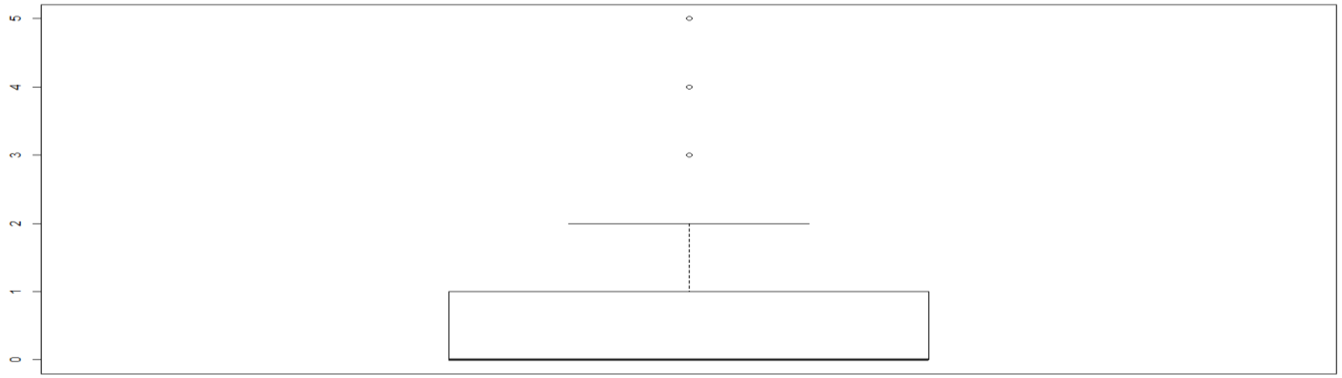
EDA (Exploratory Data Analysis)

In the following slides, we explain how credit card default varies with demographic factors such as age and income as well as credit card factors such as average credit card utilisation and also how the WOE/IV values impact the overall analysis.

For demographic analysis, performance tag 1 is taken, as we are concerned about analysing the number of defaulters w.r.t to these kind of factors.

Boxplot and Outlier Analysis

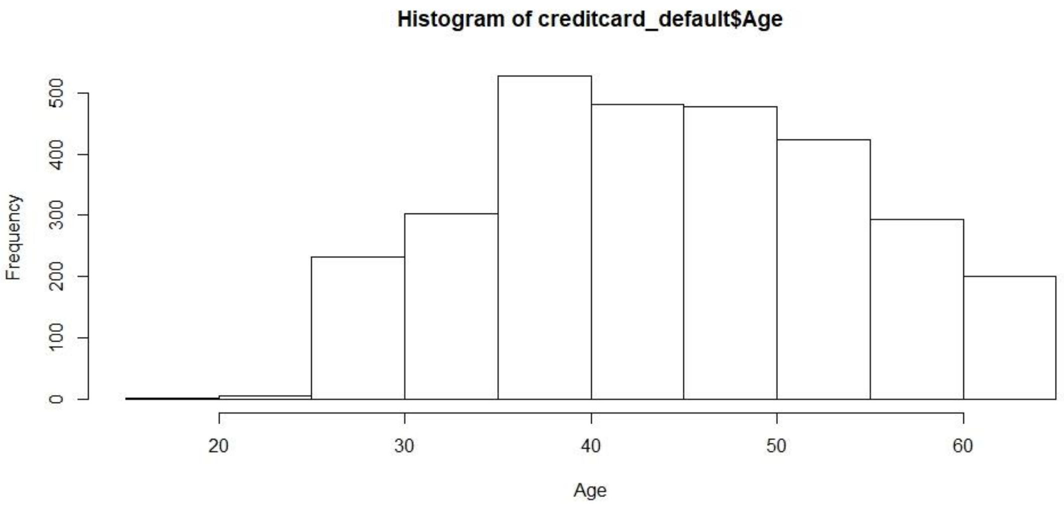
For some variables like No.of.times.90.DPD.or.worse.in.last.12.months, boxplot turns out to be like this:

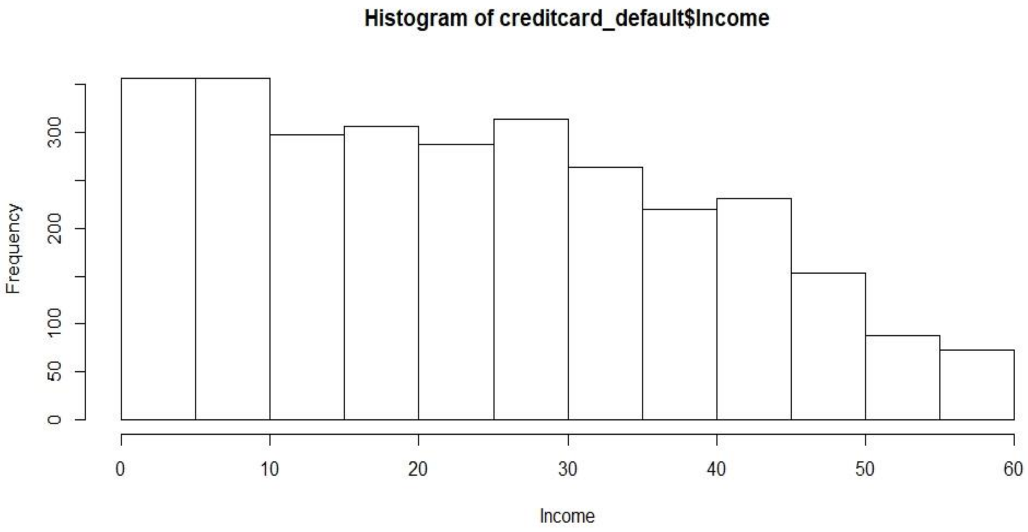


This shows the presence of outliers in the respective variables which can be removed for more accurate analysis.

NOTE: This step may be redundant as some of the outliers can be taken care of using WOE/IV values and binning of variables.

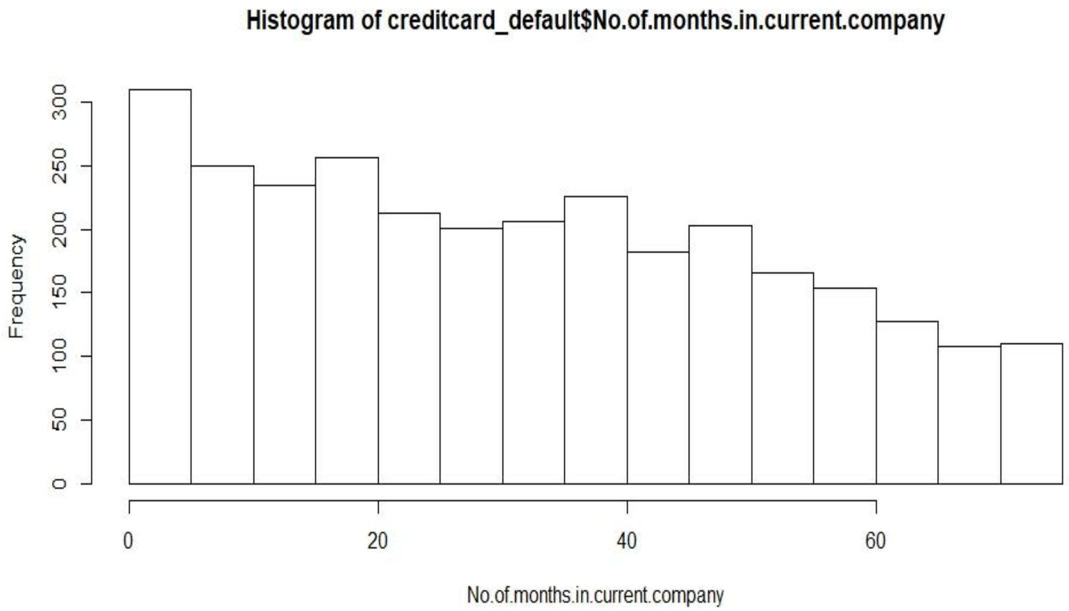
Analysis of Credit Card Default w.r.t. Demographic Factors

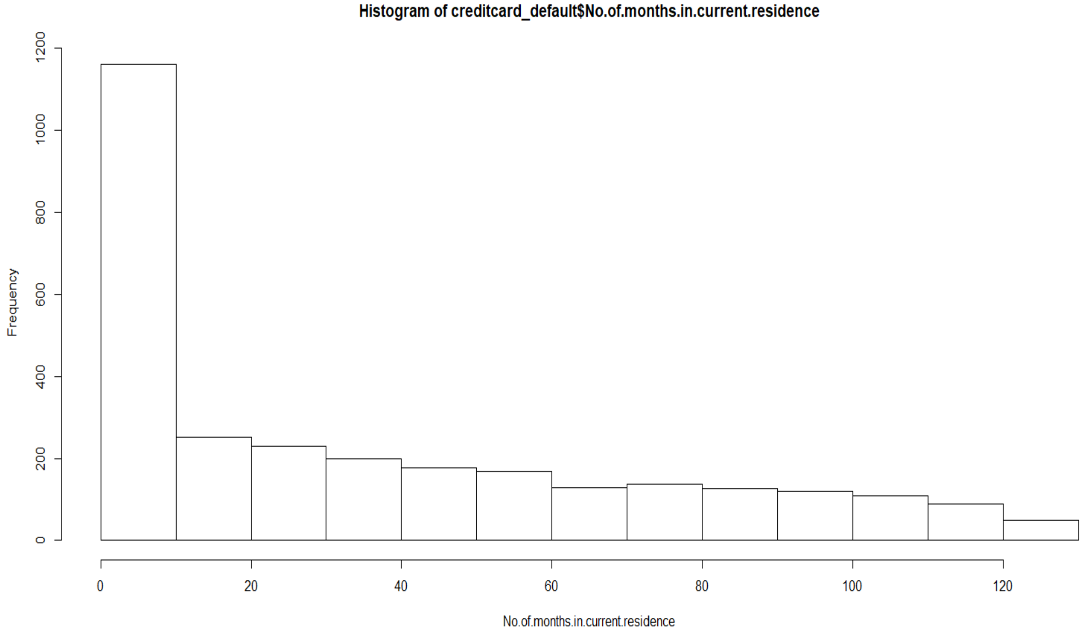




Most defaulters lie between the ages 35-50 and incomes 0-10 (lower income ranges are more prone to default).

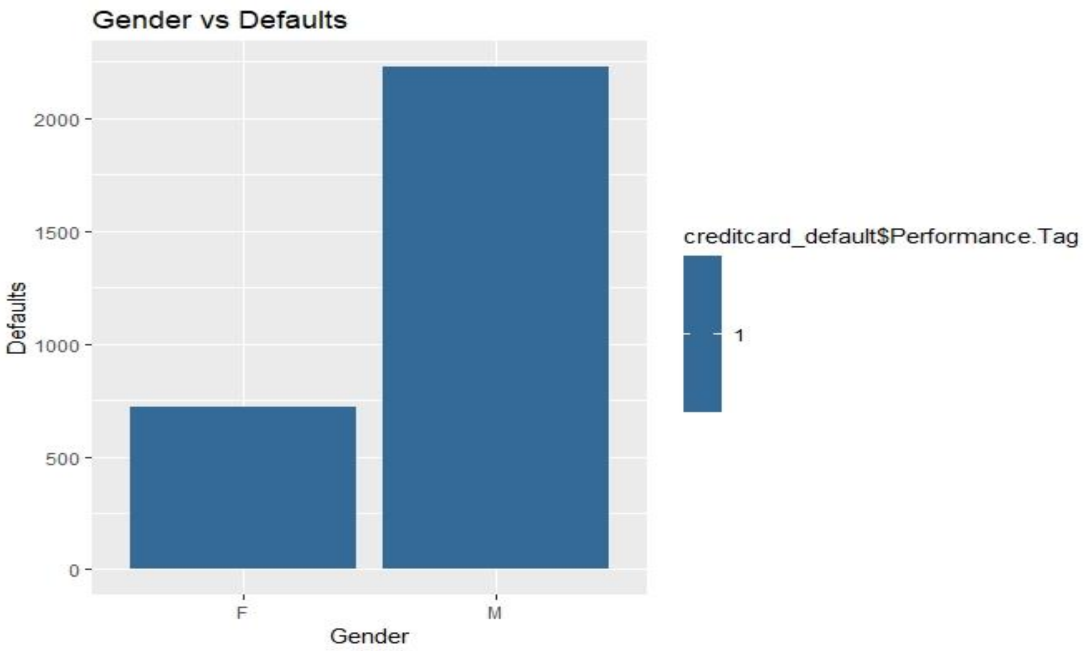
Analysis of Credit Card Default w.r.t. Demographic Factors

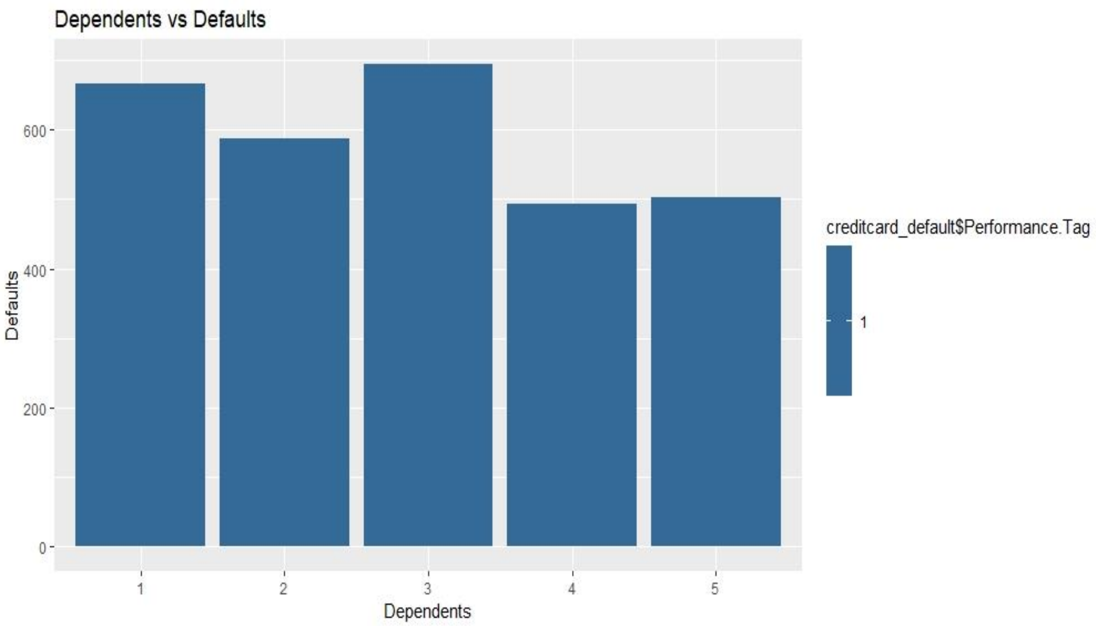




Most defaulters spend only 0-5 months in the current company they are employed and 0-10 months in the current residence they stay in.

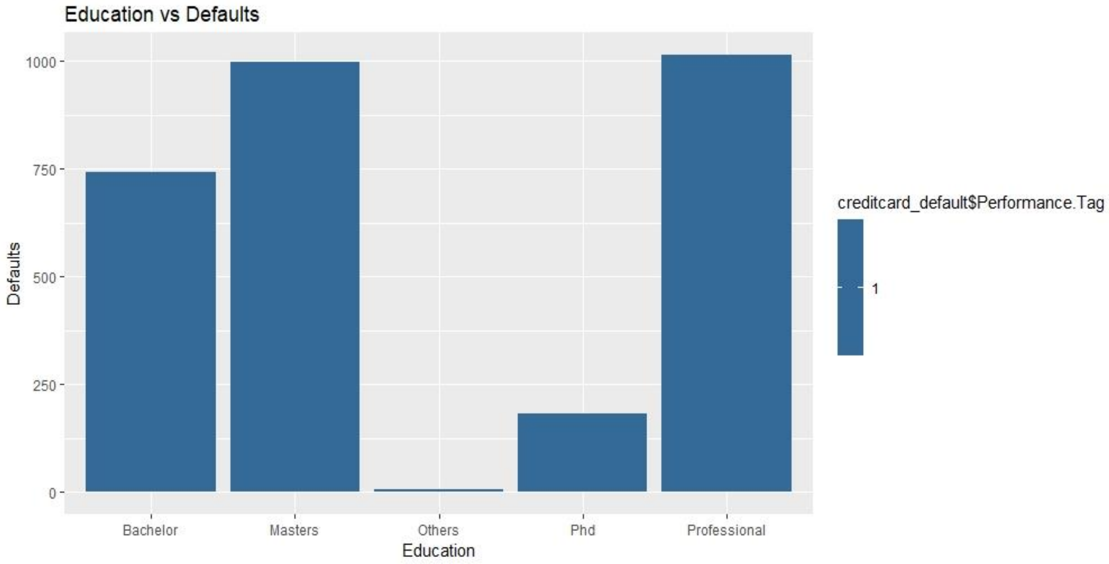
Analysis of Credit Card Default w.r.t. Demographic Factors

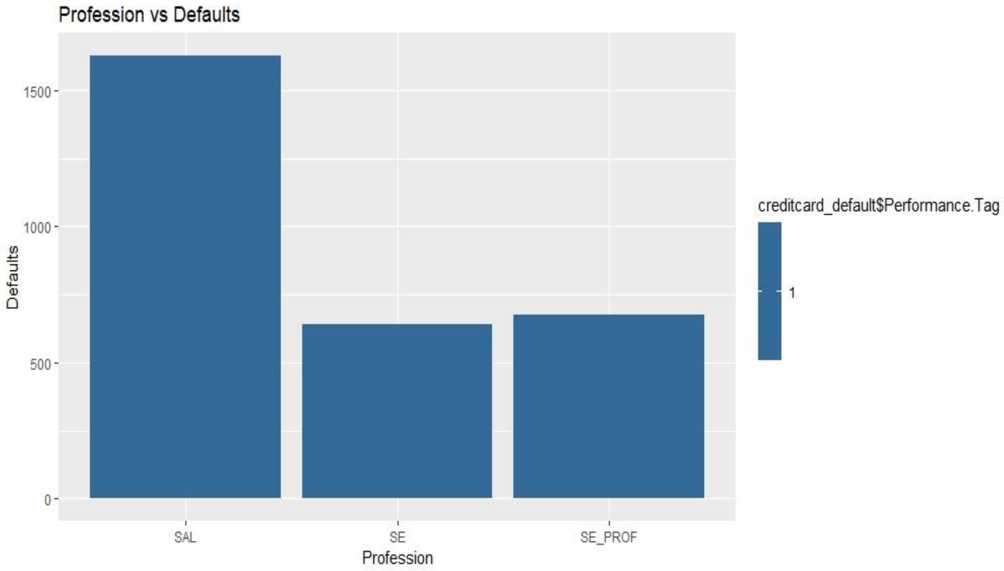




Most defaulters are male and have 3 dependents.

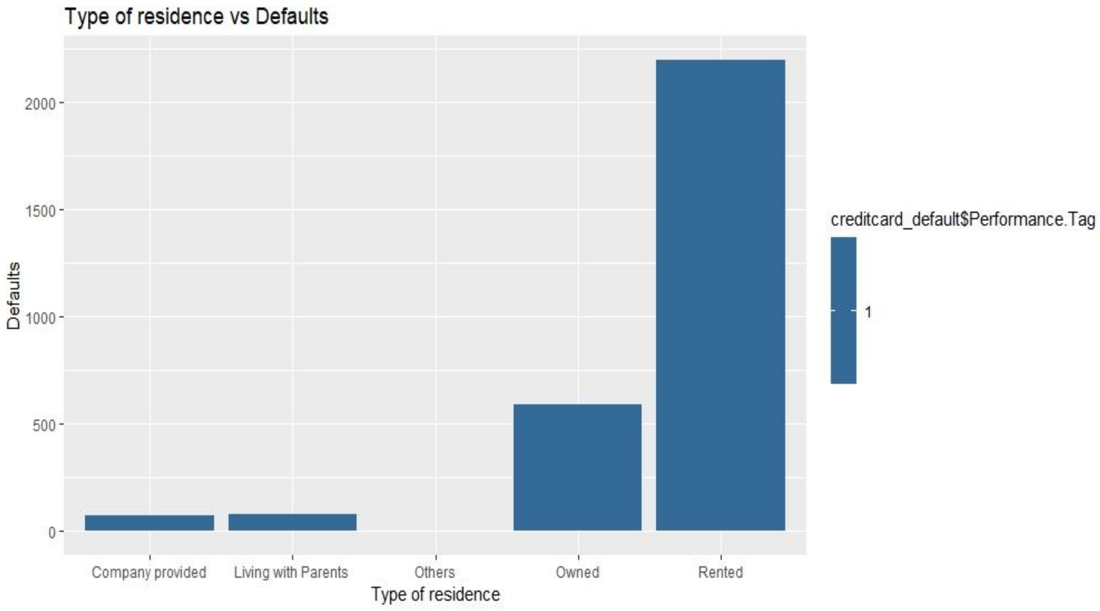
Analysis of Credit Card Default w.r.t. Demographic Factors

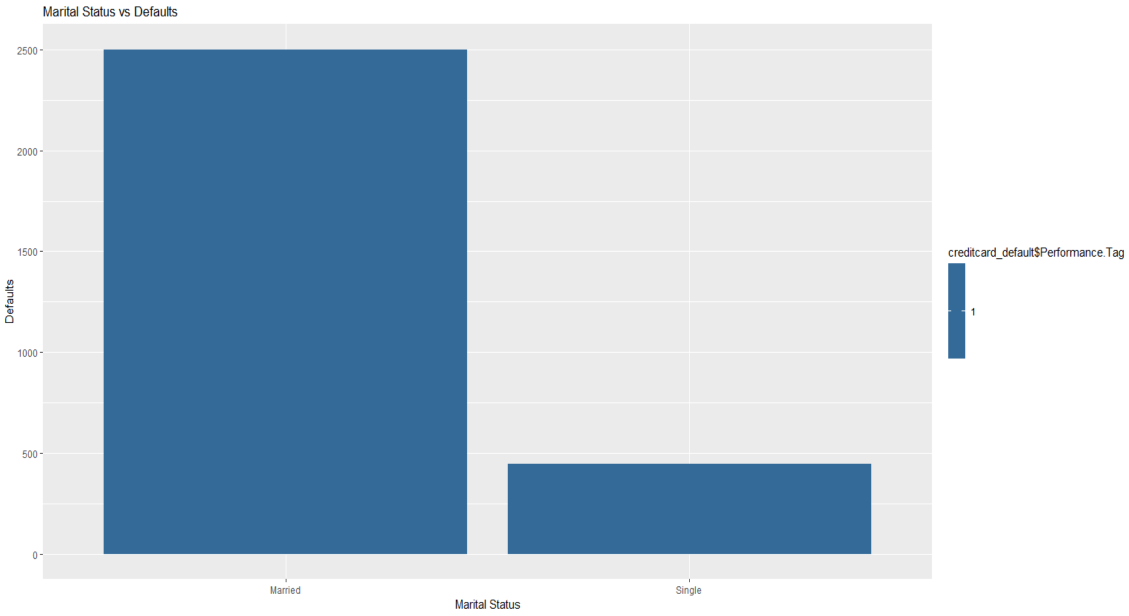




Most defaulters have education level as master/professional and have SAL as their profession.

Analysis of Credit Card Default w.r.t. Demographic Factors

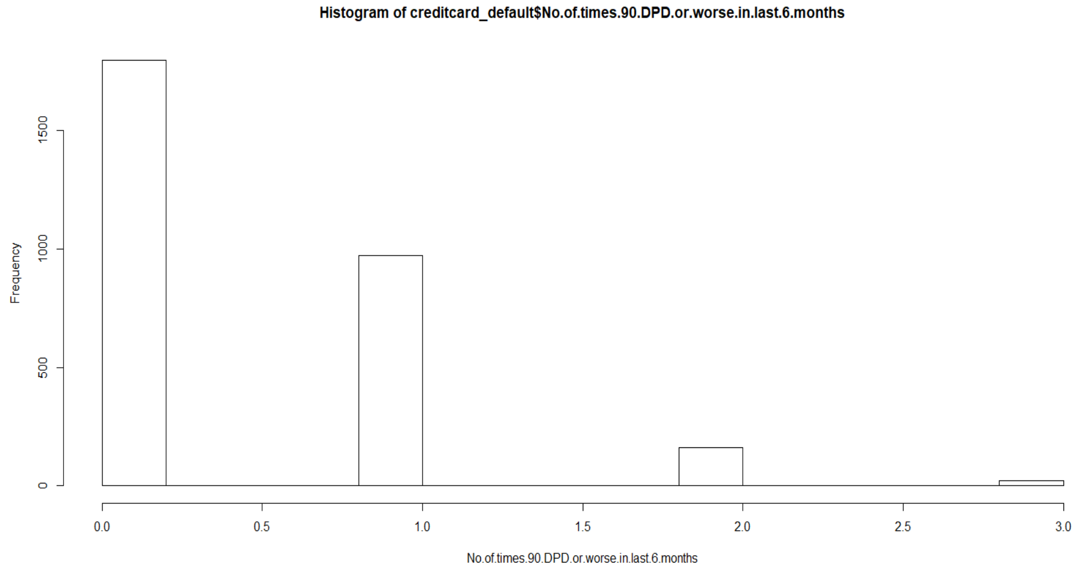


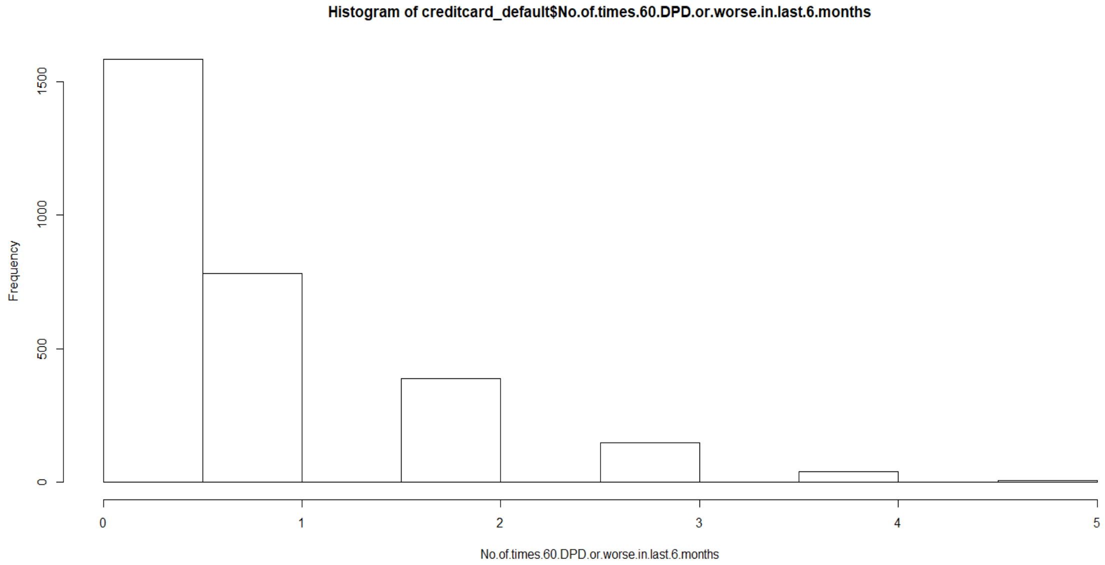


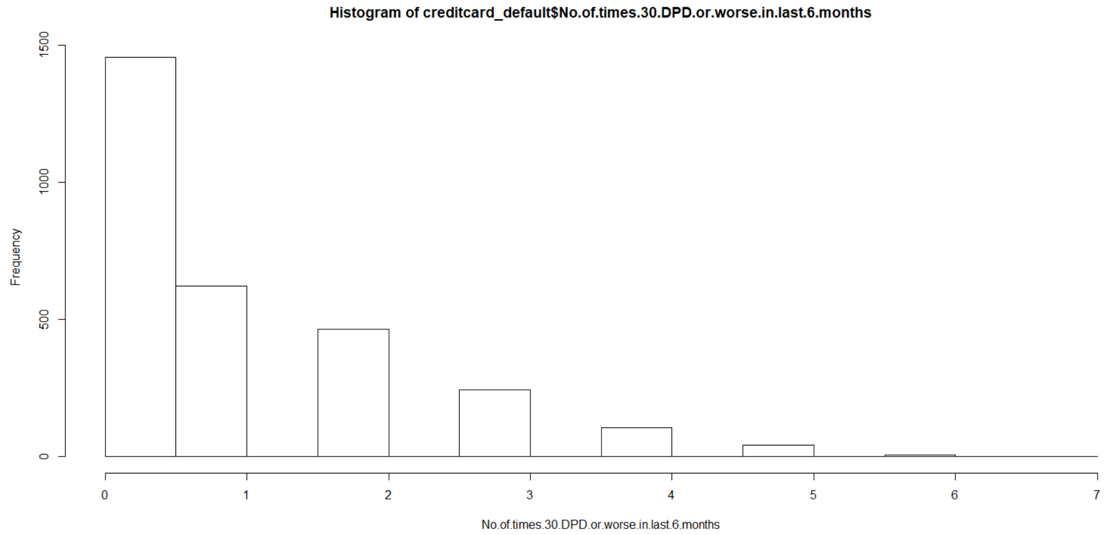
Most defaulters live in rented residences and are married.

From the graphs in this section, it is clear that **lower demographic factors like income, number of months in current company and residence, significantly increase the default rate**.

Analysis of Credit Card Default – Default Over 6 Months

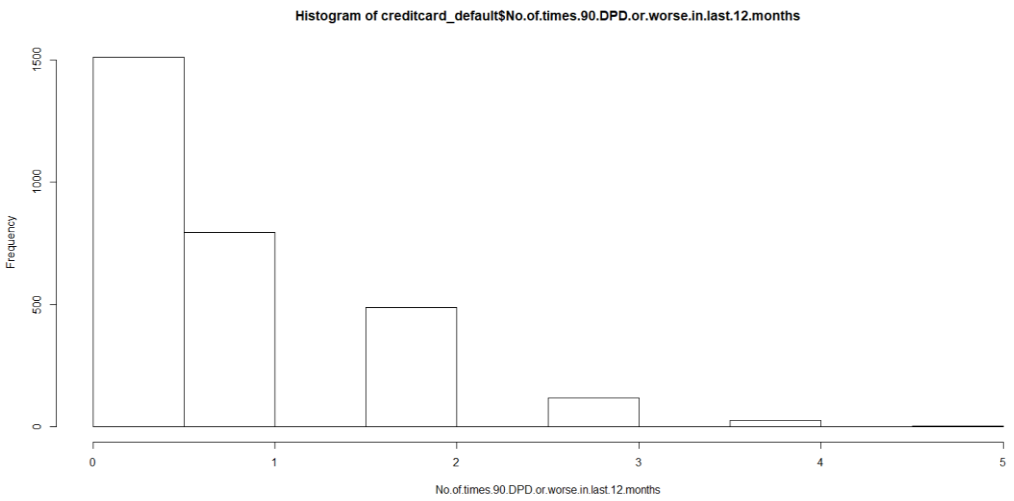

Histogram of creditcard_default$No.of.times.90.DPD.or.worse.in.last.6.months

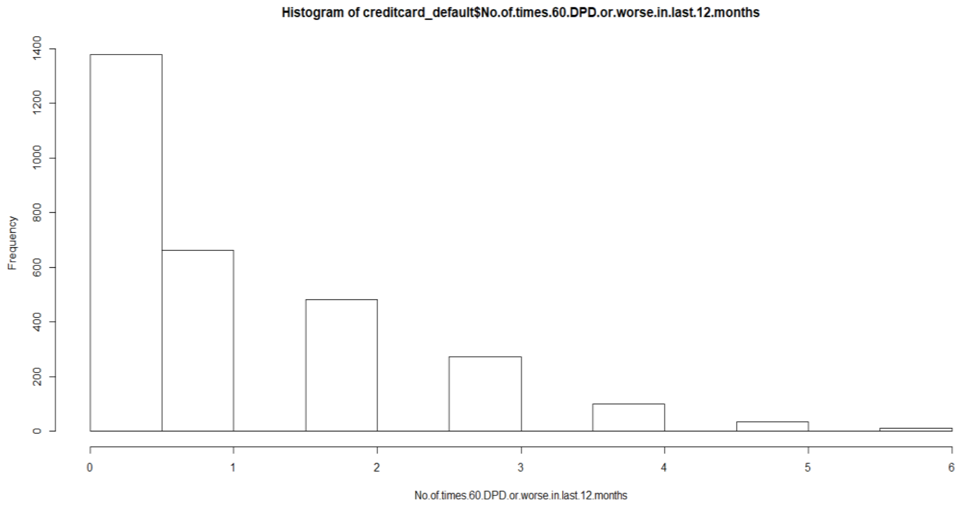

Histogram of creditcard_default$No.of.times.60.DPD.or.worse.in.last.6.months

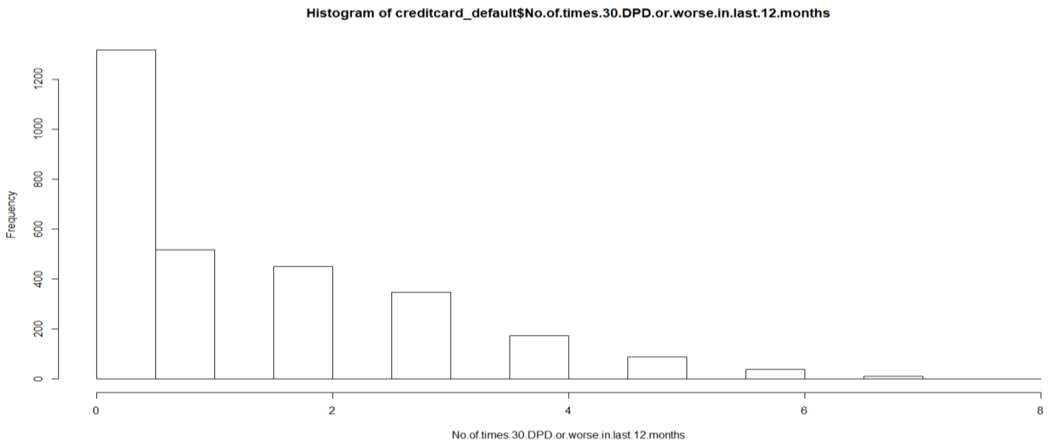

Histogram of creditcard_default$No.of.times.30.DPD.or.worse.in.last.6.months

From the given histograms, it is clear that lower values, lead to higher defaulters.

Analysis of Credit Card Default – Default Over 12 months

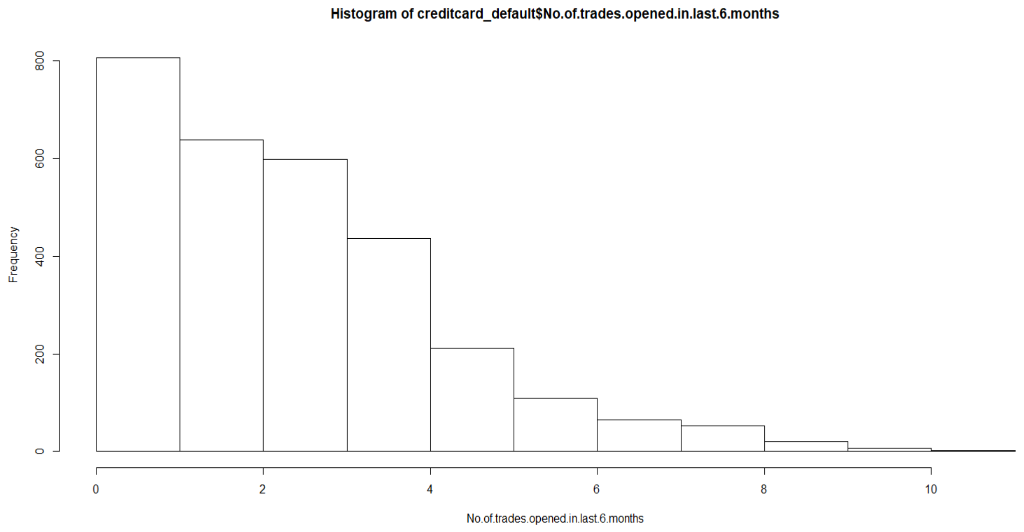

Histogram of creditcard_default$No.of.times.90.DPD.or.worse.in.last.12.months

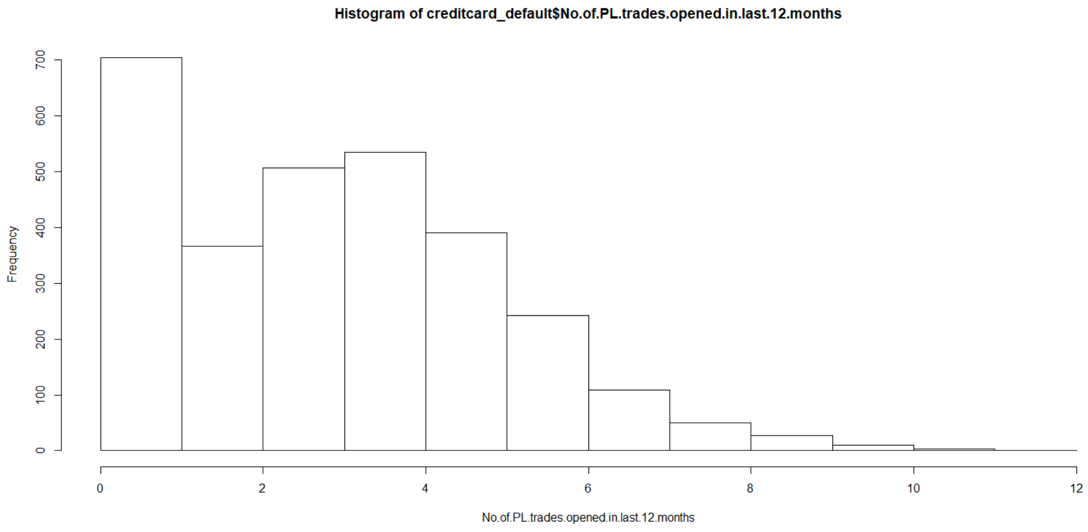

Histogram of creditcard_default$No.of.times.60.DPD.or.worse.in.last.12.months


Histogram of creditcard_default$No.of.times.30.DPD.or.worse.in.last.12.months

From the given histograms, it is clear that lower values, lead to higher defaulters.

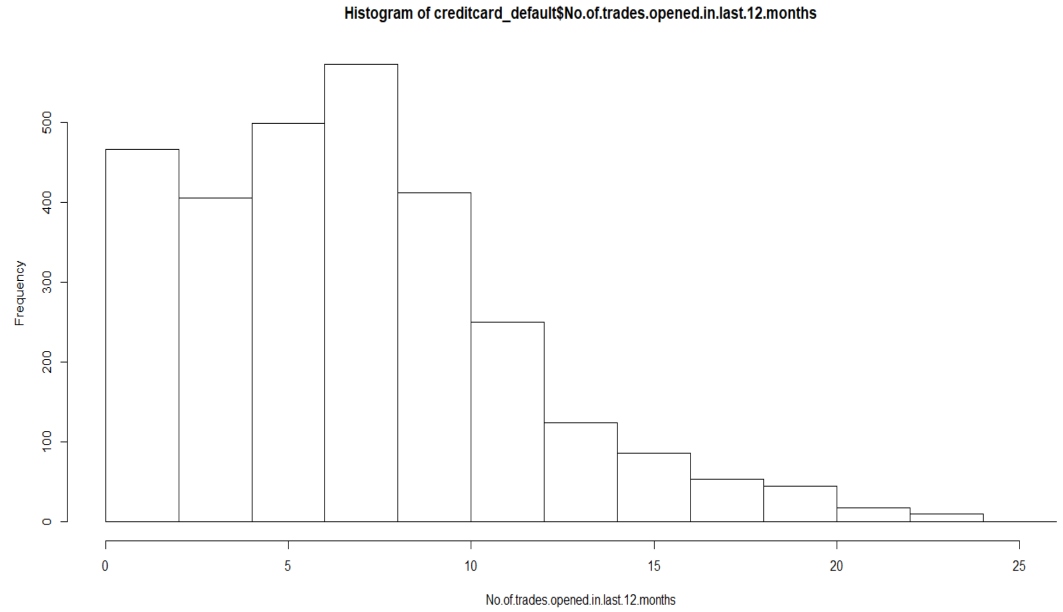
Analysis of Credit Card Default – No. of Trades Opened

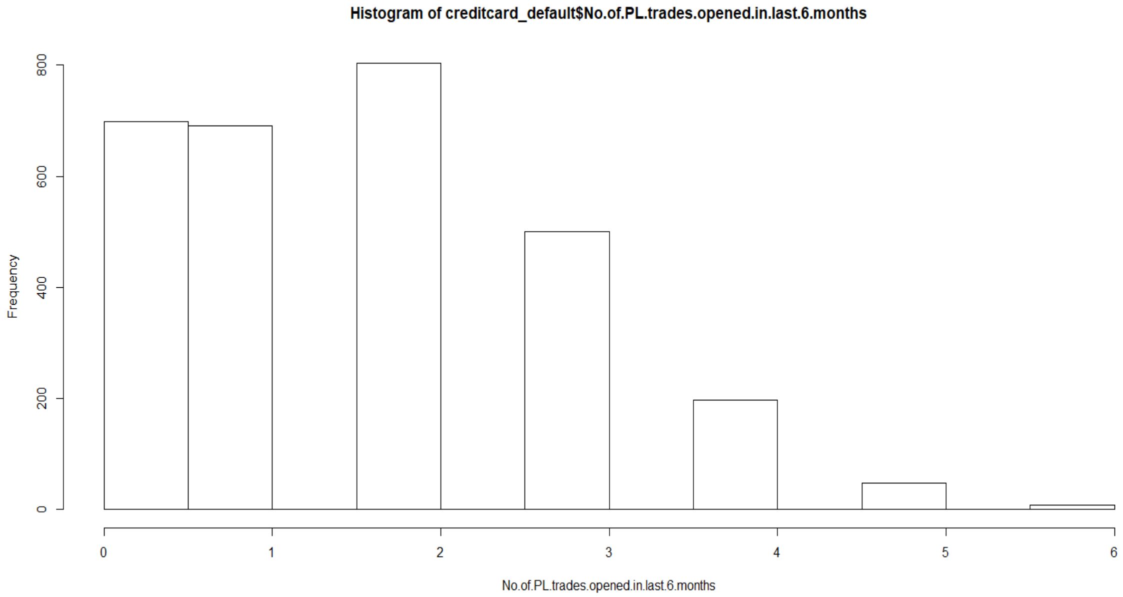




From the given histograms, it is clear that lower values lead to higher defaulters.

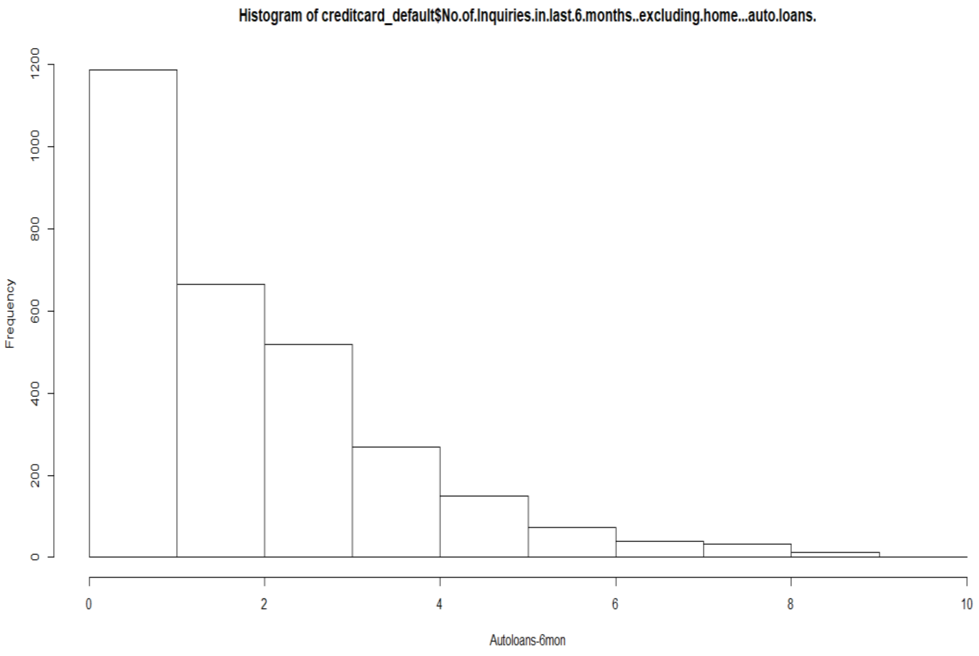
Analysis of Credit Card Default – No. of PL Trades Opened

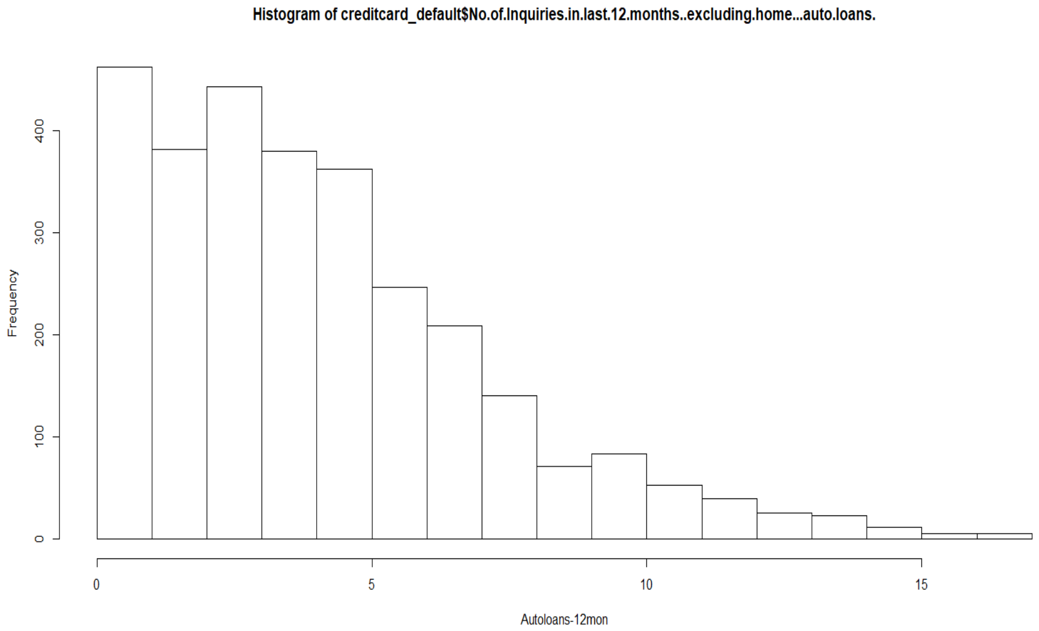

Histogram of creditcard_default$No.of.trades.opened.in.last.12.months


Histogram of creditcard_default$No.of.PL.trades.opened.in.last.6.months

From the given histograms, it is clear that lower values lead to higher defaulters.

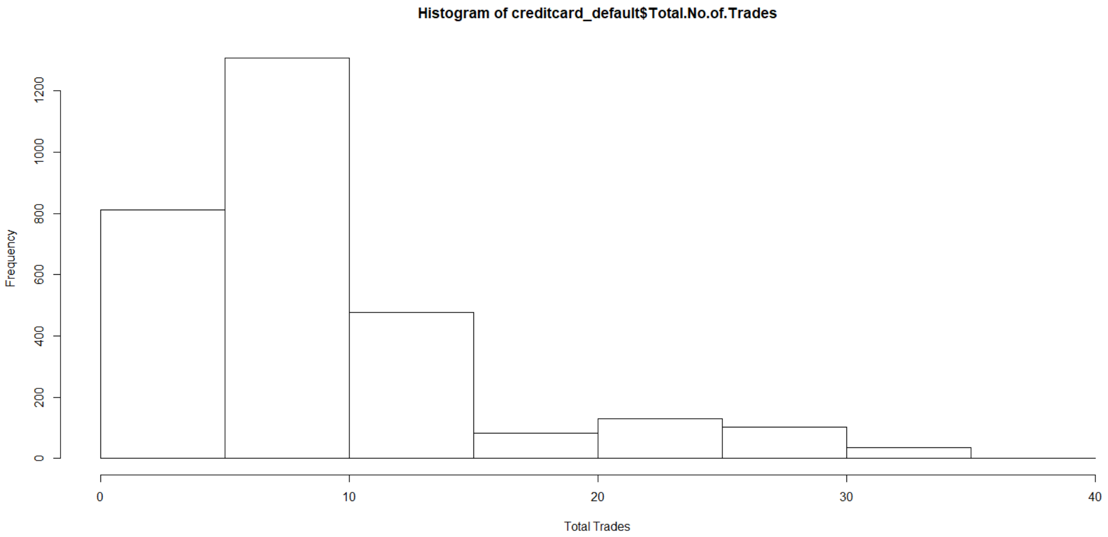
Analysis of Credit Card Default – No. of Inquiries

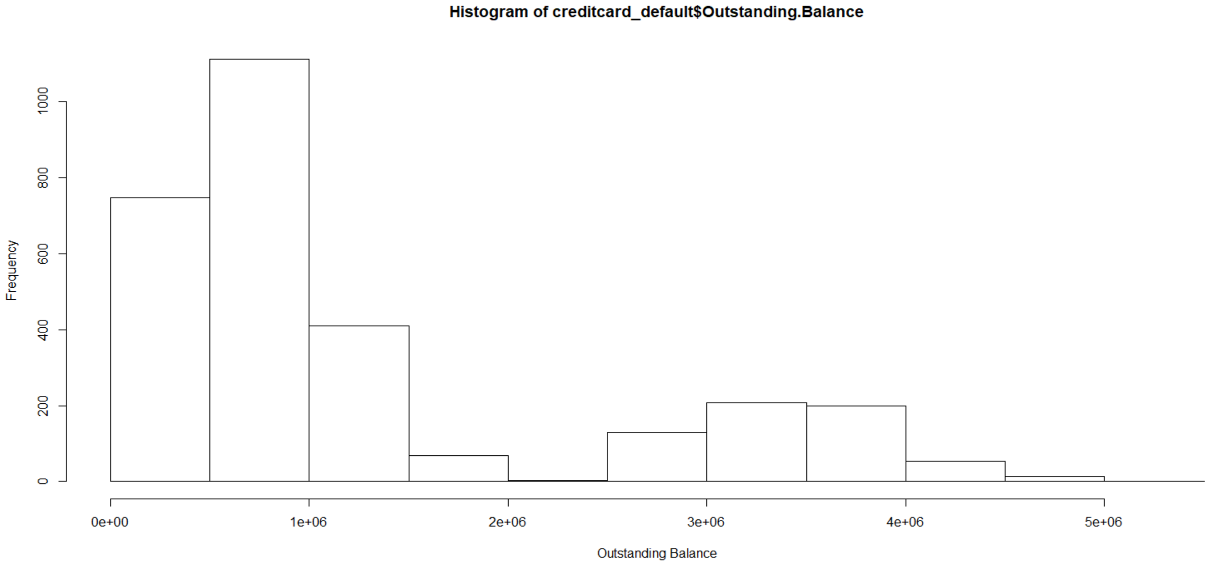


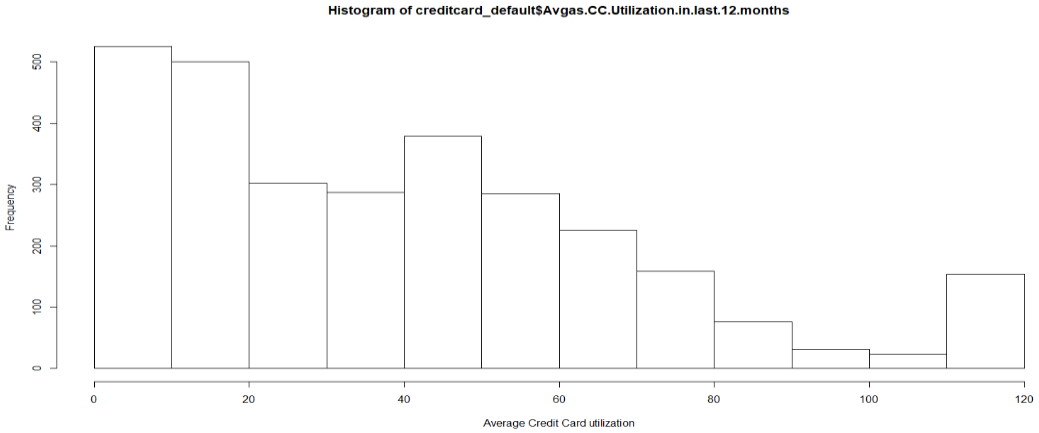


From the given histograms, it is clear that lower values lead to higher defaulters.

Analysis of Credit Card Default – Summary



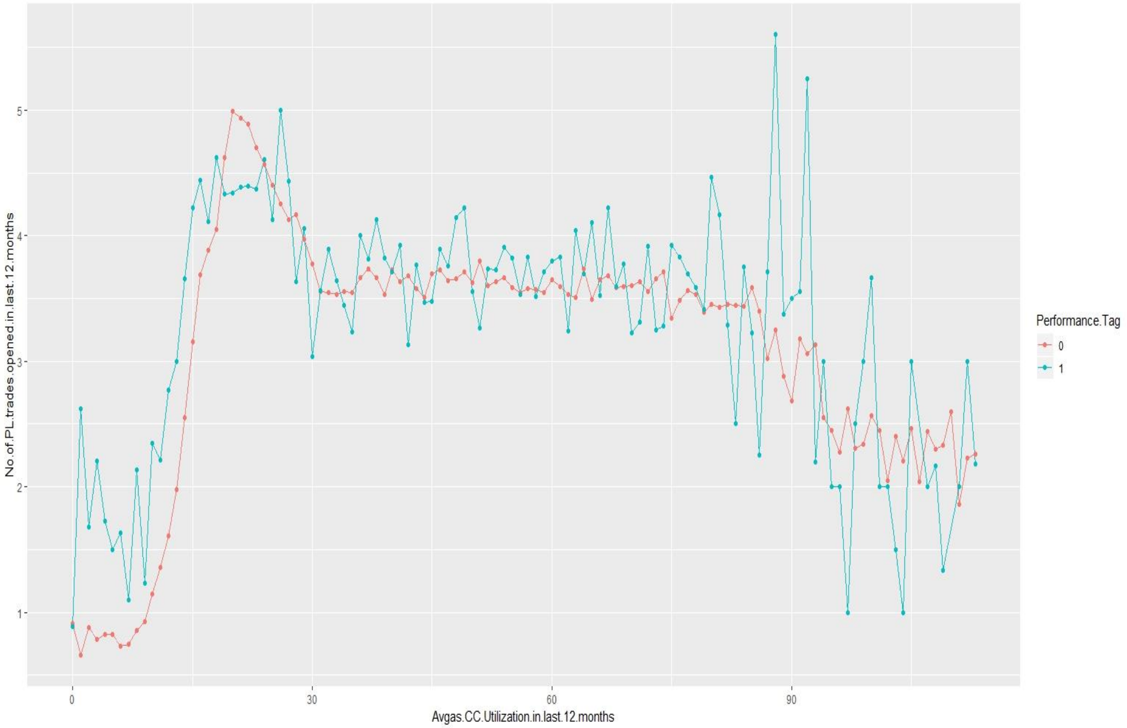


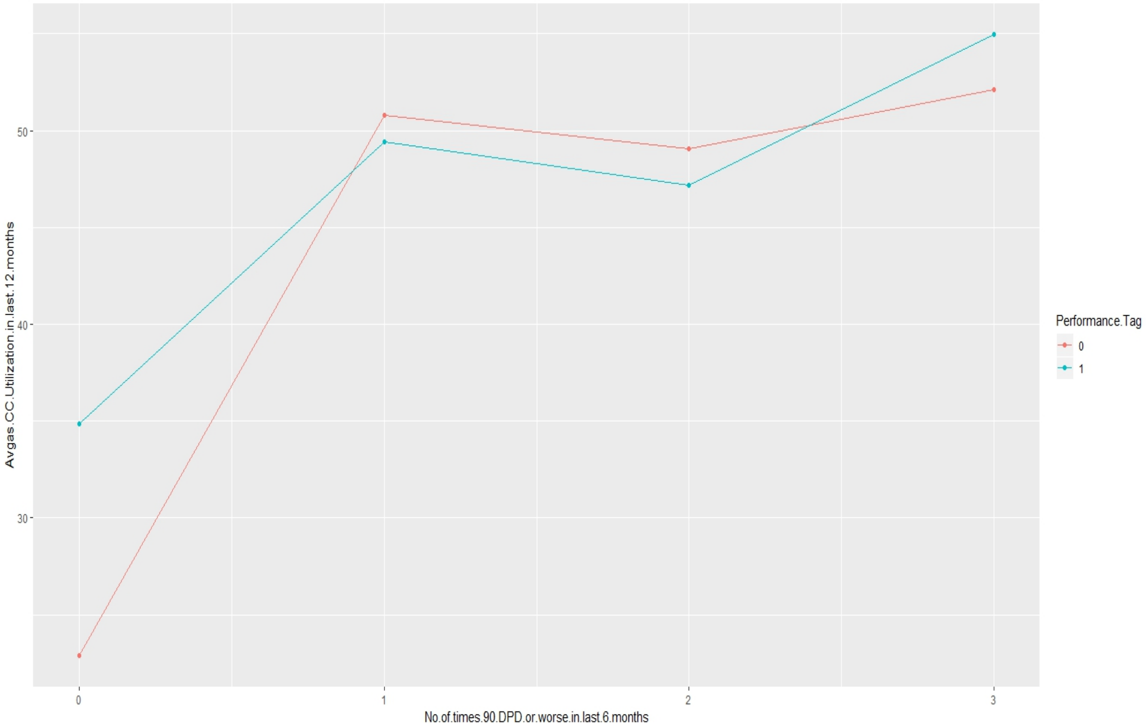


From the given histograms, it is clear that lower values lead to higher defaulters.

From the graphs in this section, it is clear that **low credit factors like outstanding balance, total number of trades and average credit card utilization significantly increase the risk of default**.

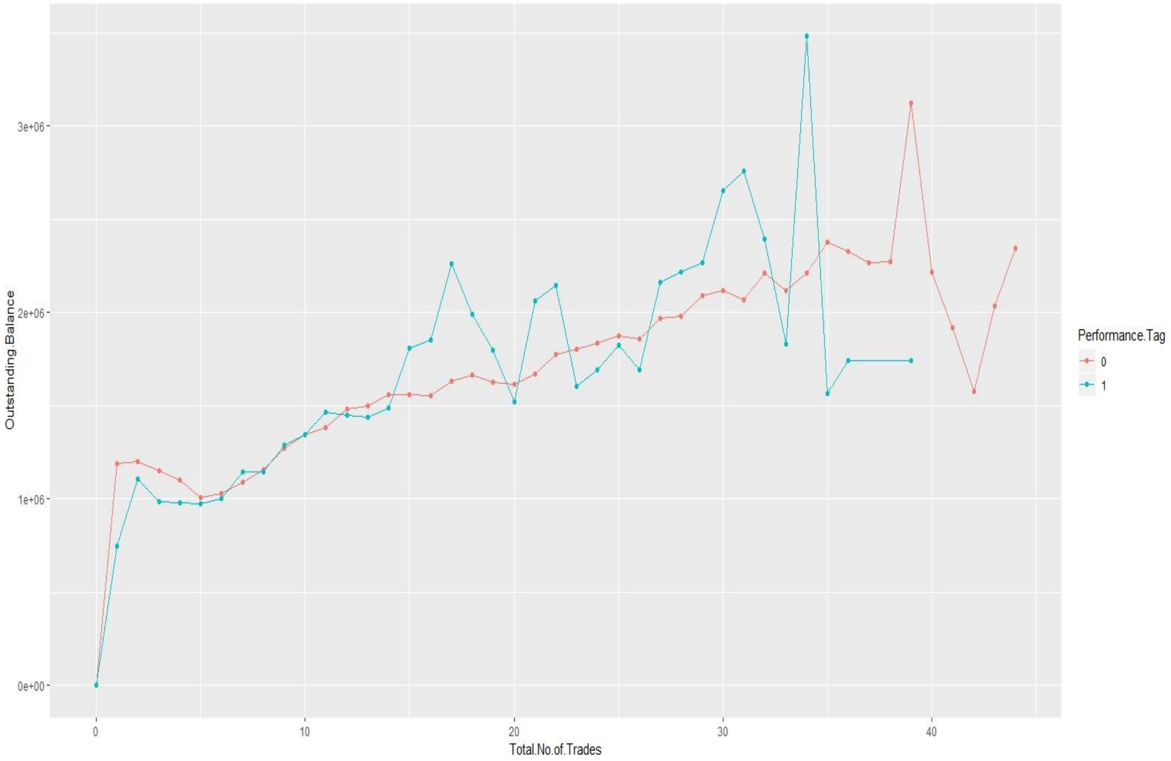
Analysis of Credit Card Default – Bivariate Analysis w.r.t. Average Credit Card Utilisation

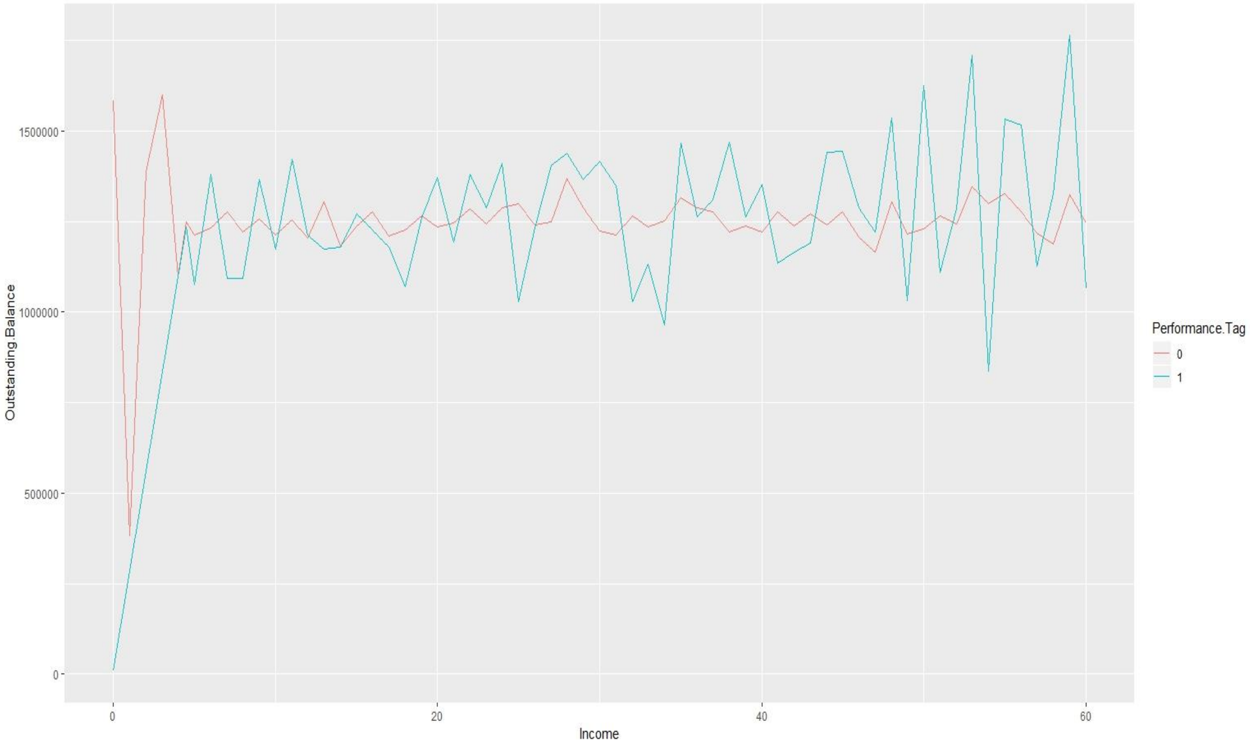




From the graphs, it is clear that **number of PL trades opened** is **relatively higher** for defaulting users.
The **Average Credit Card Utilization tends to be higher** for defaulting users. It tends to **increase** when **number of DPD increases**.

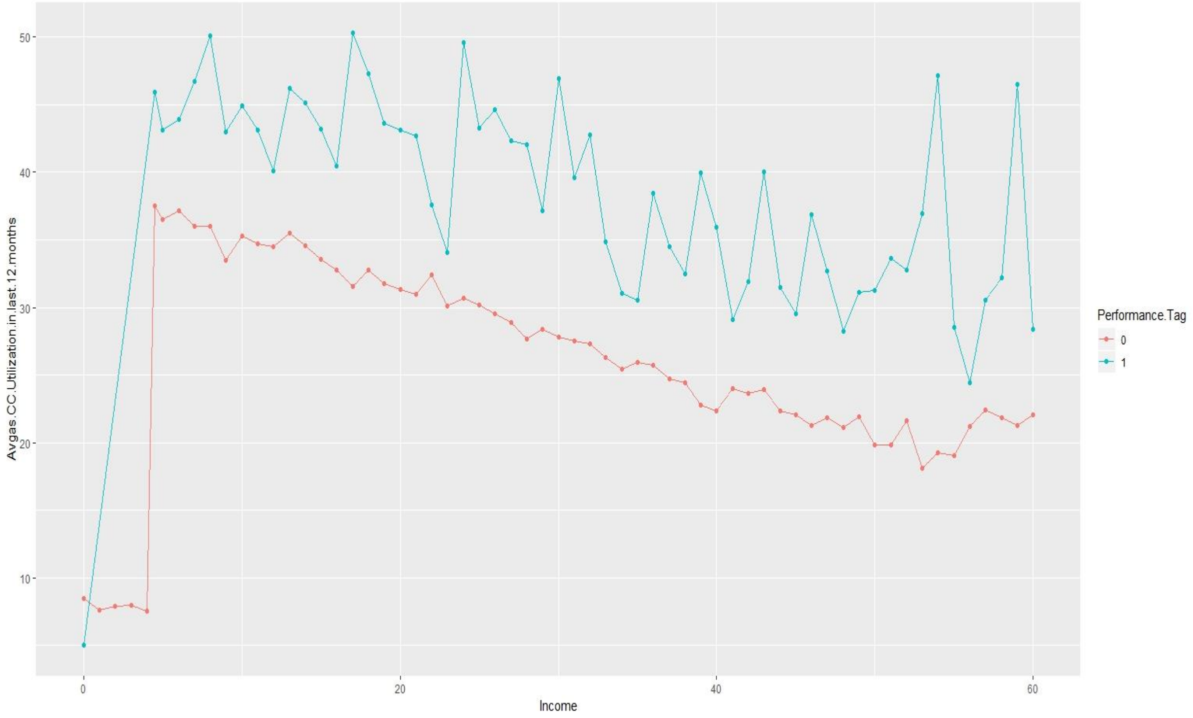
Analysis of Credit Card Default – Bivariate Analysis w.r.t. Outstanding Balance

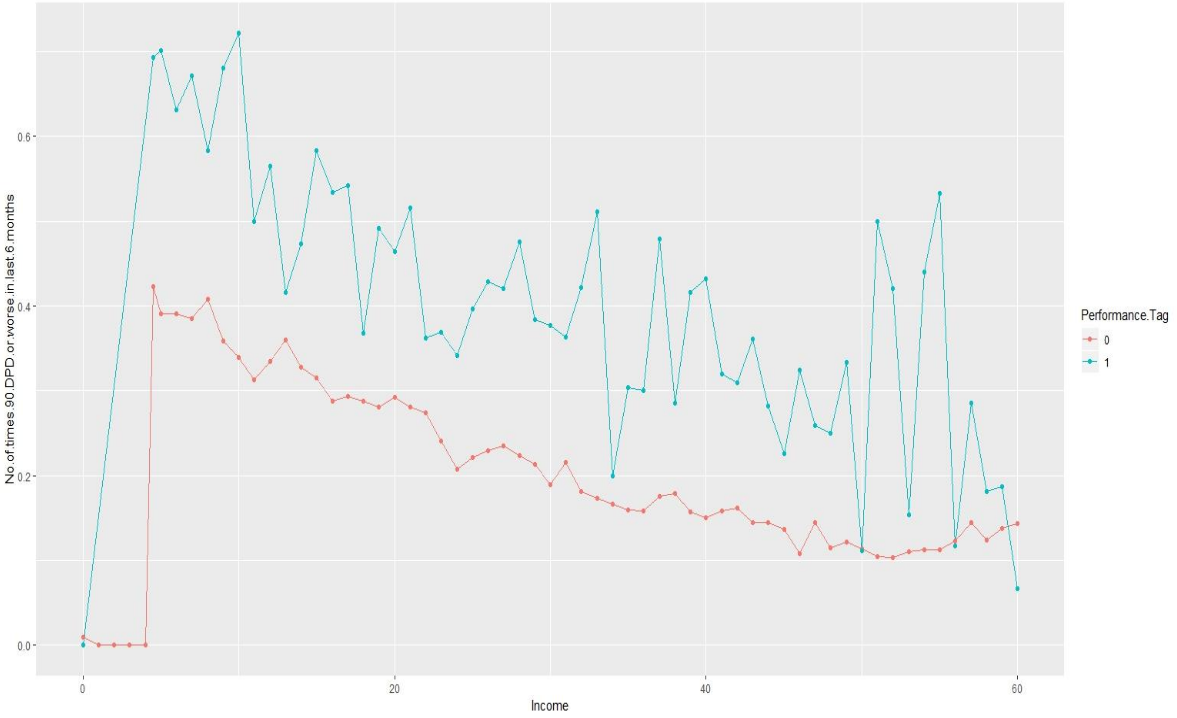




From the graphs, it is clear that **total number of trades** is **relatively higher** for defaulting users.
The **outstanding balance tends to be higher** for defaulting users. It tends to **increase** when **income increases** for default users.

Analysis of Credit Card Default – Bivariate Analysis w.r.t. Income

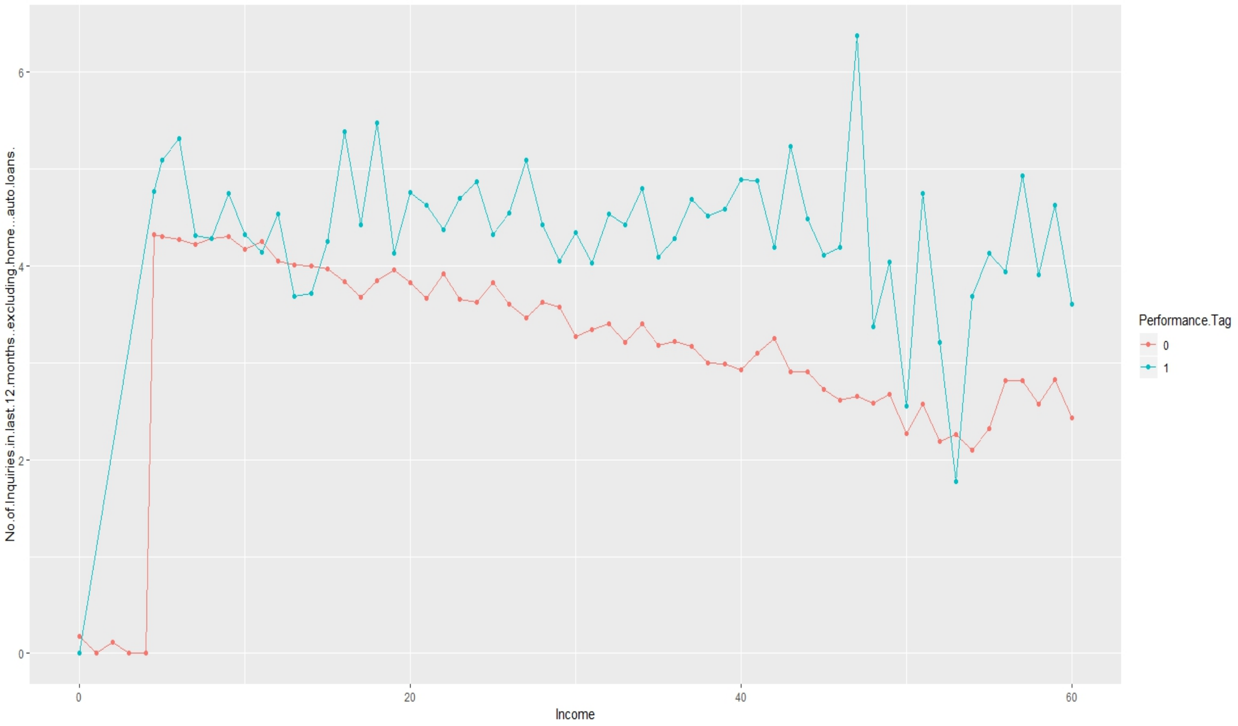


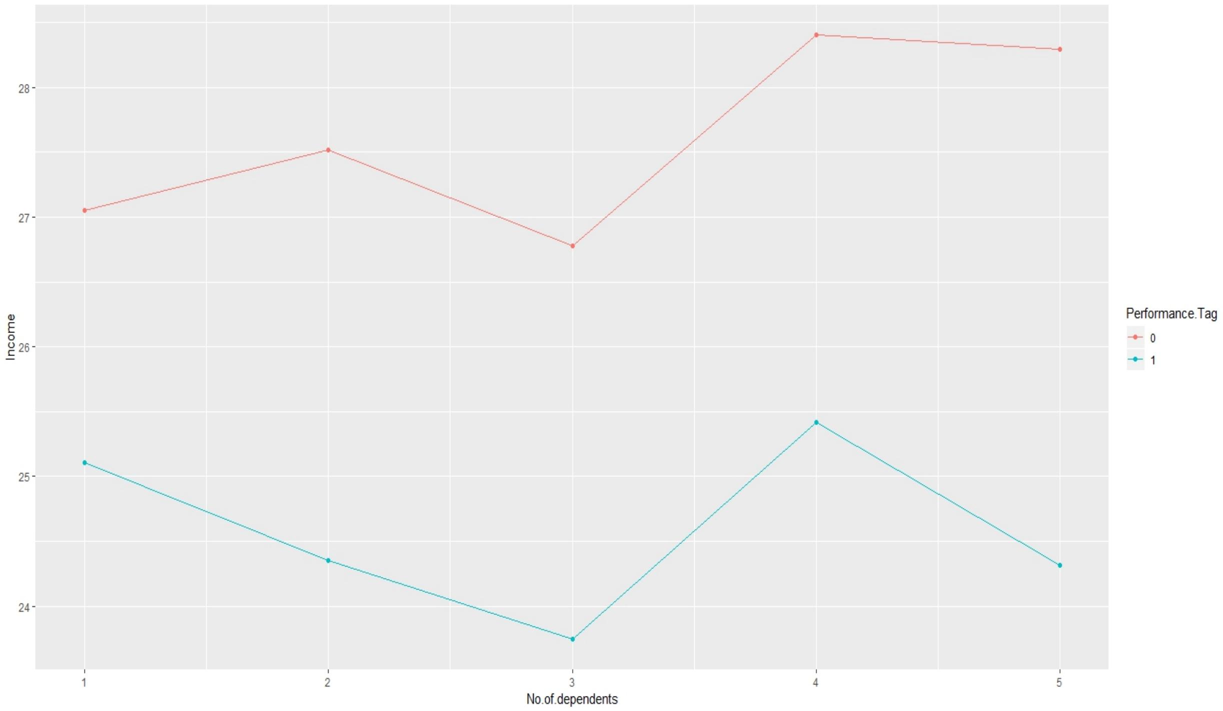


From the graphs, it is clear that as **income increases, average credit card utilization tends to increase** for defaulting users. If **average credit card utilization** is >40 for a low income, >30 for middle income, >25 for higher income, they should be looked at.

Also, as **income increases, number of DPDs tends to be way higher.** High number of defaulting users are in **lower to medium income range**.

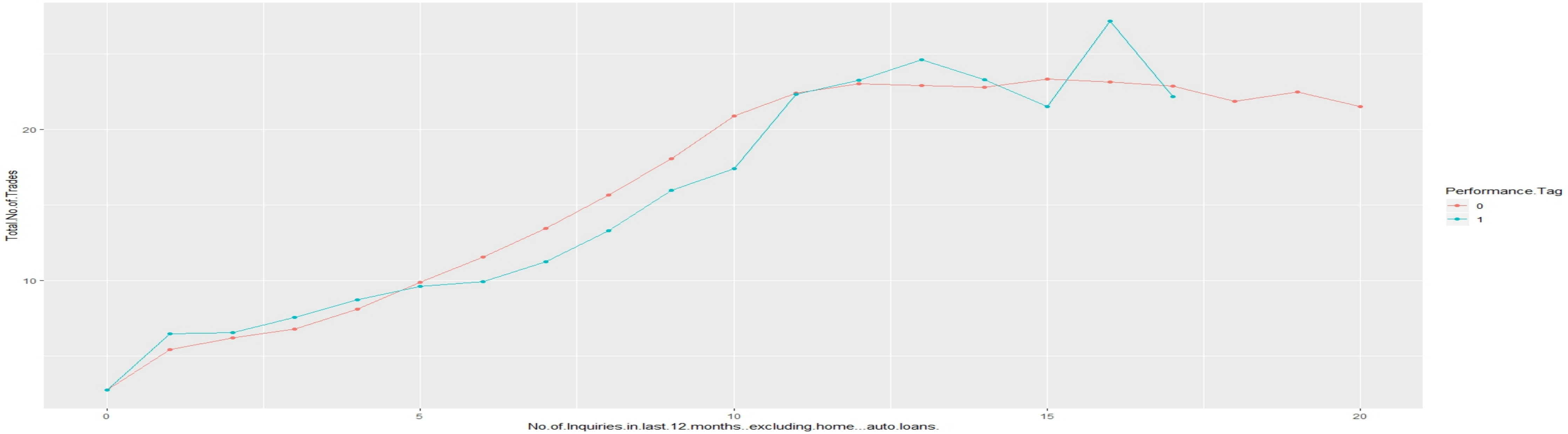
Analysis of Credit Card Default – Bivariate Analysis w.r.t. Income





From the graphs, it is clear that as **income increases, number of inquiries tends to increase** for defaulting users. Also, **income per number of dependents is much lower** for defaulting users.

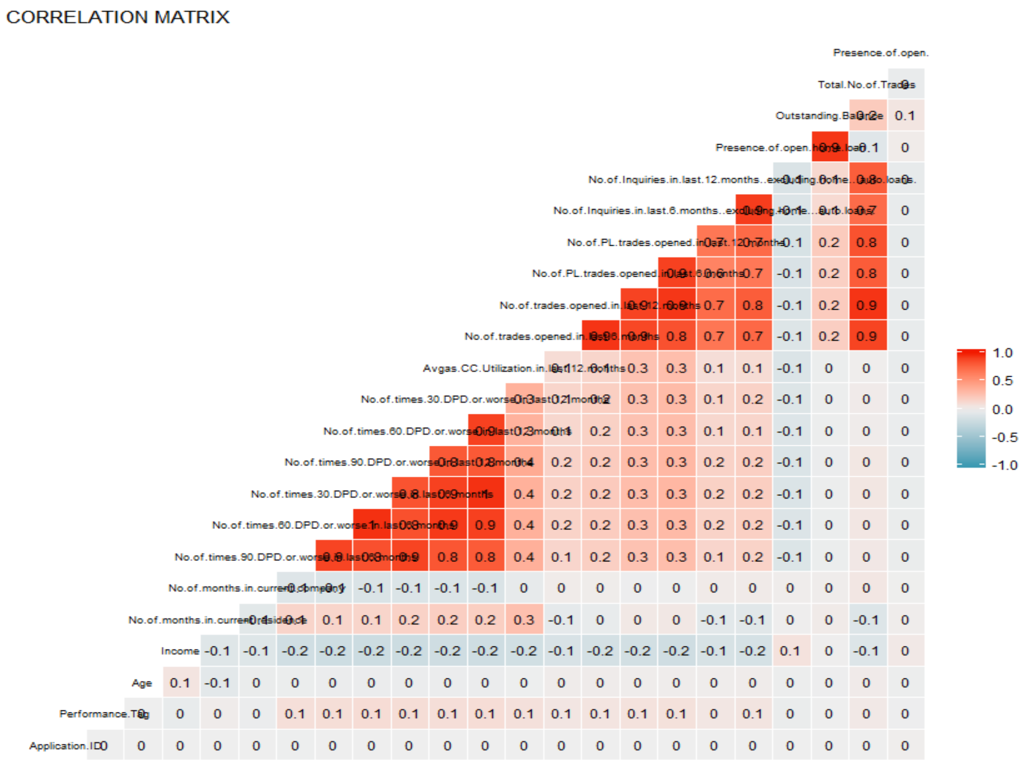
Analysis of Credit Card Default – Bivariate Analysis w.r.t. Number of Inquiries



From the graphs, it is clear that the **total number of trades tend to be higher** for defaulting users.
Also, as **number of inquiries increases**, **total no of trades increases**, then **gradually tends to become constant**.

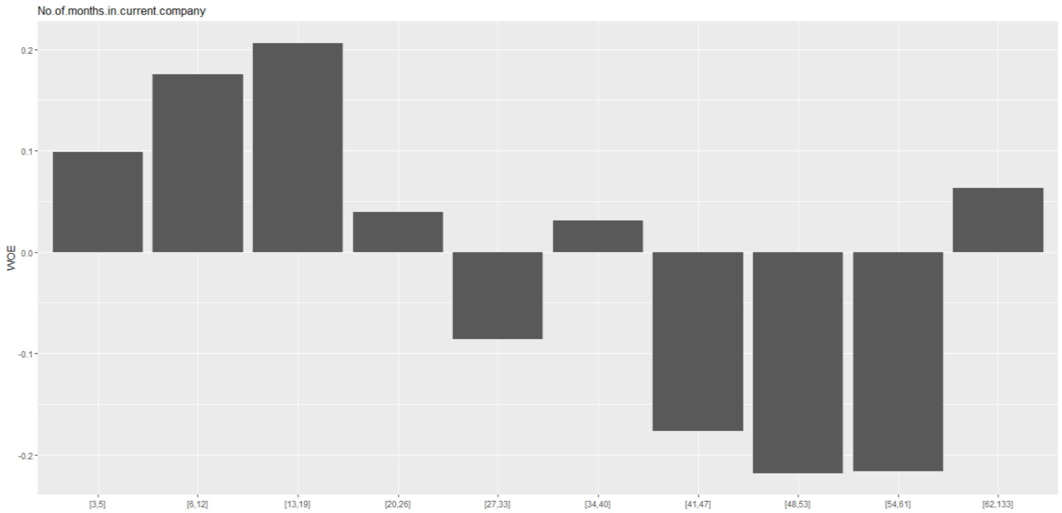
From the bivariate analysis and the resulting graphs obtained, it is clear that **credit factors like outstanding balance, total number of trades and average credit card utilization** have **a greater influence over the risk of default** as compared to demographic factors.

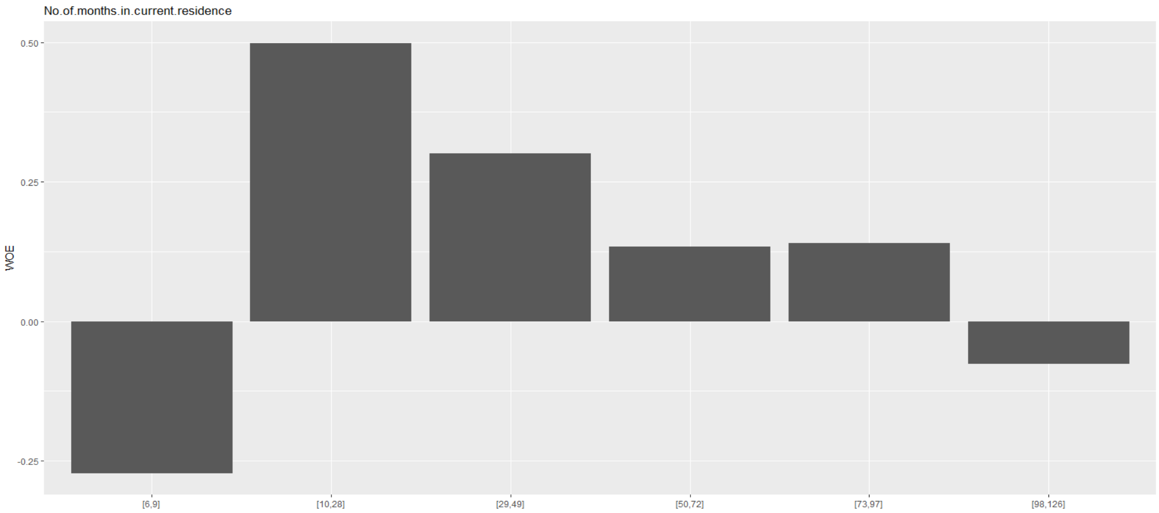
Multivariate Analysis using Correlation Matrix

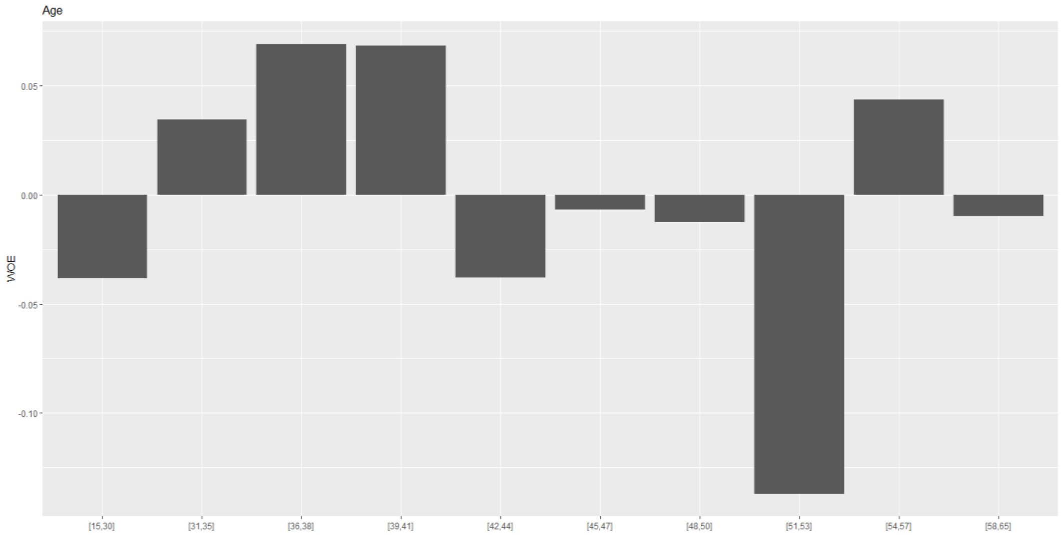


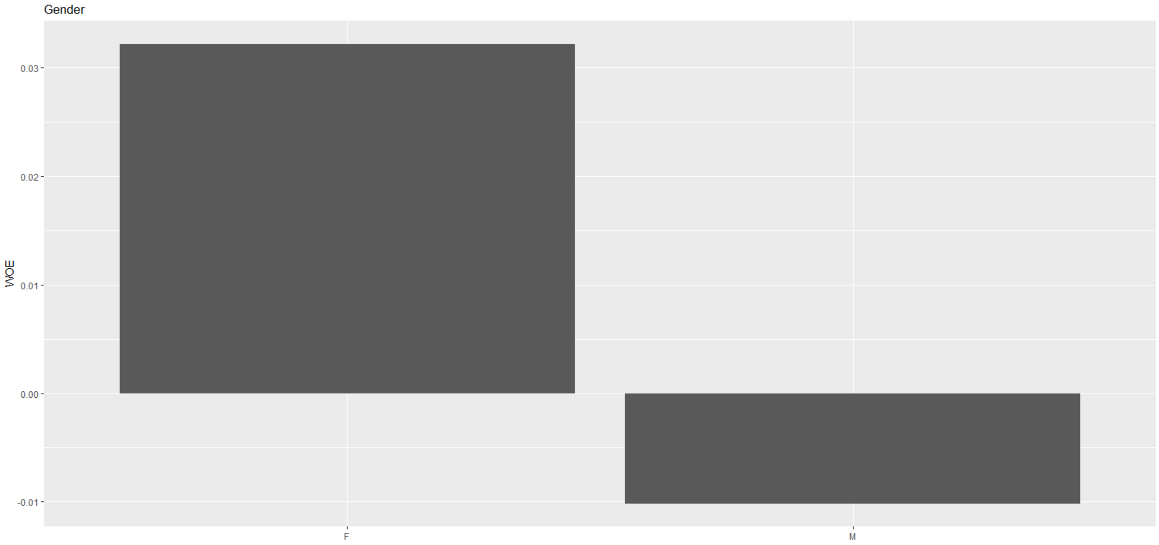
From the correlation matrix it is clear that **credit factors like total number of trades and average credit card utilization have a greater influence on rate of defaulters** as compared to demographic factors in the correlation matrix.

WOE / IV analysis for Demographic Factors

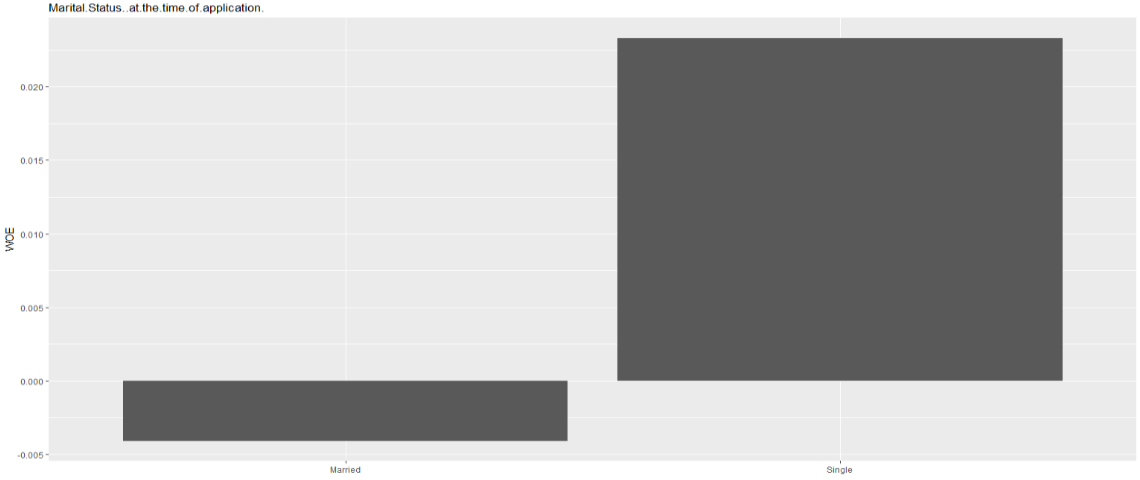




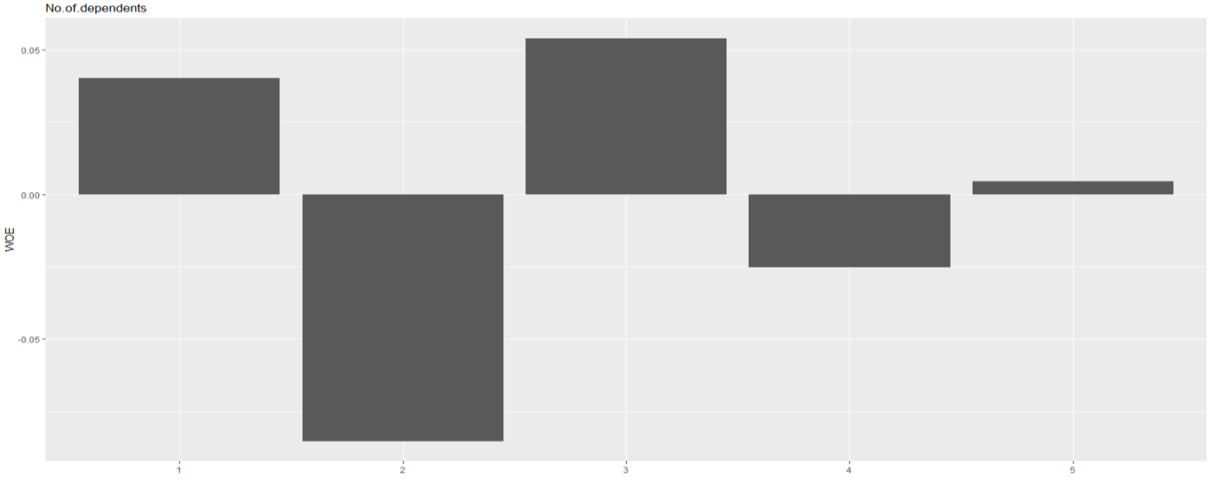




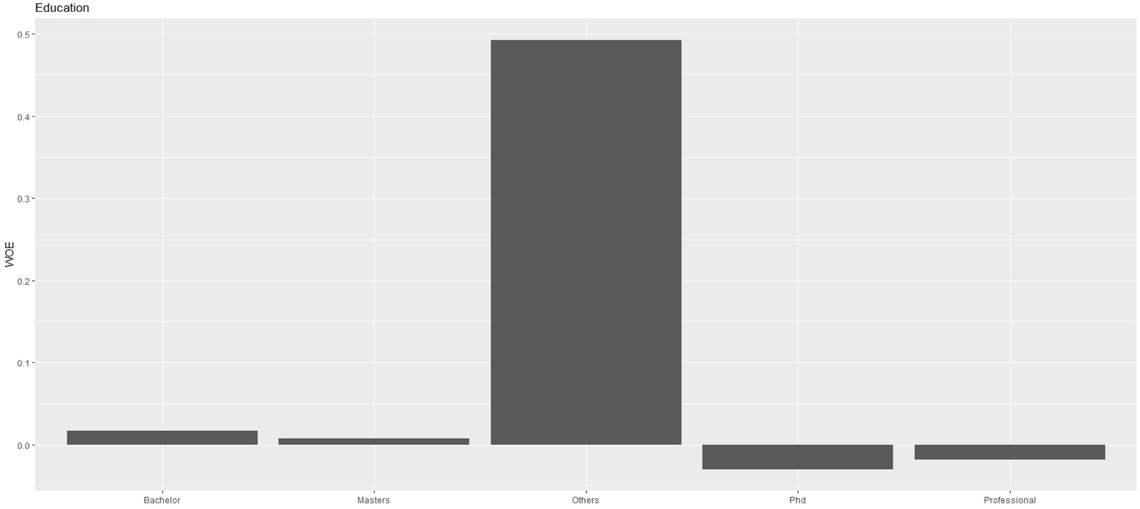
WOE / IV analysis for Demographic Factors



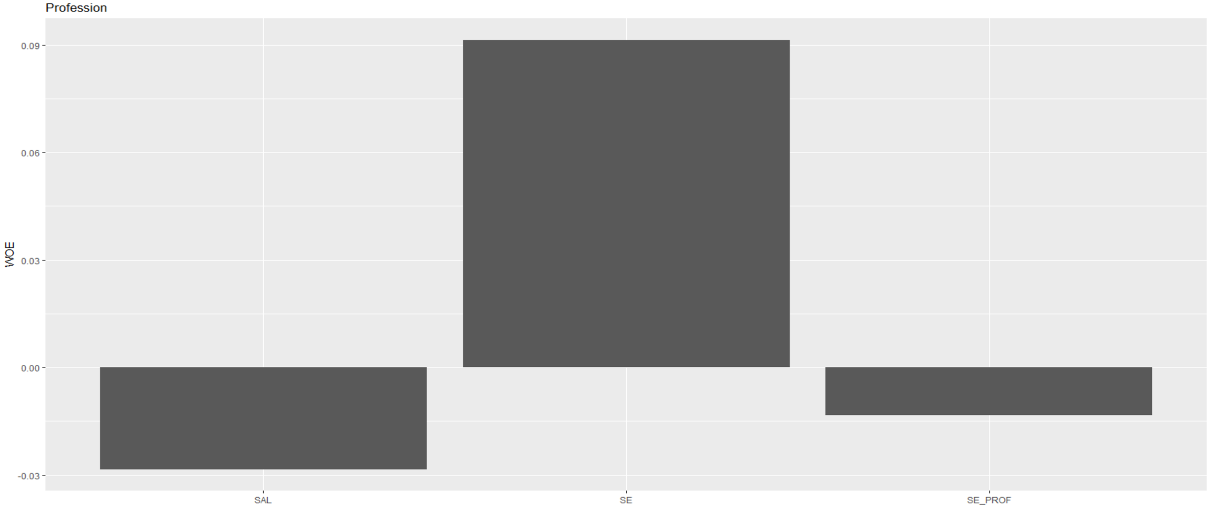
Marital.Status..at.the.time.of.application.



No.of.dependents

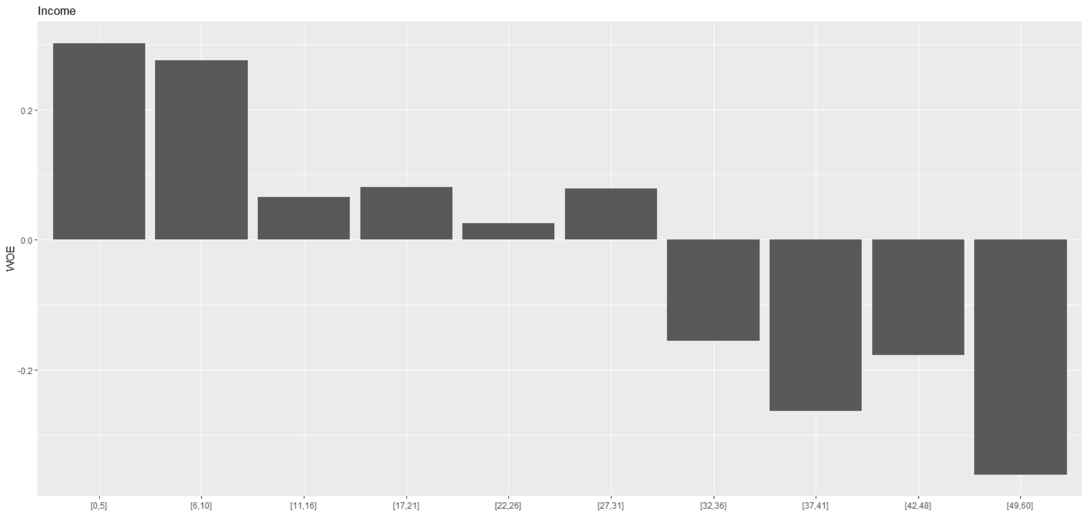


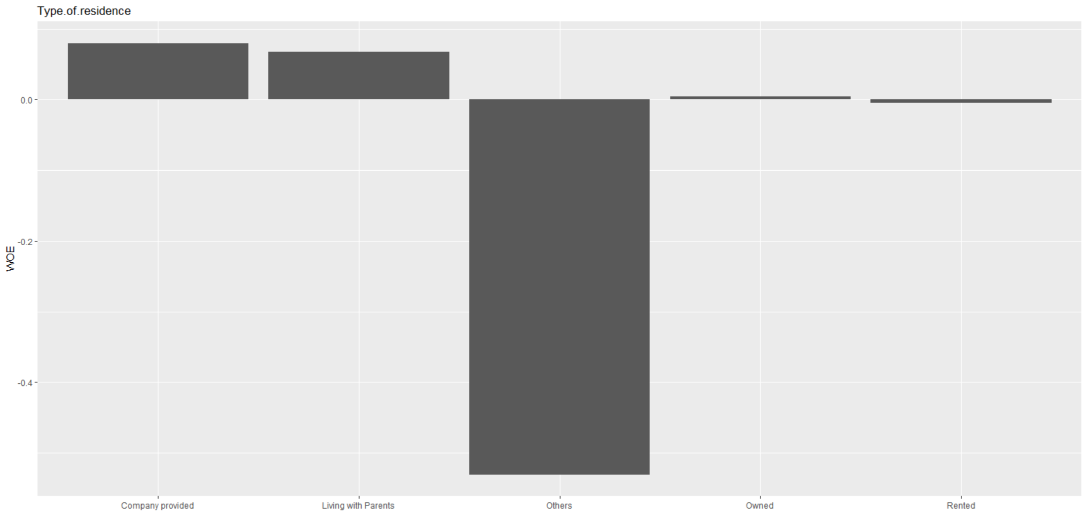
Education



Profession

WOE / IV analysis for Demographic Factors



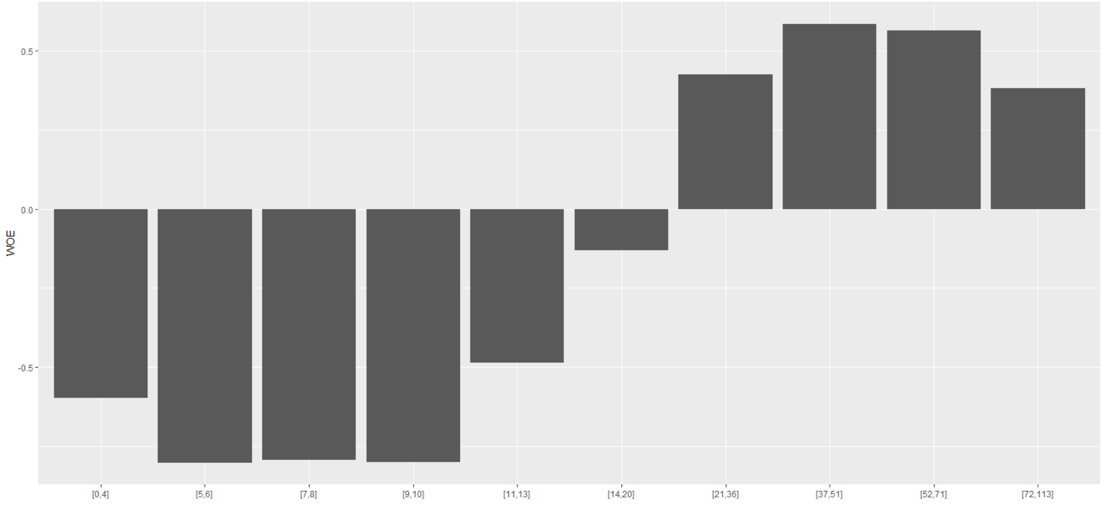


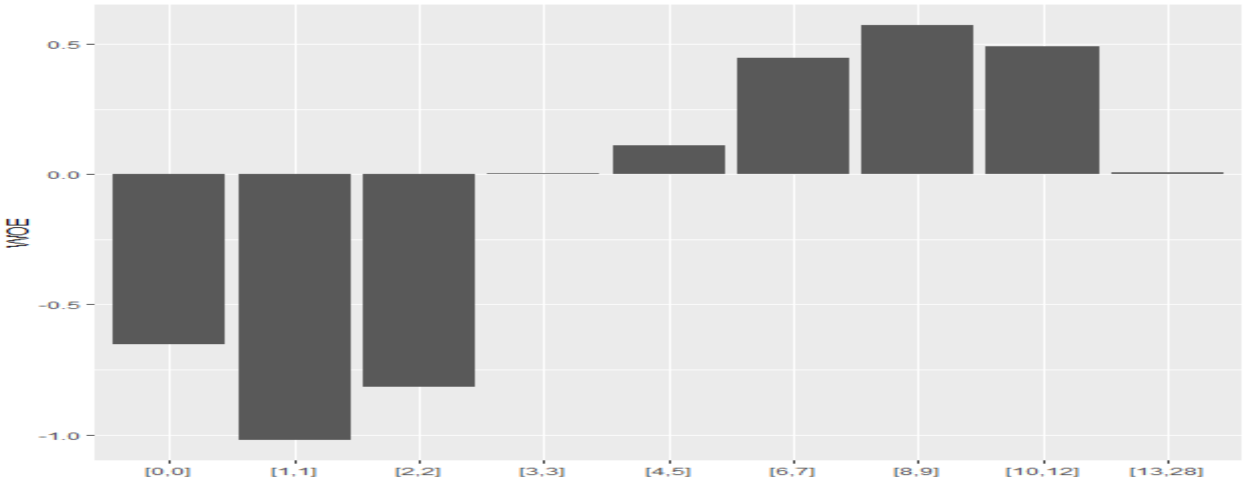
WOE/IV Value	Significance
< 0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Suspicious or too good

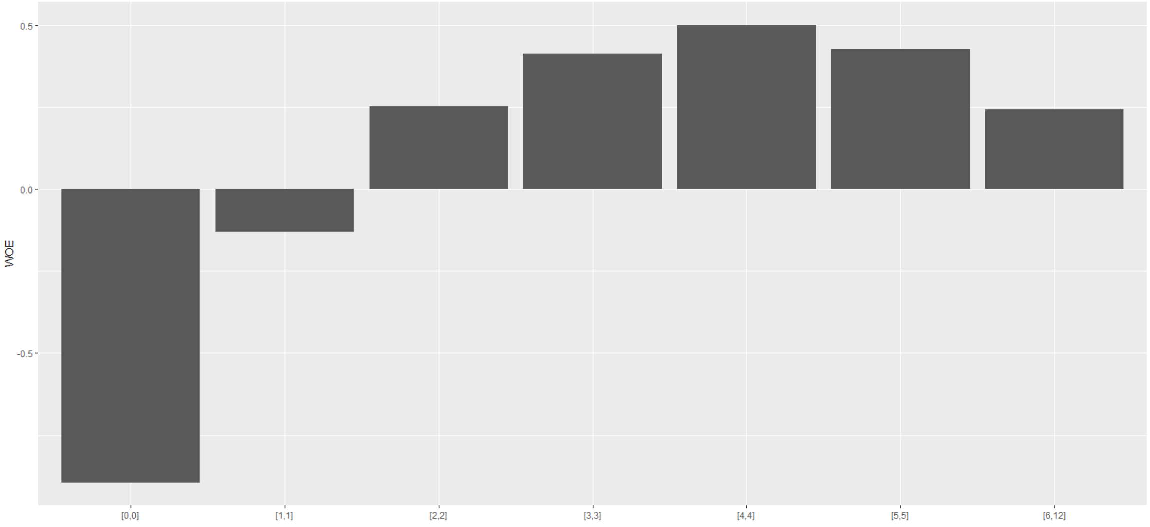
NOTE: The table to the left shows the significance of WOE/IV on predictor variables.

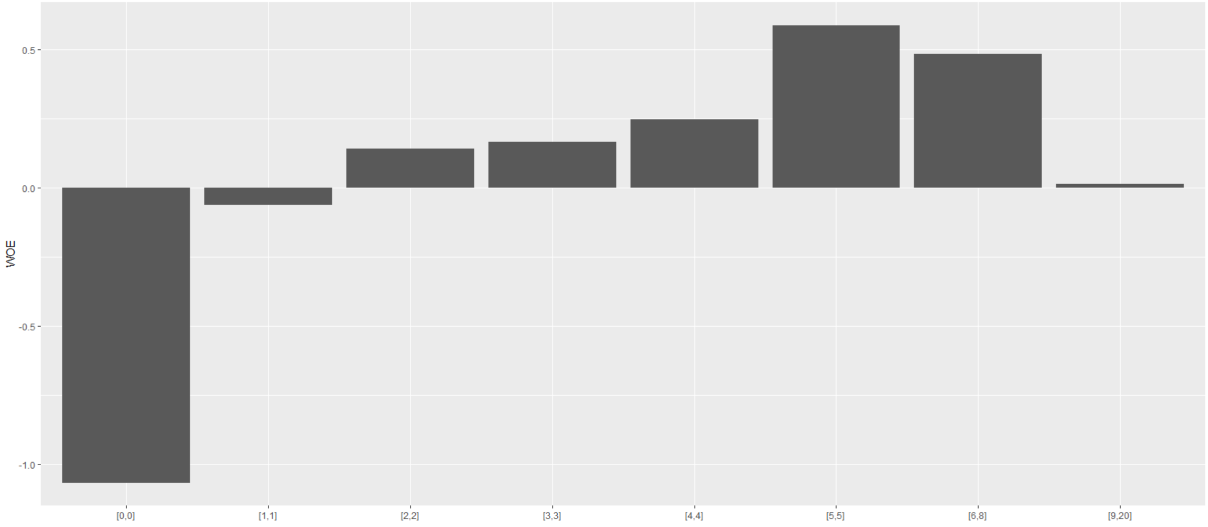
Thus, the graphs in slides 24, 25, 26 and 27 show the variation of predictor demographic variables w.r.t. various bins and WOE/IV values.

WOE/IV analysis for Credit Factors

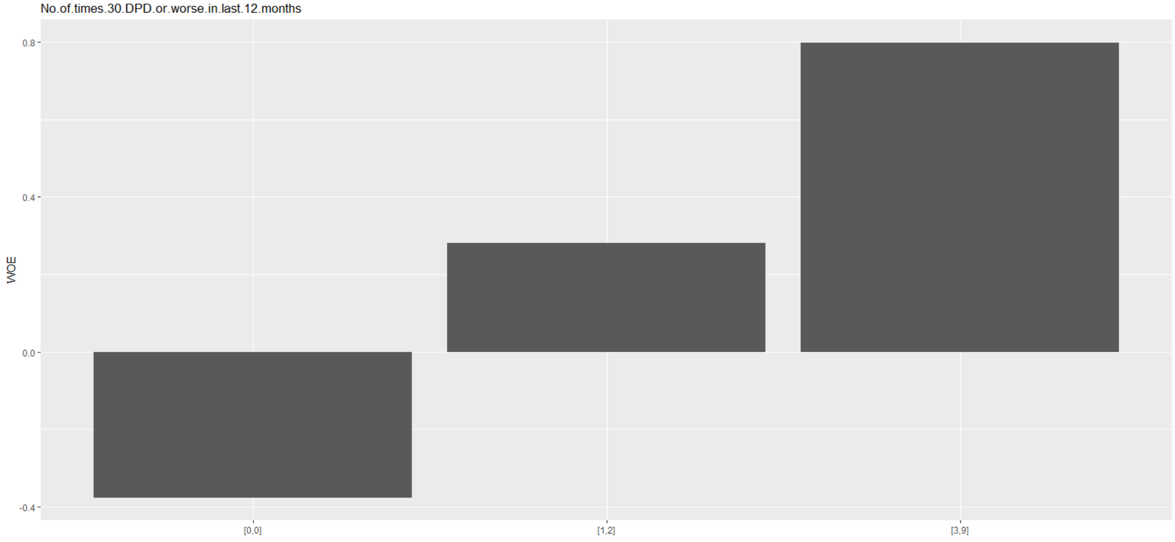

Avgas.CC.Utilization.in.last.12.months

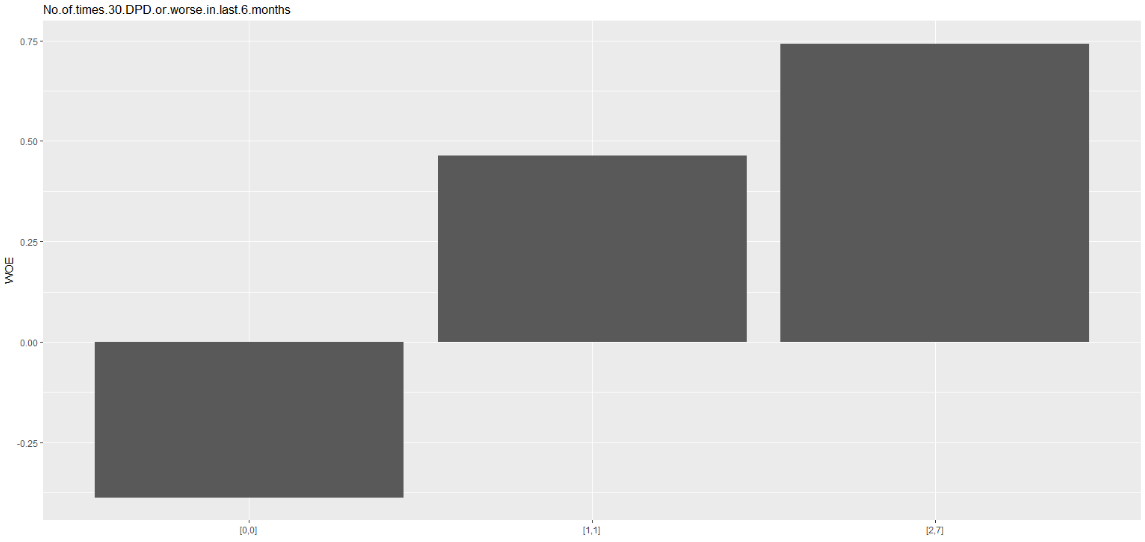

No.of.trades.opened.in.last.12.months

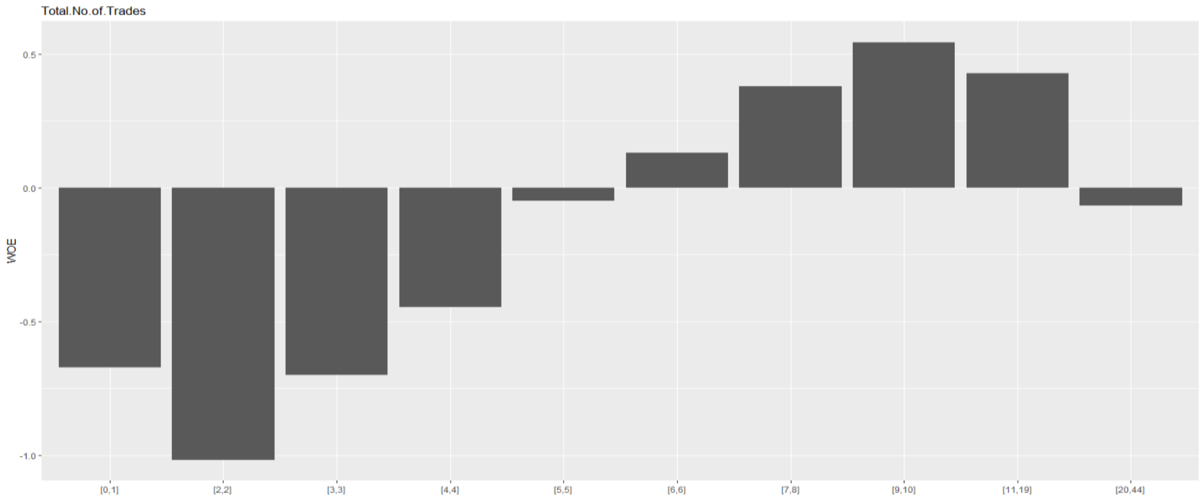

No.of.PL.trades.opened.in.last.12.months

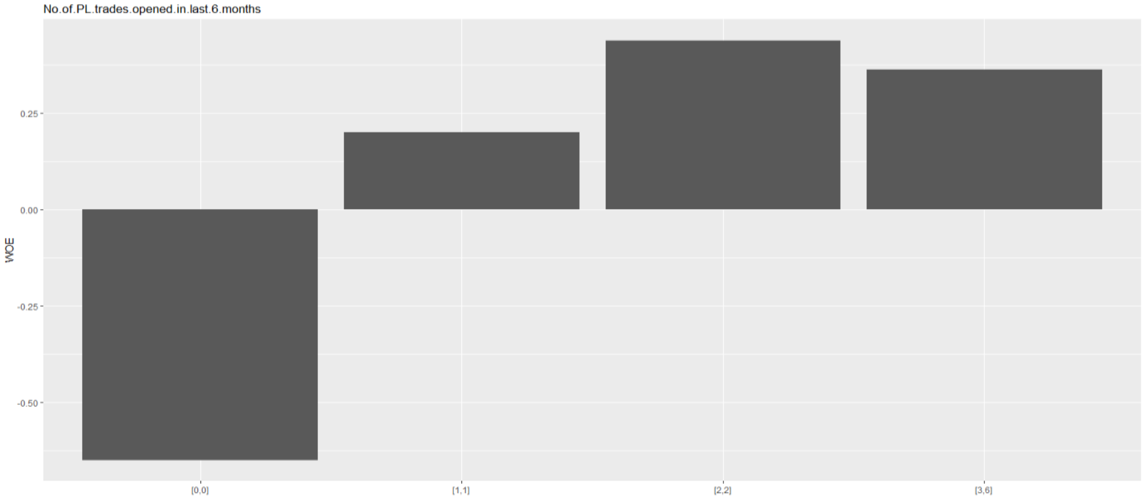

No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.

WOE/IV analysis for Credit Factors

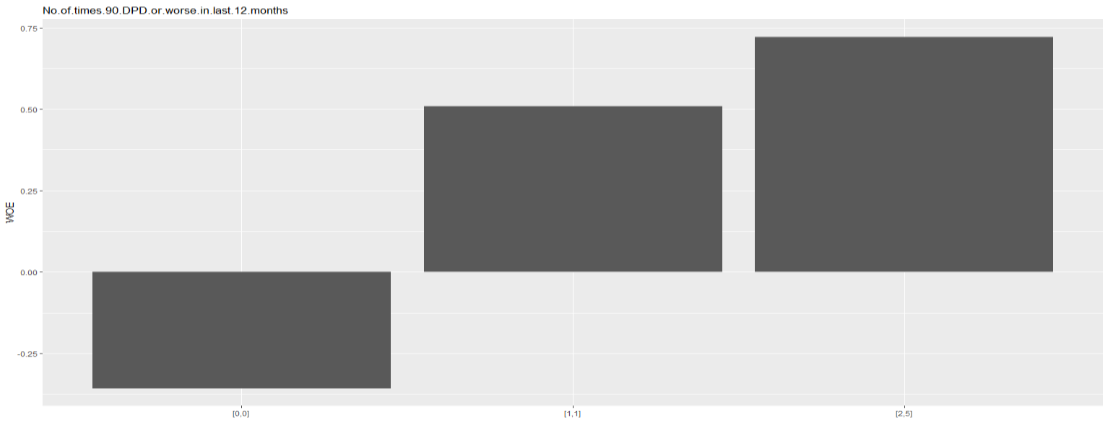


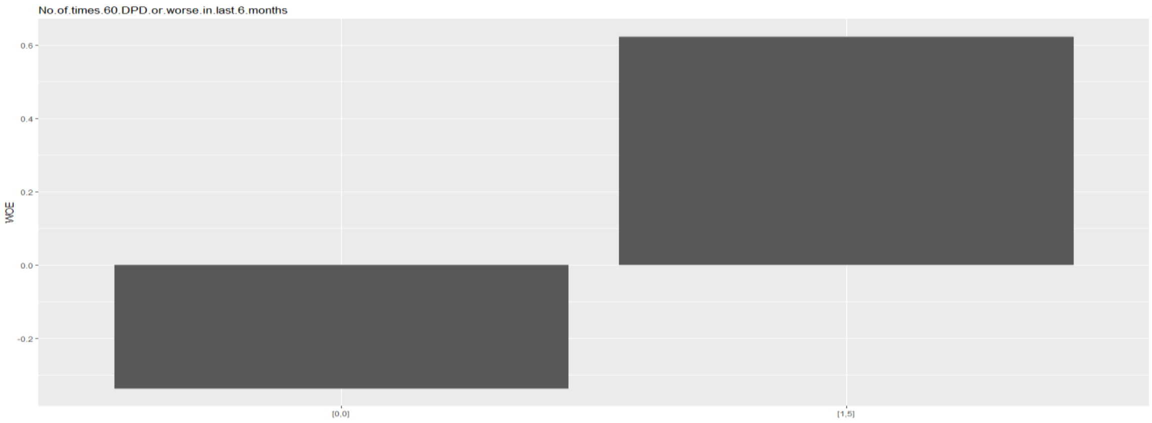


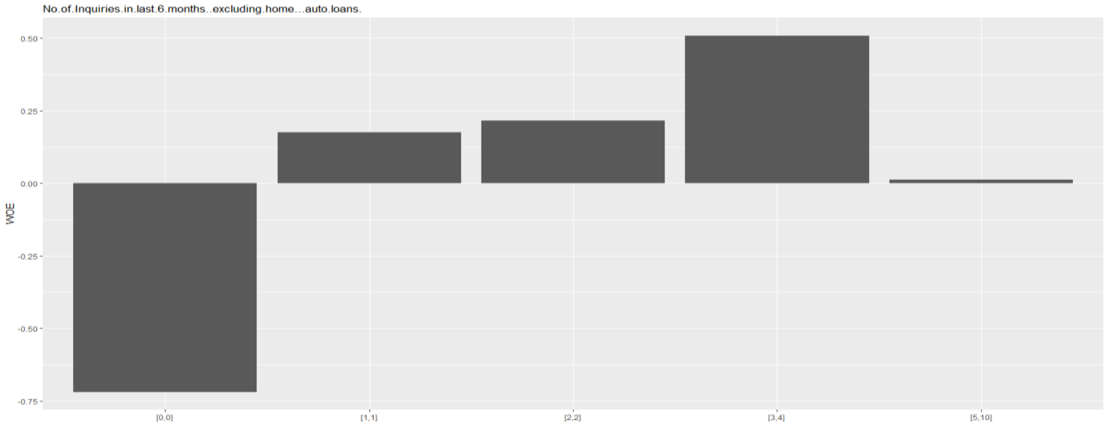


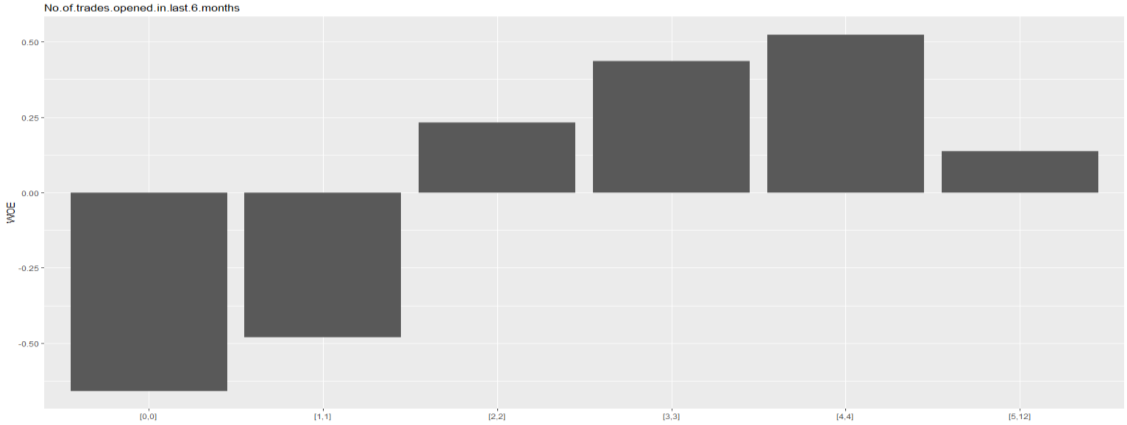


WOE/IV analysis for Credit Factors

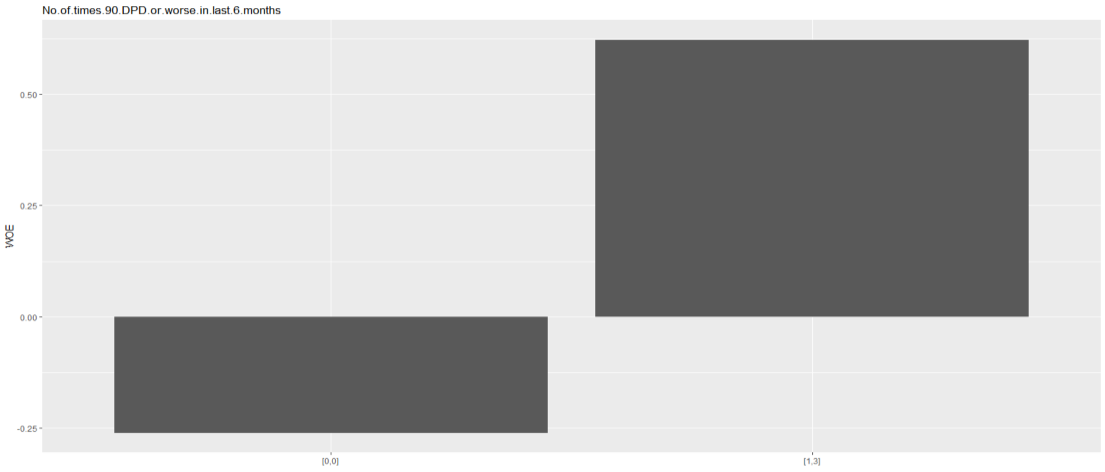


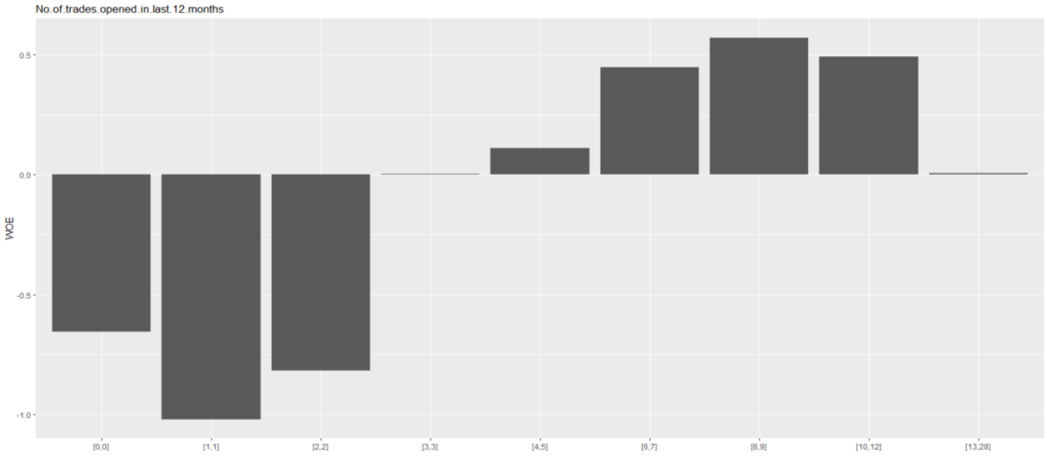


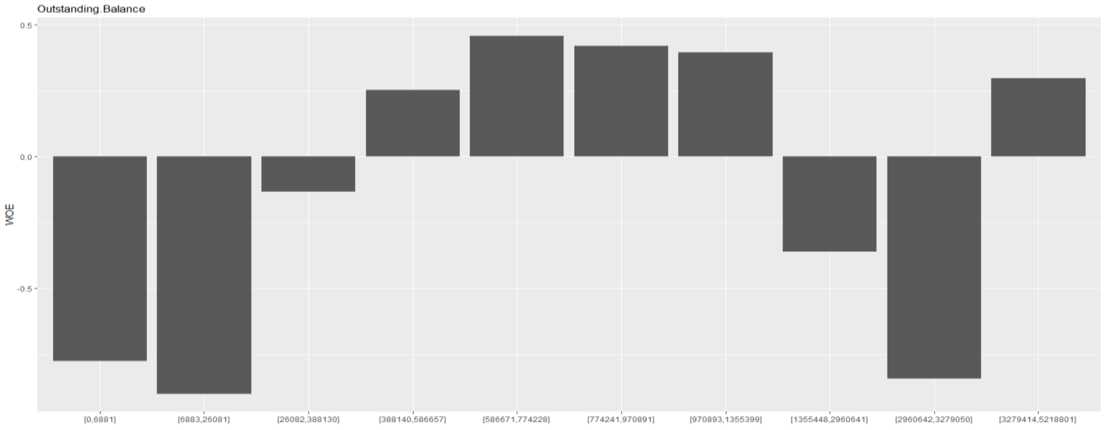


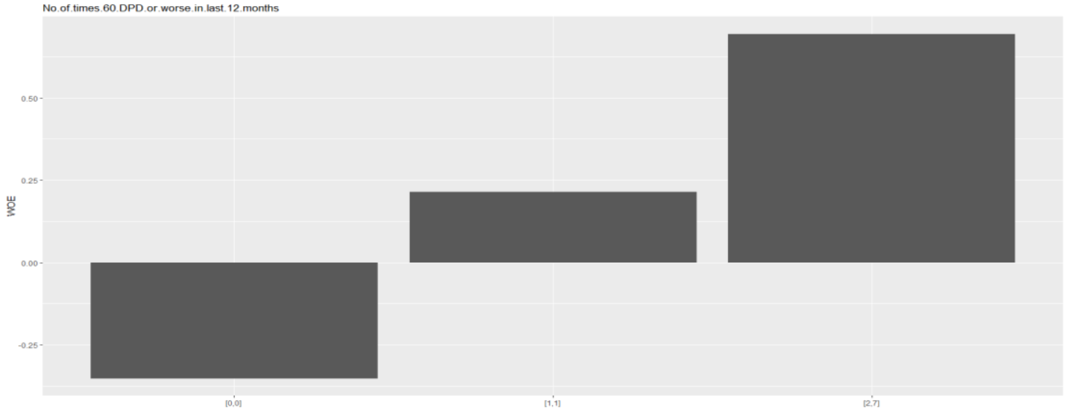


WOE/IV analysis for Credit Factors

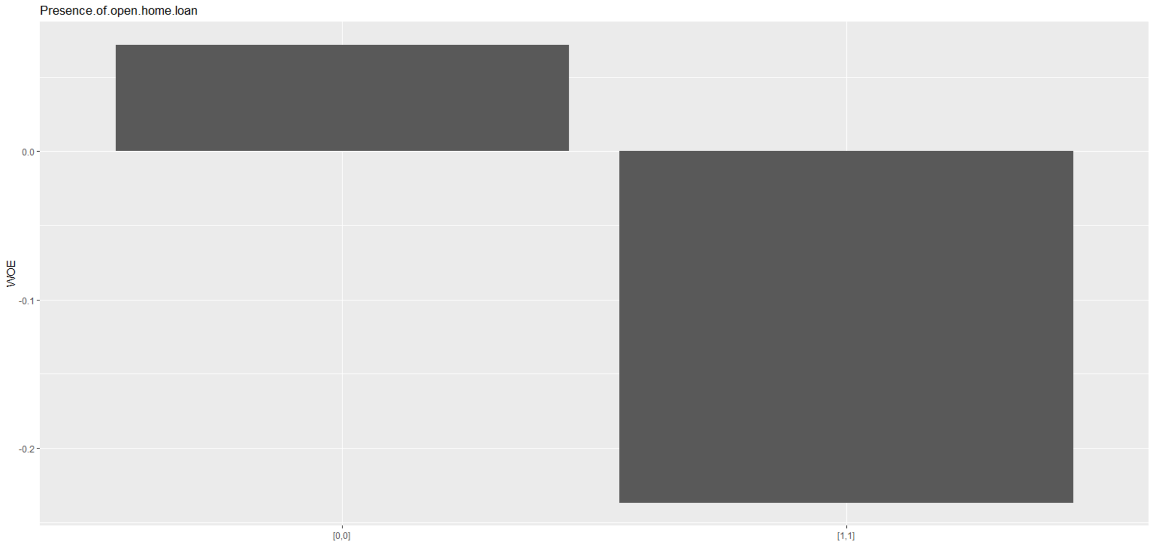


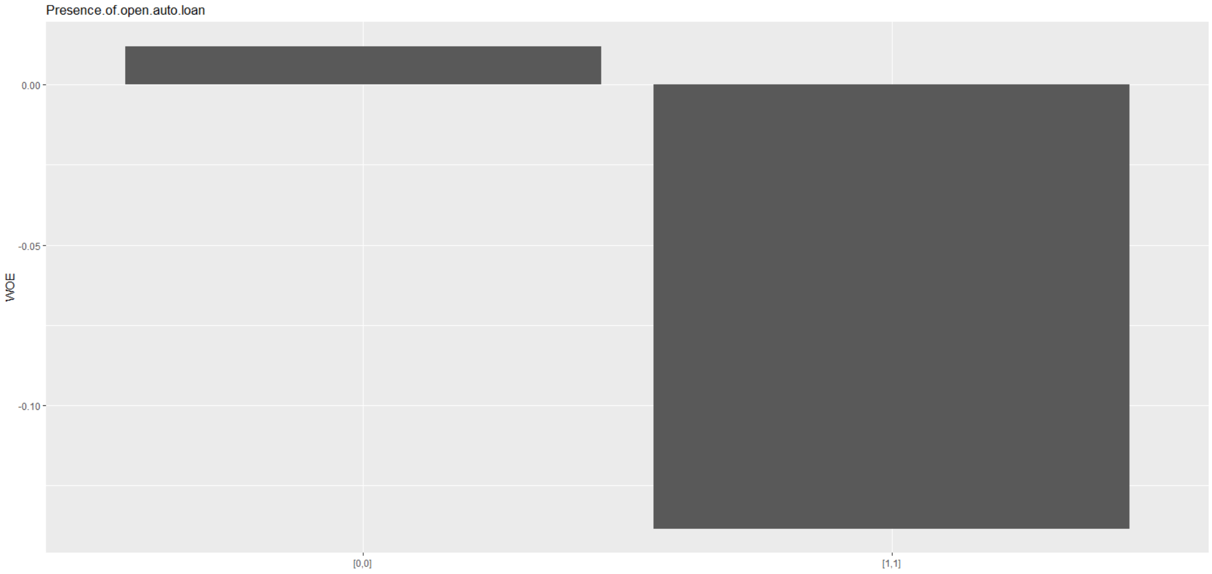






WOE/IV analysis for Credit Factors





WOE/IV Value	Significance
< 0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Suspicious or too good

NOTE: The table to the left shows the significance of WOE/IV on predictor variables.

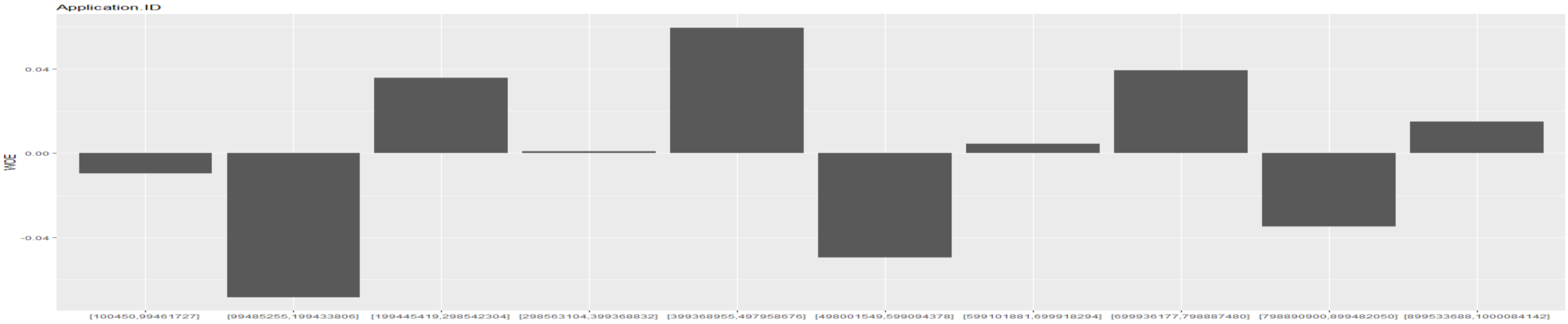
Thus, the graphs in slides 28, 29, 30, 31 and 32 show the variation of predictor credit variables w.r.t. various bins and WOE/IV values.

WOE/IV Analysis

- The **top 10 predictor variables** obtained according to WOE/IV analysis are:
  - Avgas.CC.Utilization.in.last.12.months
  - No.of.trades.opened.in.last.12.months
  - No.of.PL.trades.opened.in.last.12.months
  - No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
  - No.of.times.30.DPD.or.worse.in.last.12.months
  - No.of.times.30.DPD.or.worse.in.last.6.months
  - Total.No.of.Trades
  - No.of.PL.trades.opened.in.last.6.months
  - No.of.times.90.DPD.or.worse.in.last.12.months
  - No.of.times.60.DPD.or.worse.in.last.6.months

WOE/IV Analysis

- This shows that **credit factors comparatively have a greater influence over rate of defaulters over demographic factors** in the EDA as well as in the WOE Analysis.
- **Application.ID** as is obvious and known to us has zero predictive power and is also **highly non-monotonic** as shown in the graph below:



NOTE: The predictive power of individual variables are dependent heavily on the bins that have been considered while implementing WOE IV in our dataset. A few bins/clusters/groups in a particular variable are found to be more helpful then others.

Procedure Used in Model Building

The following diagram illustrates the methodology used in this project.

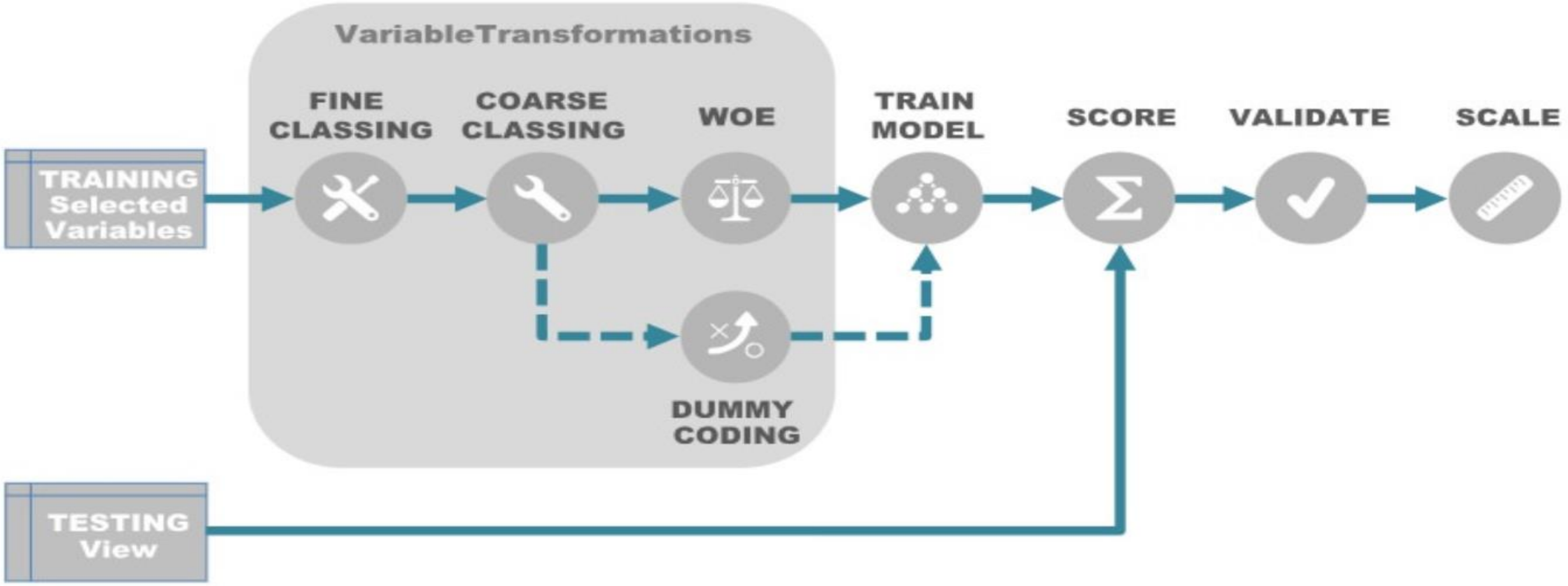


Figure 1. Standard Scorecard Development Process

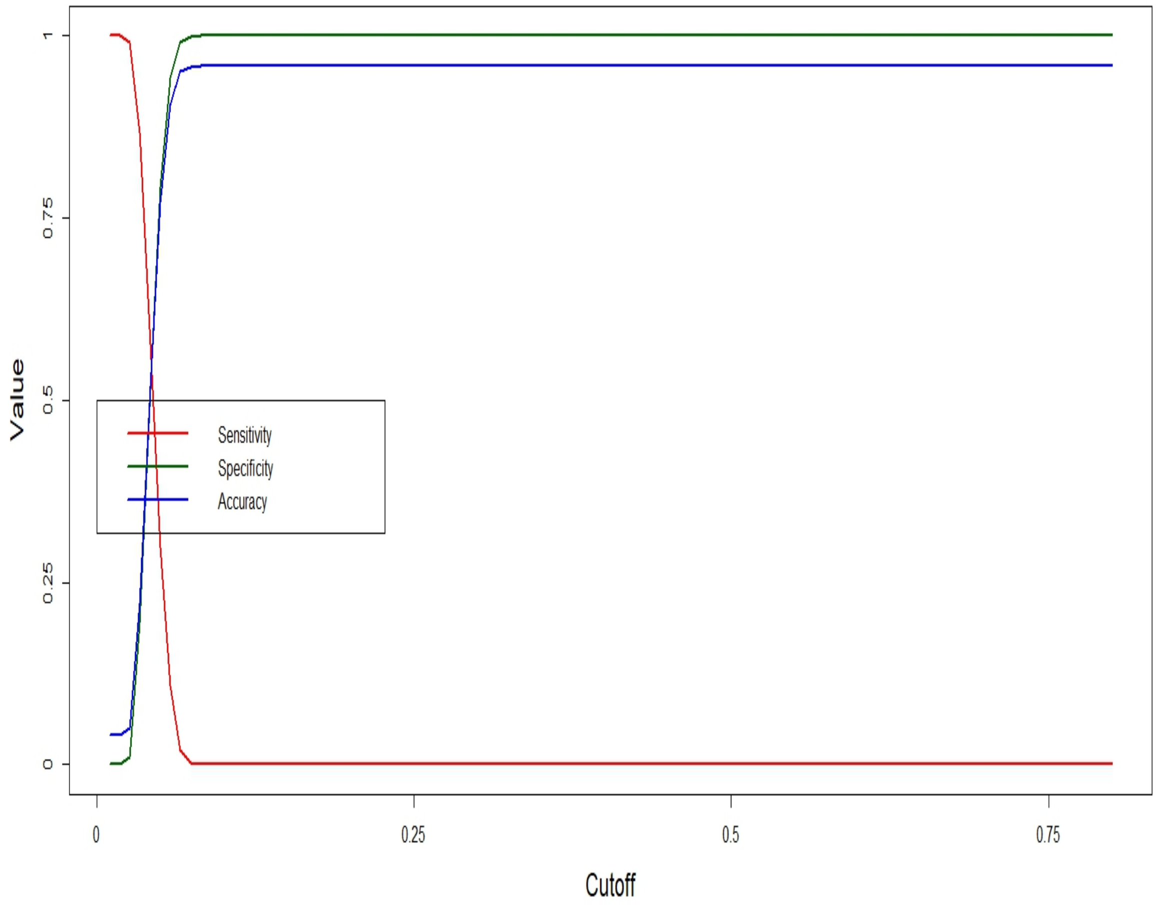
Model Building

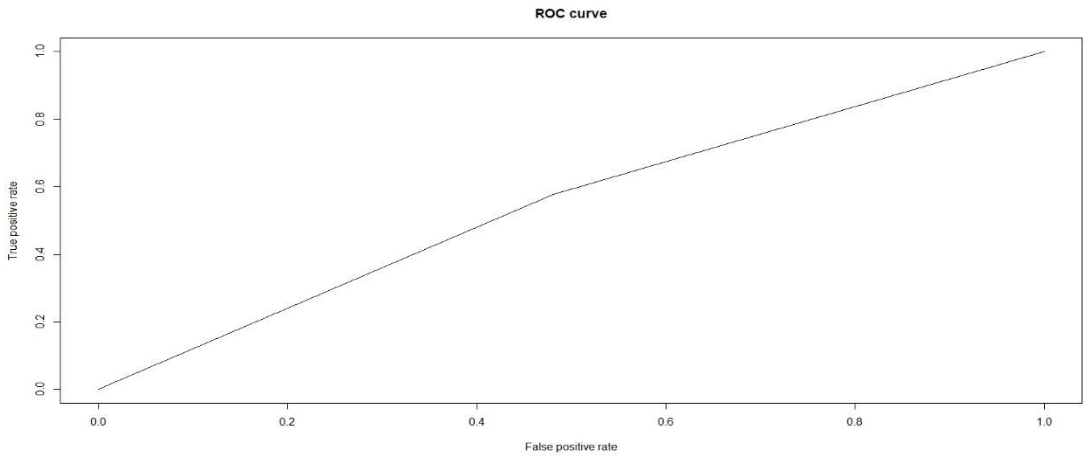
- Considering the classification problem of dividing the applicants in two categories based on the performance tag, Defaulters and Non Defaulters, we can use the following different models for the combined demographic and credit data.
  - Logistic Regression
  - Logistic Regression (with SMOT)
  - Random Forest
  - Random Forest (with SMOT)
- **SVM is not used** as the **amount of data is quite large**.
- The data in each of the models is seperated into test and train sets using **70:30 train:test ratio** before the actual model building processes.
- Thus, six models are built in total (two using only the demographic data and four using the entire merged data).

Model Building

- Two logistic regression models are **built only using the demographic dataset** for reference and comparision purpose.
- However, the **final model** is based on the **combined dataset**.
- **Removing the non significant variables** in logistic regression model is done on the **basis of VIF and p-values**.
- In **random forest models** we need to **vary the number of trees, min number of buckets and min number of leaves in a node**.
- **SMOT (Synthetic Minority Oversampling Technique)** is used to provid**e better model results** and **handle class imbalance issues**.
- SMOT algorithm creates **artificial data based on feature space** (rather than data space) similarities from minority samples. **It generates a random set of minority class observations** to **shift the classifier learning bias towards minority class**.

Logistic Regression Model using Only Demographic Data (without using SMOT)

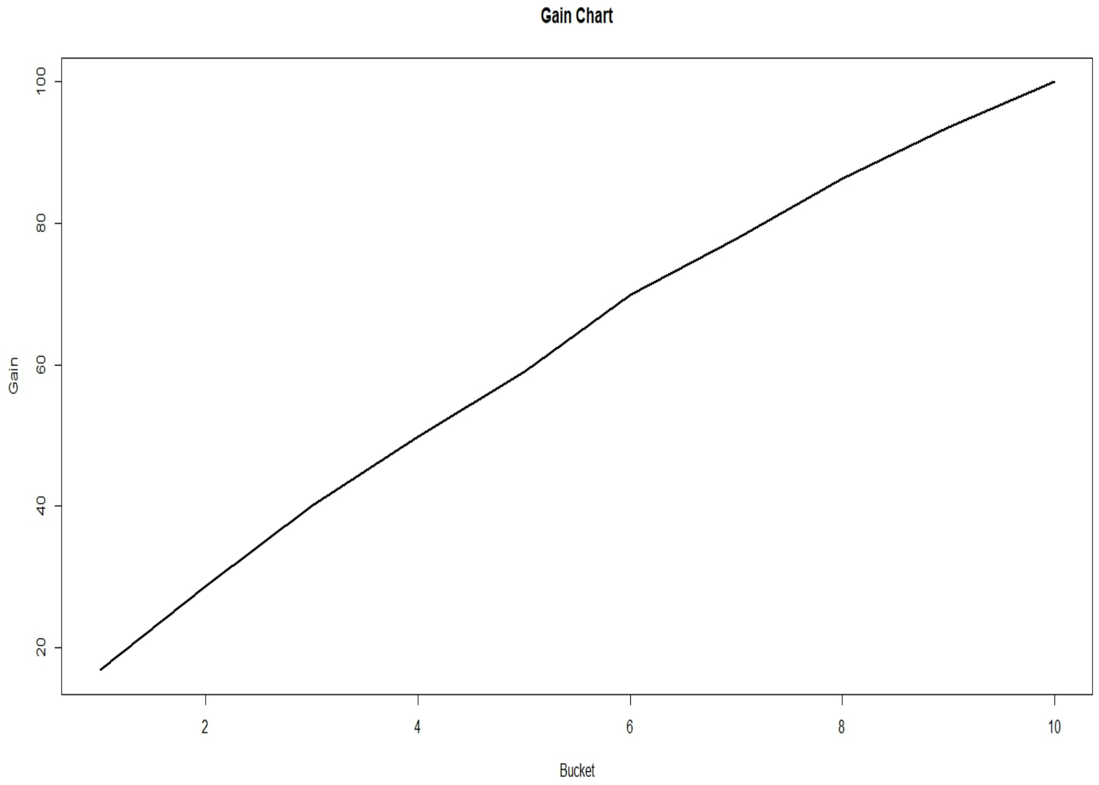


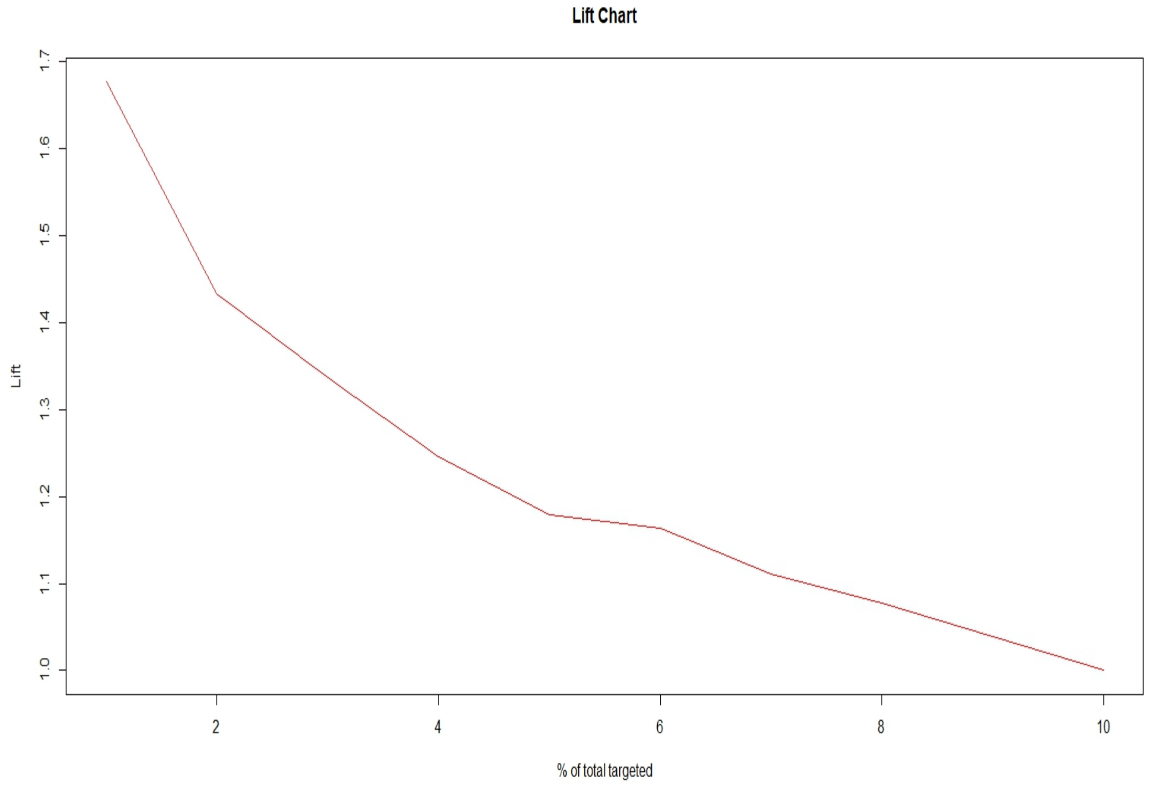


The following are the important predictor variables obtained using this model:

- Income
- No.of.months.in.current.residence
- No.of.months.in.current.company
- Profession.xSE

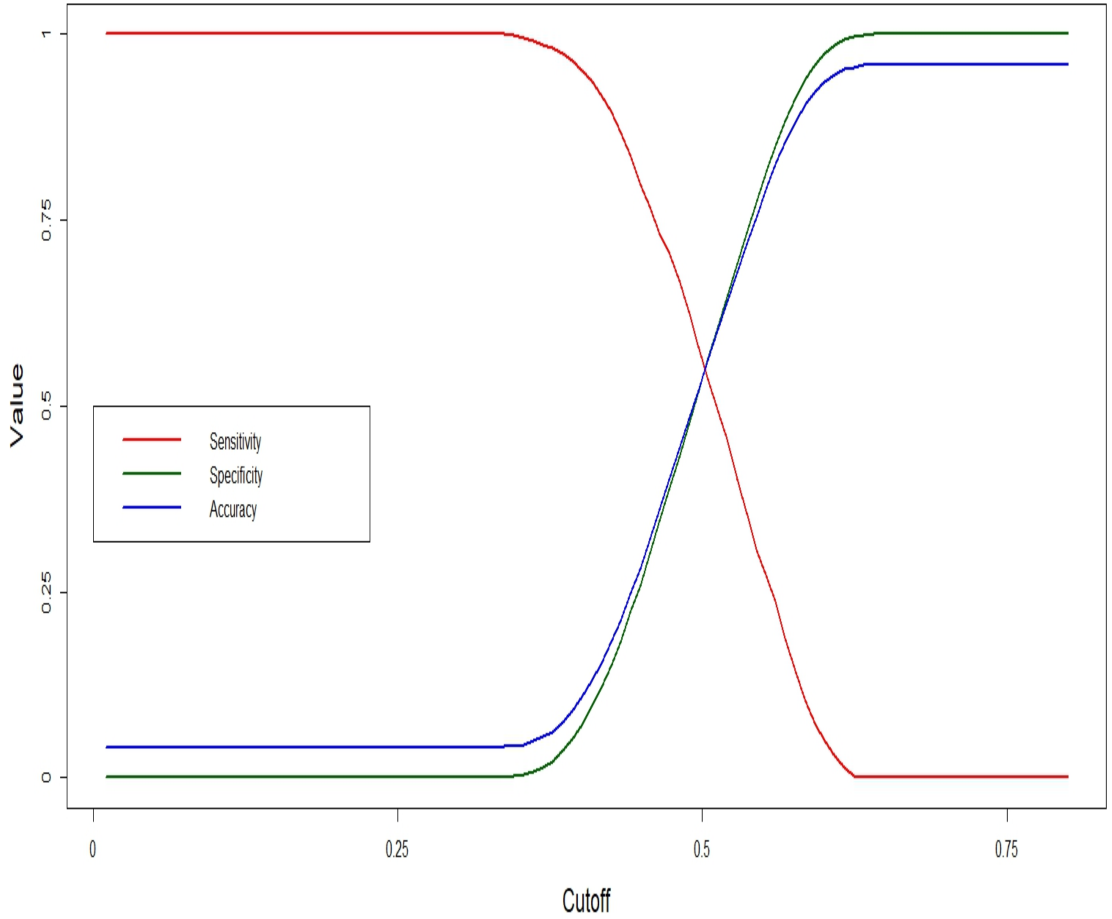
Logistic Regression Model using Only Demographic Data (without using SMOT)

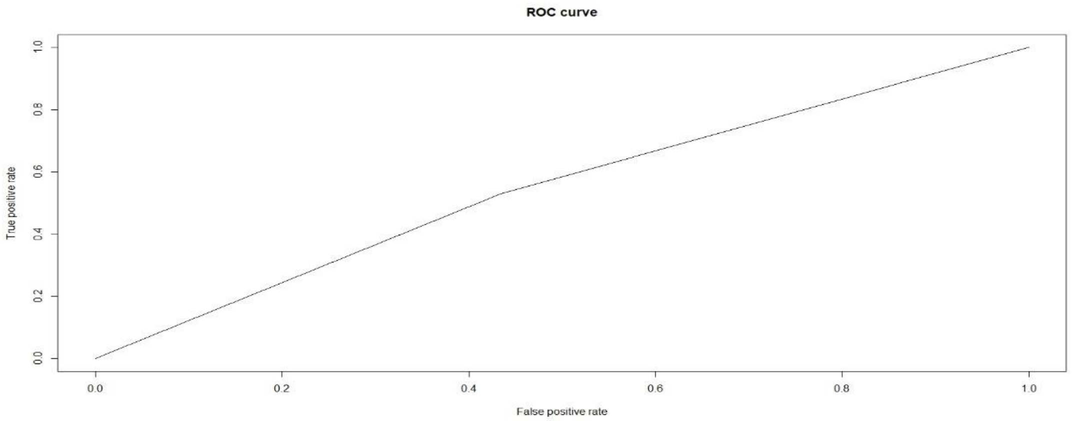




From the lift gain chart, it is clear that most of the values (70-75%) are accurately predicted by the 6th decile.

Logistic Regression Model using Only Demographic Data (using SMOT)

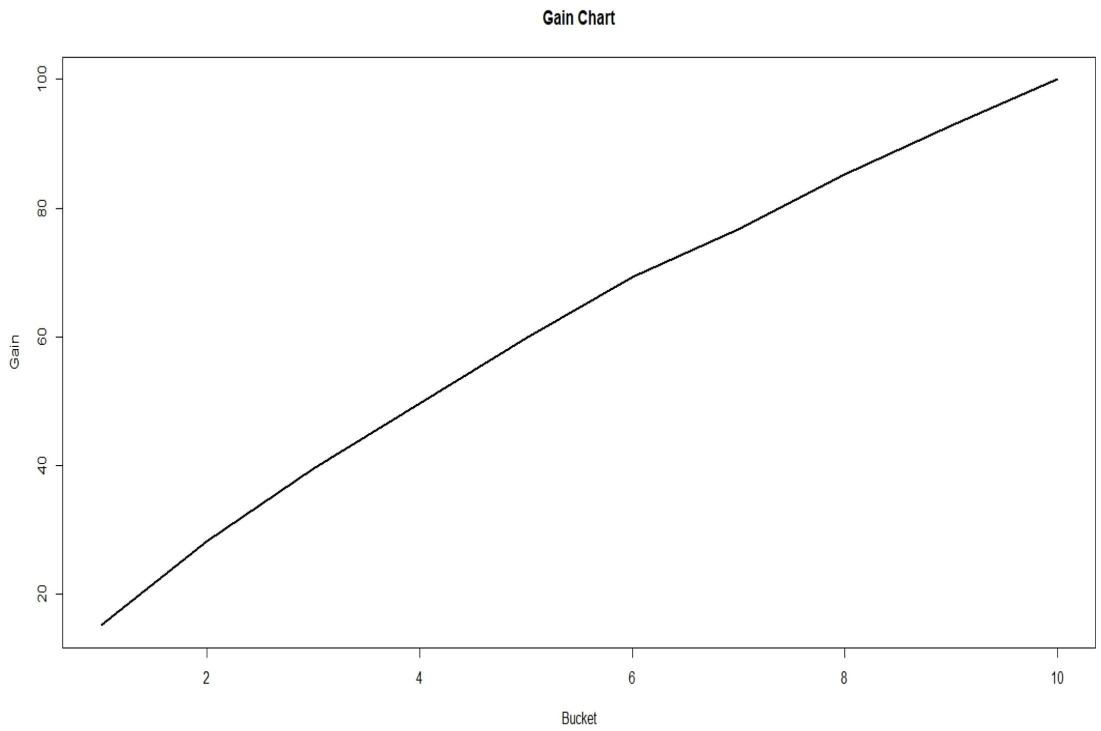


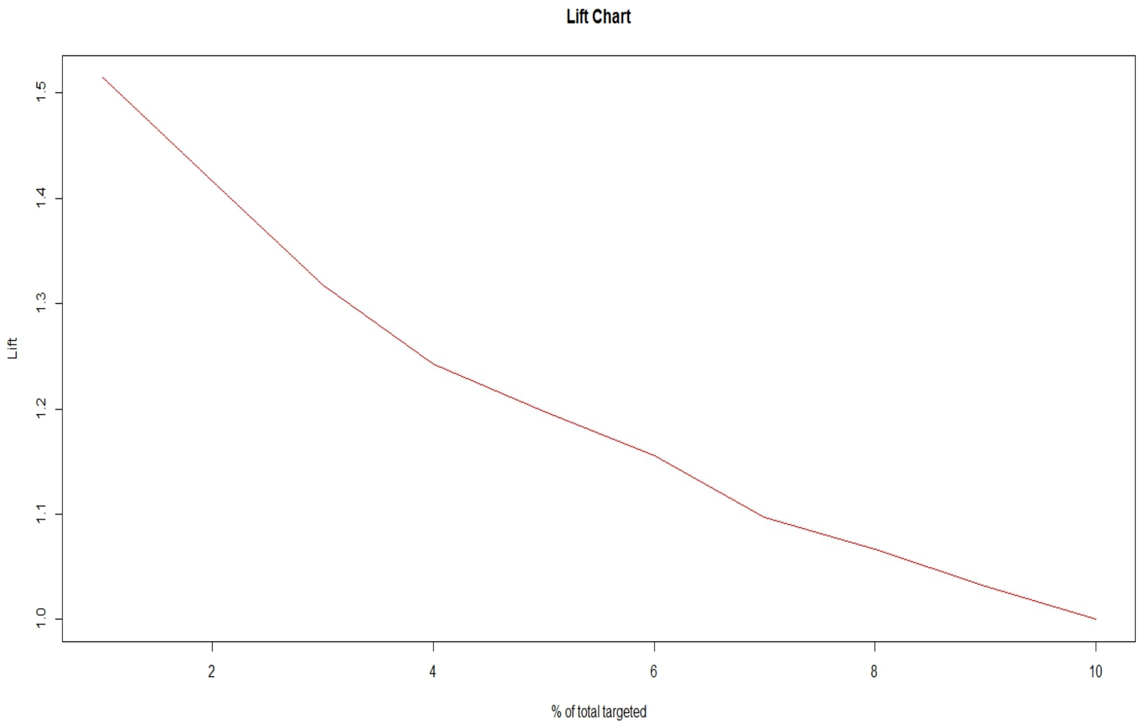


The following are the important predictor variables obtained using this model:

- Income
- No.of.months.in.current.residence
- No.of.months.in.current.company
- Education.xPhd
- Education.xProfessional

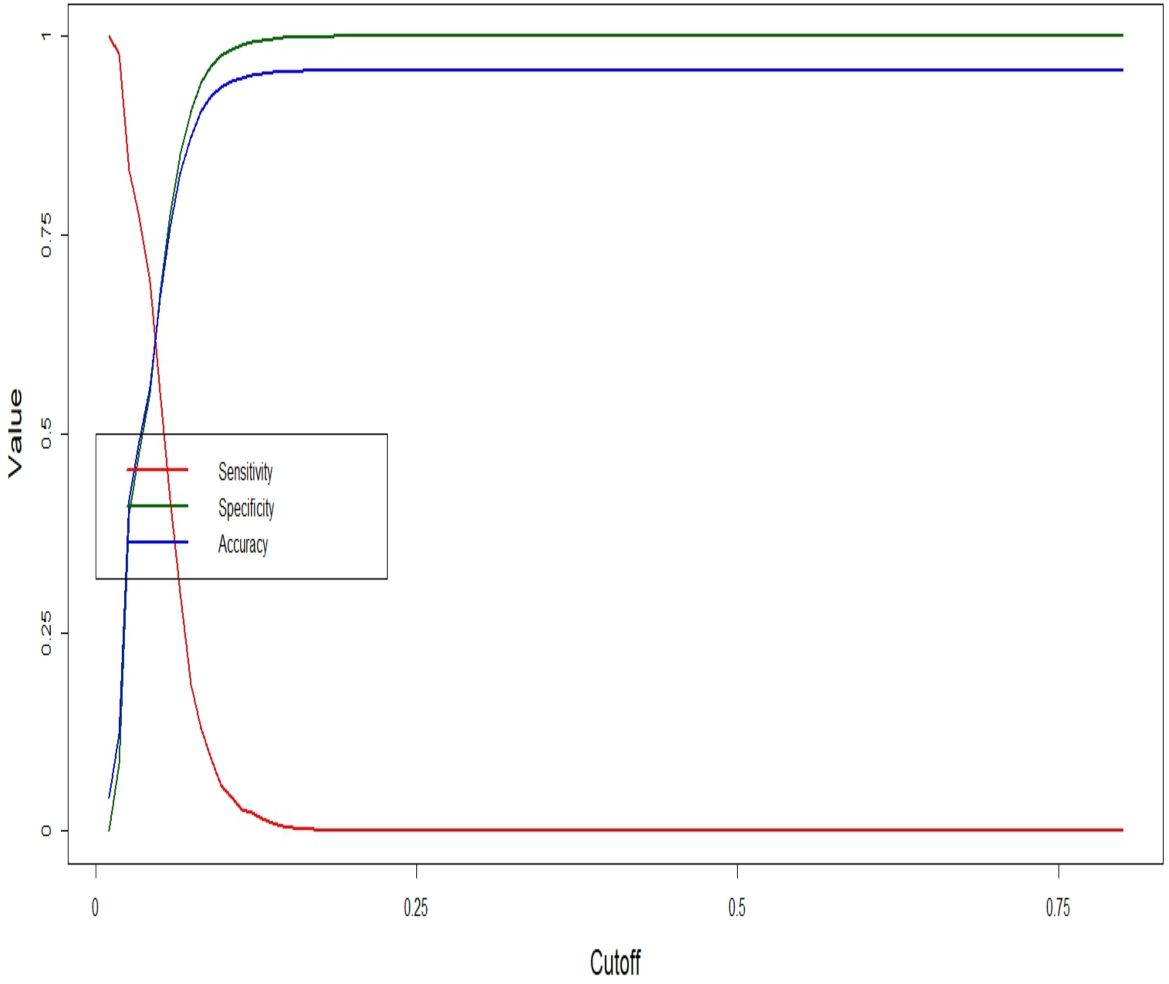
Logistic Regression Model using Only Demographic Data (using SMOT)

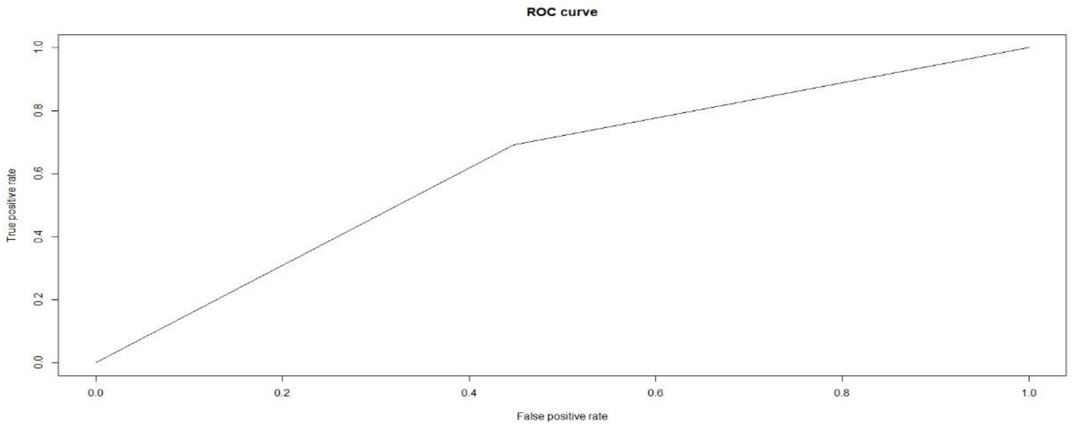




From the lift gain chart, it is clear that most of the values (70-75%) are accurately predicted by the 6th decile.

Logistic Regression Model (without using SMOT)

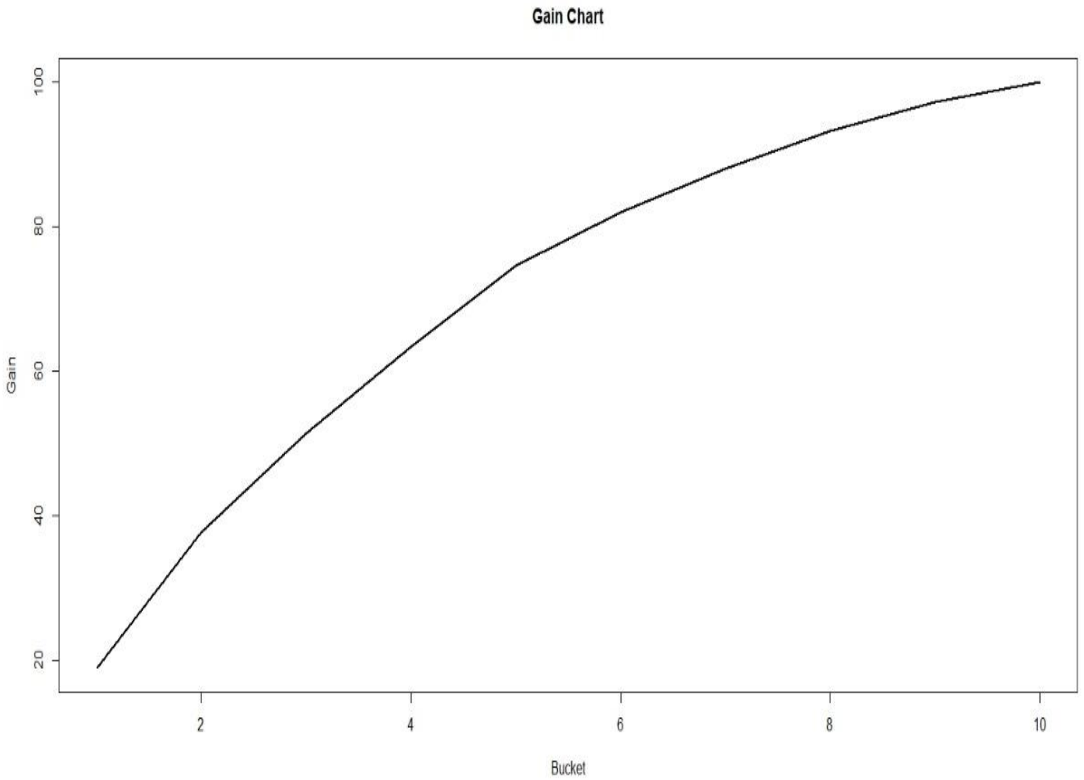


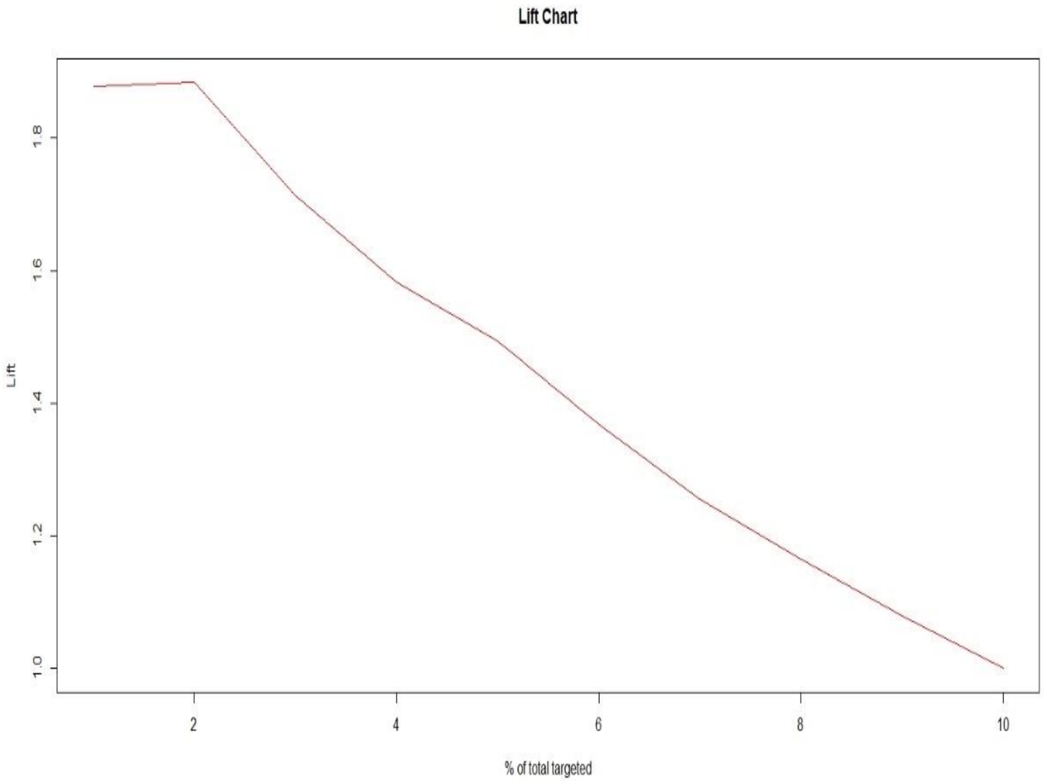


The following are the important predictor variables obtained using this model:

- No.of.months.in.current.residence
- Outstanding.Balance
- Avgas.CC.Utilization.in.last.12.months
- No.of.times.30.DPD.or.worse.in.last.6.months
- No.of.PL.trades.opened.in.last.6.months
- Presence.of.open.auto.loan

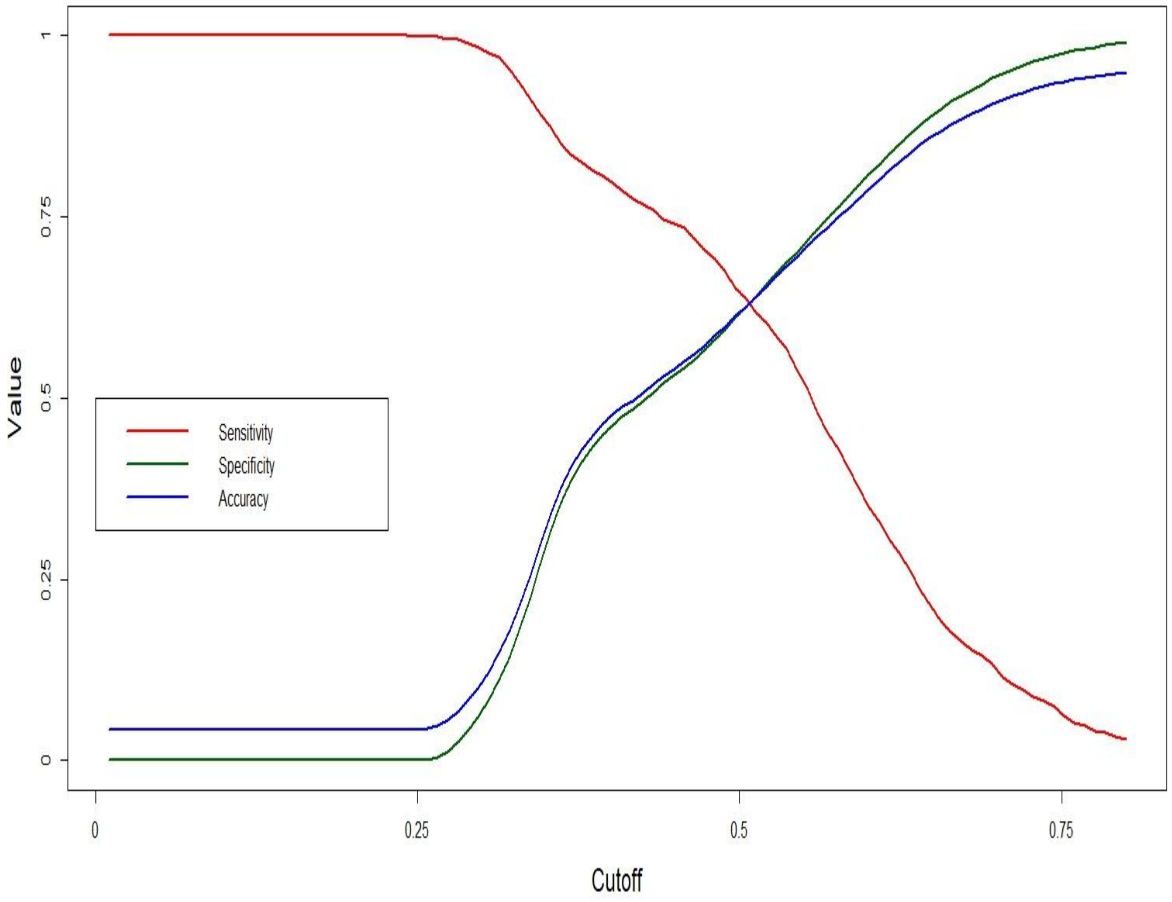
Logistic Regression Model (without using SMOT)

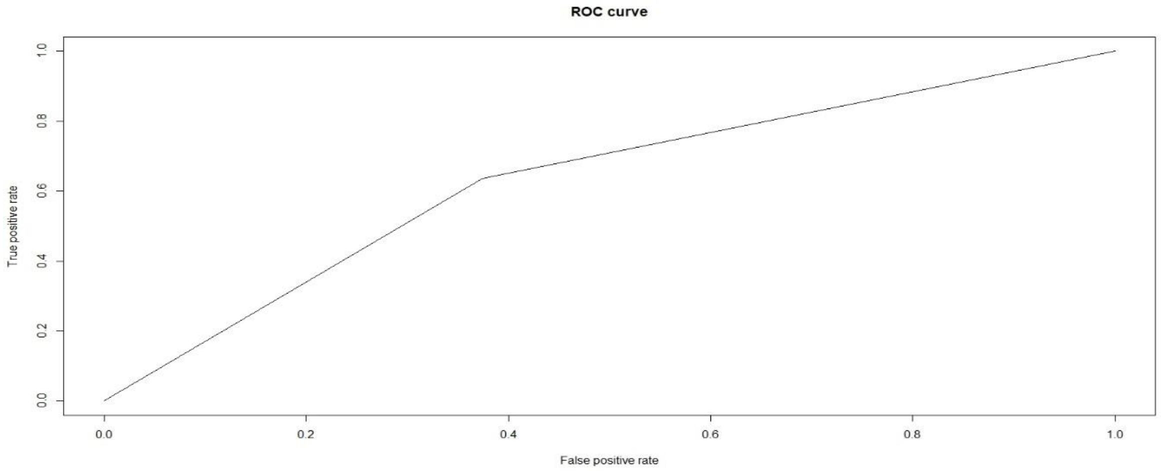




From the lift gain chart, it is clear that most of the values (80%) are accurately predicted by the 6th decile.

Logistic Regression Model (using SMOT)

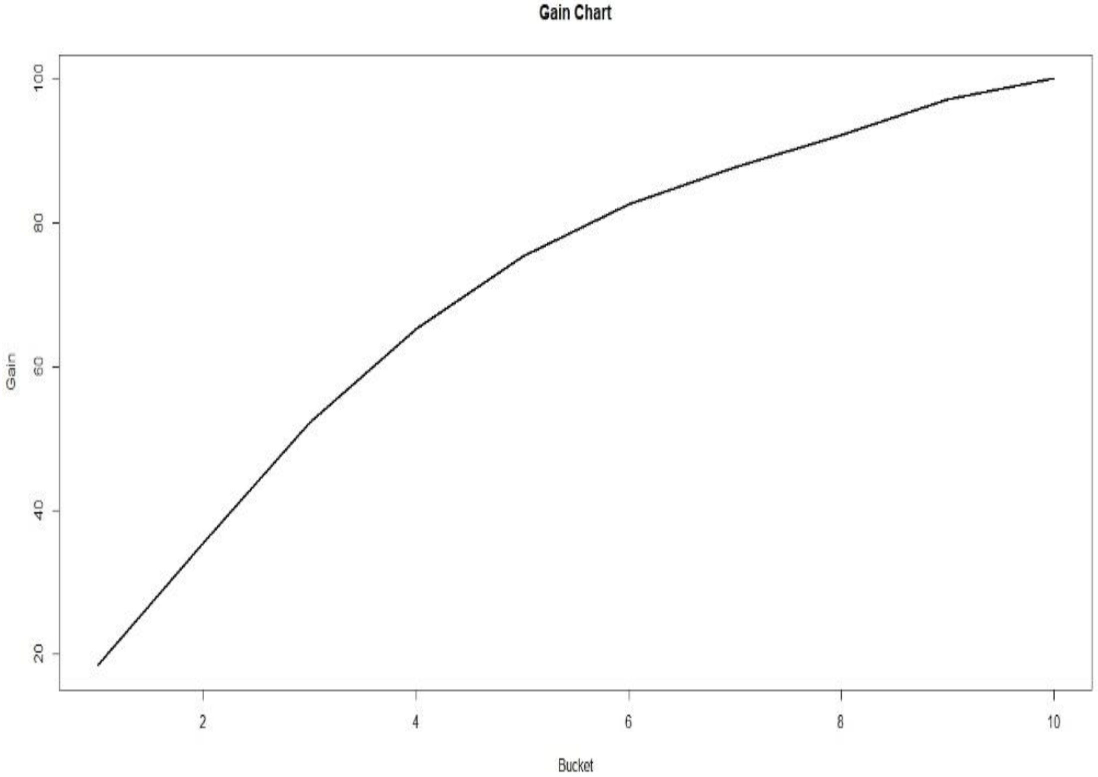


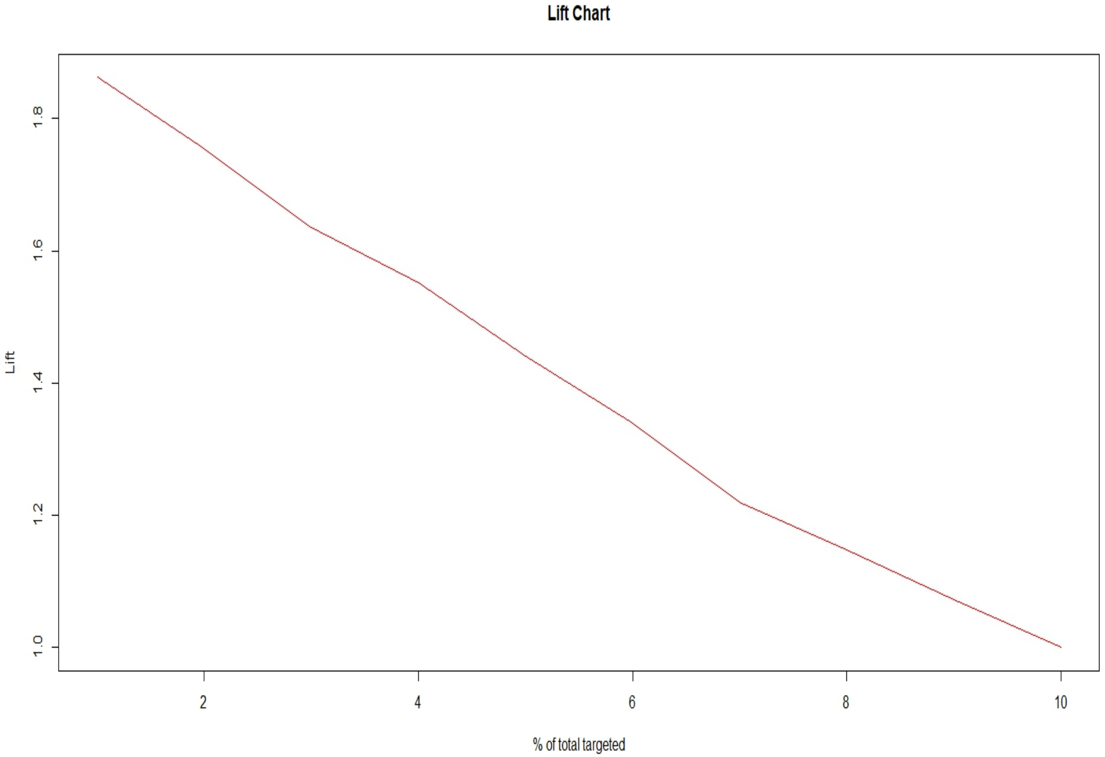


The following are the important predictor variables obtained using this model:

- Income
- No.of.months.in.current.residence
- No.of.months.in.current.company
- Avgas.CC.Utilization.in.last.12.months
- No.of.times.30.DPD.or.worse.in.last.6.months
- No.of.PL.trades.opened.in.last.6.months

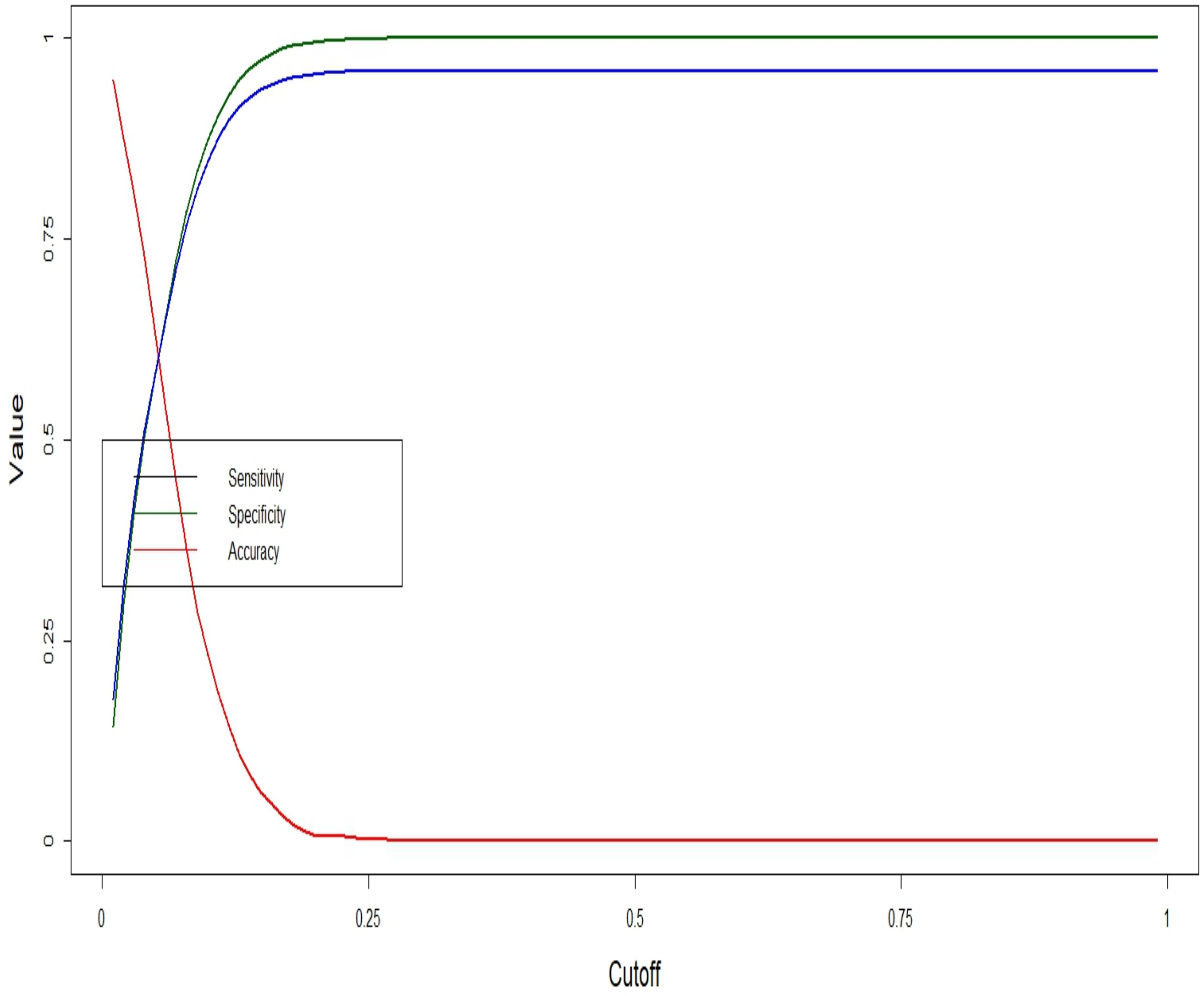
Logistic Regression Model (using SMOT)

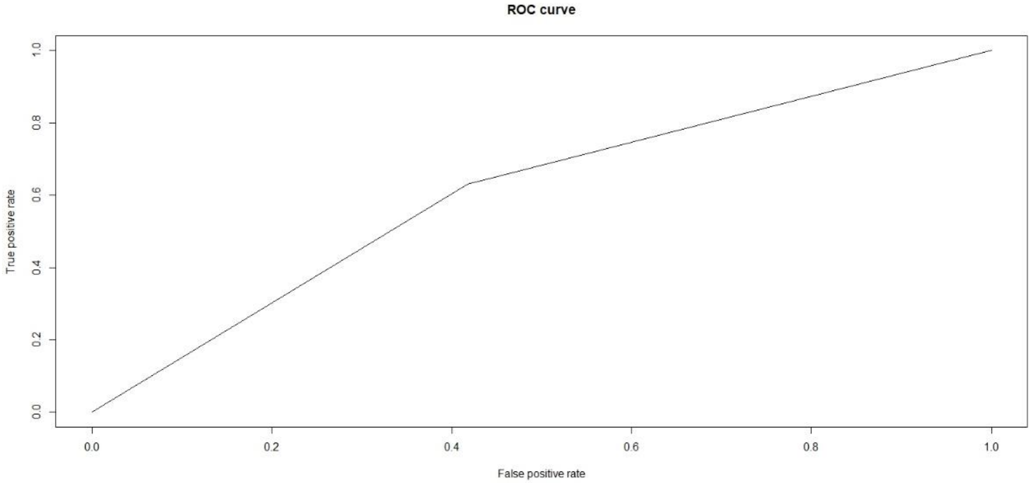




From the lift gain chart, it is clear that most of the values (80%) are accurately predicted by the 6th decile.

Random Forest Model (without using SMOT)

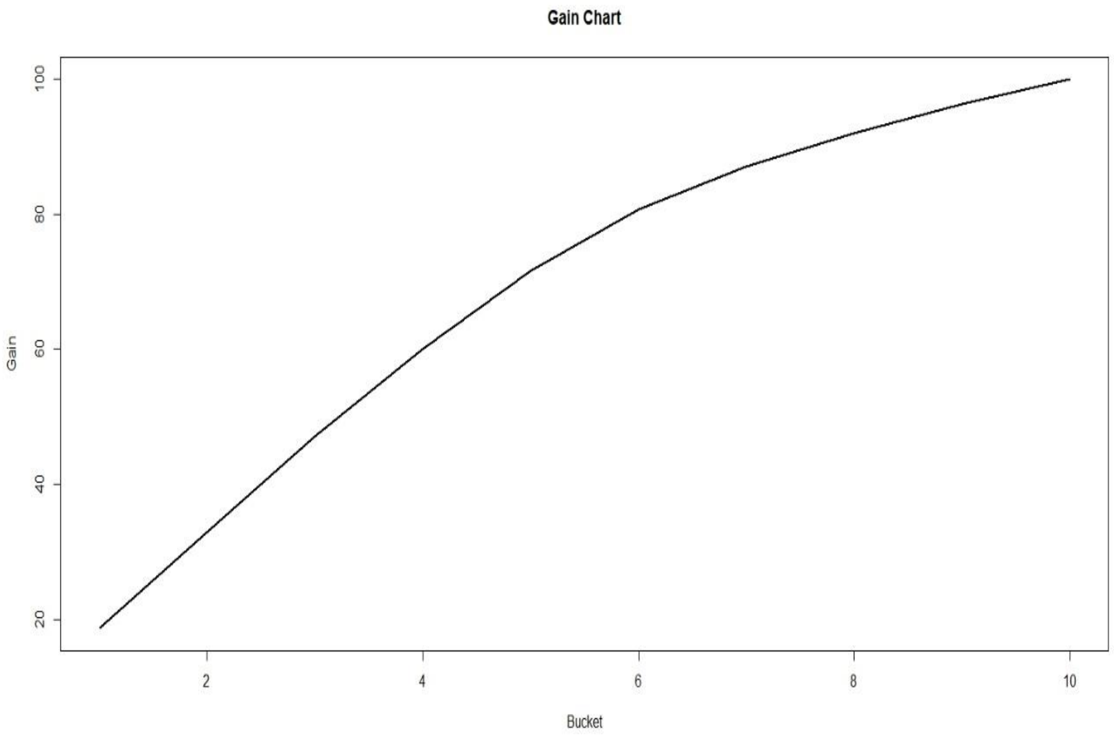


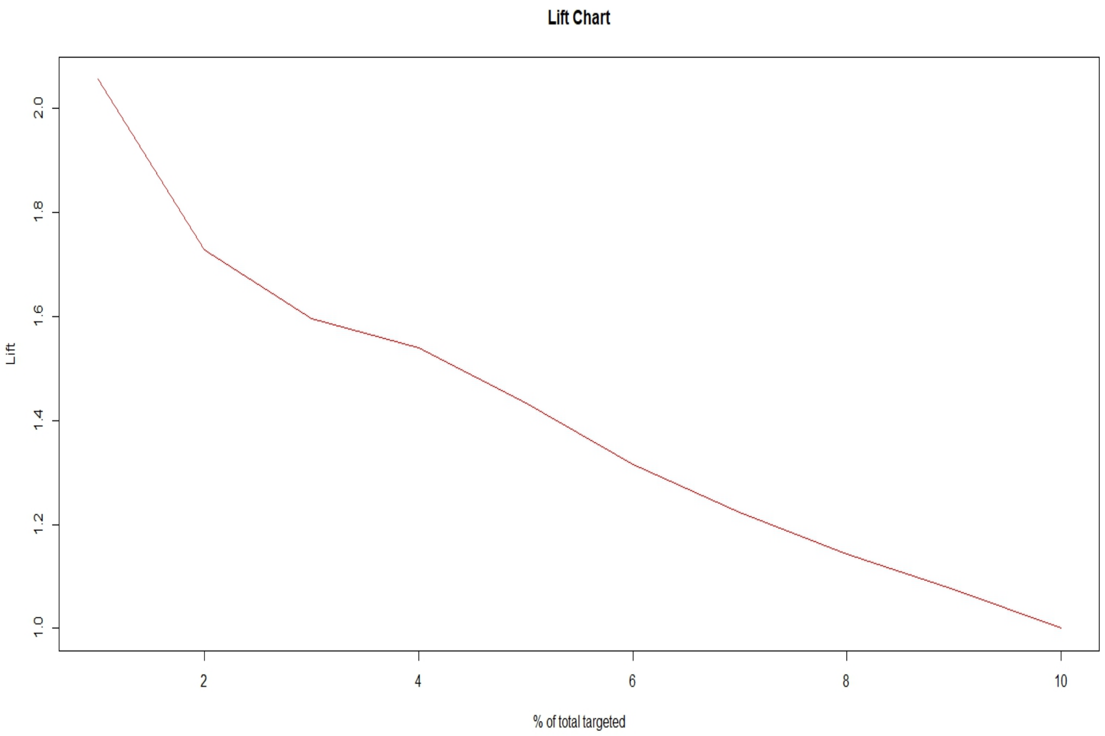


The following are the important predictor variables obtained using this model:

- No.of.months.in.current.company
- Income
- Avgas.CC.Utilization.in.last.12.months
- Age
- No.of.months.in.current.residence
- Total.No.of.Trades

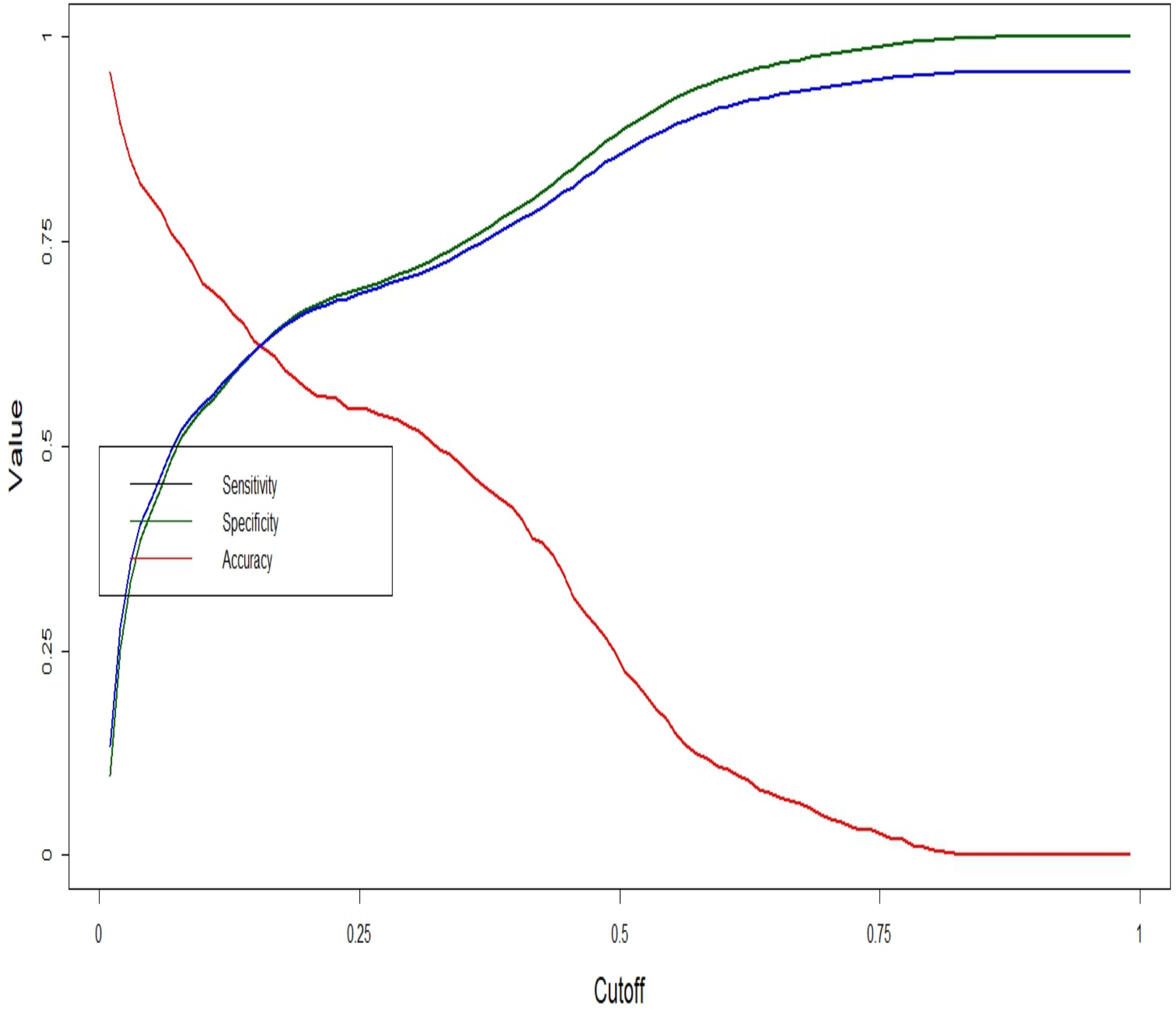
Random Forest Model (without using SMOT)

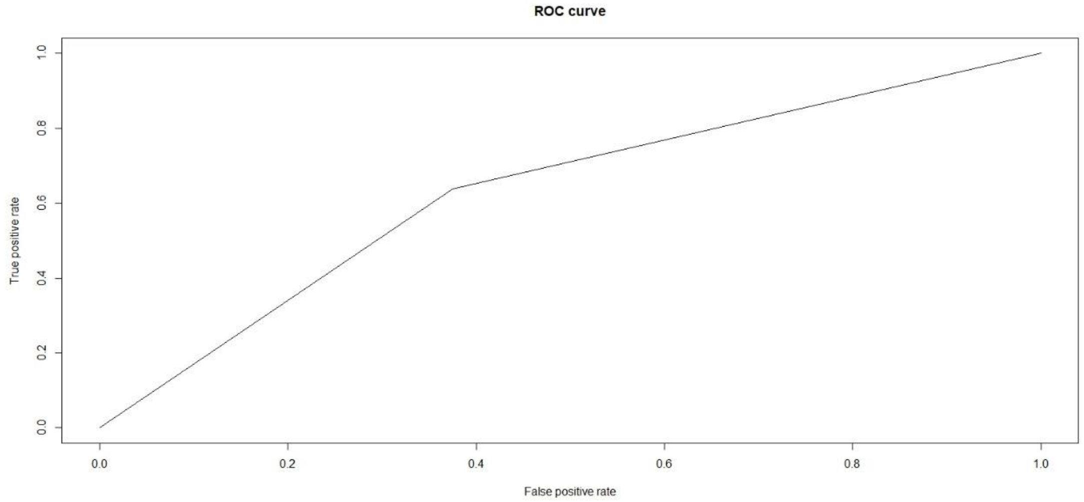




From the lift gain chart, it is clear that most of the values (80%) are accurately predicted by the 6th decile.

Random Forest Model (using SMOT)

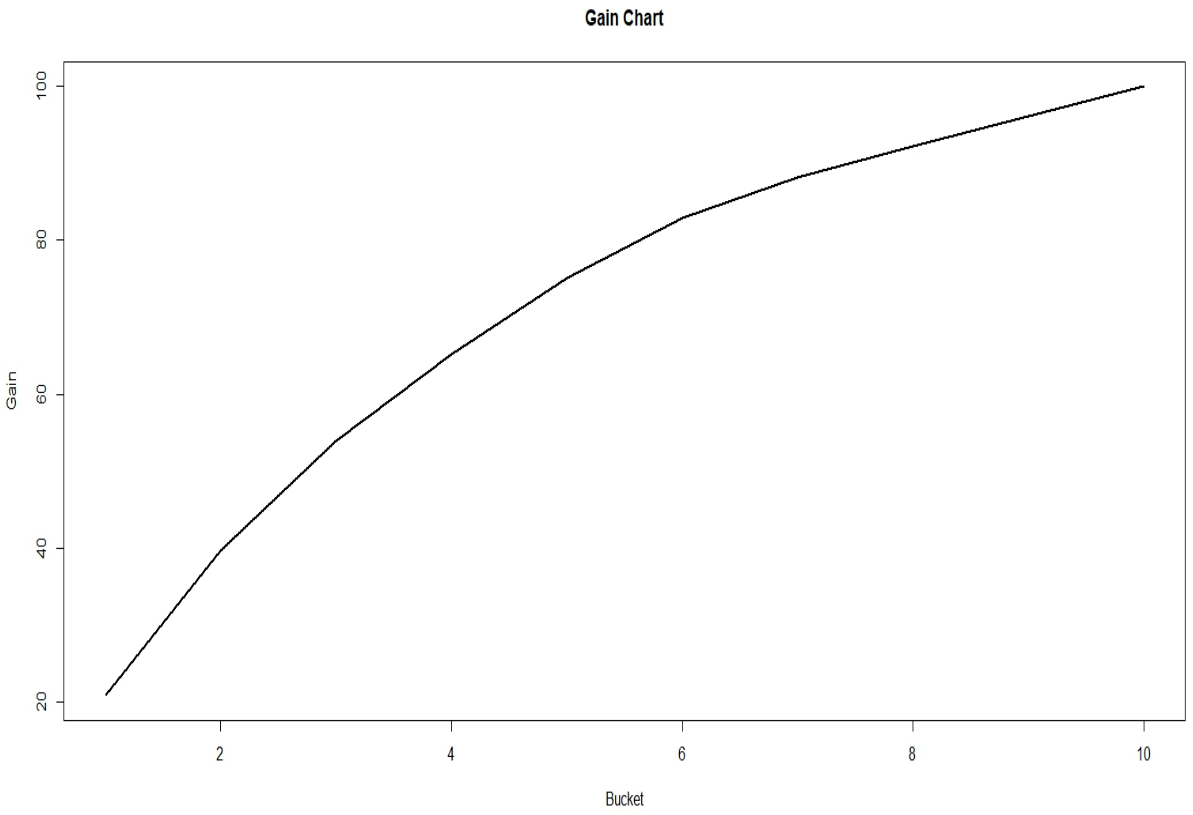


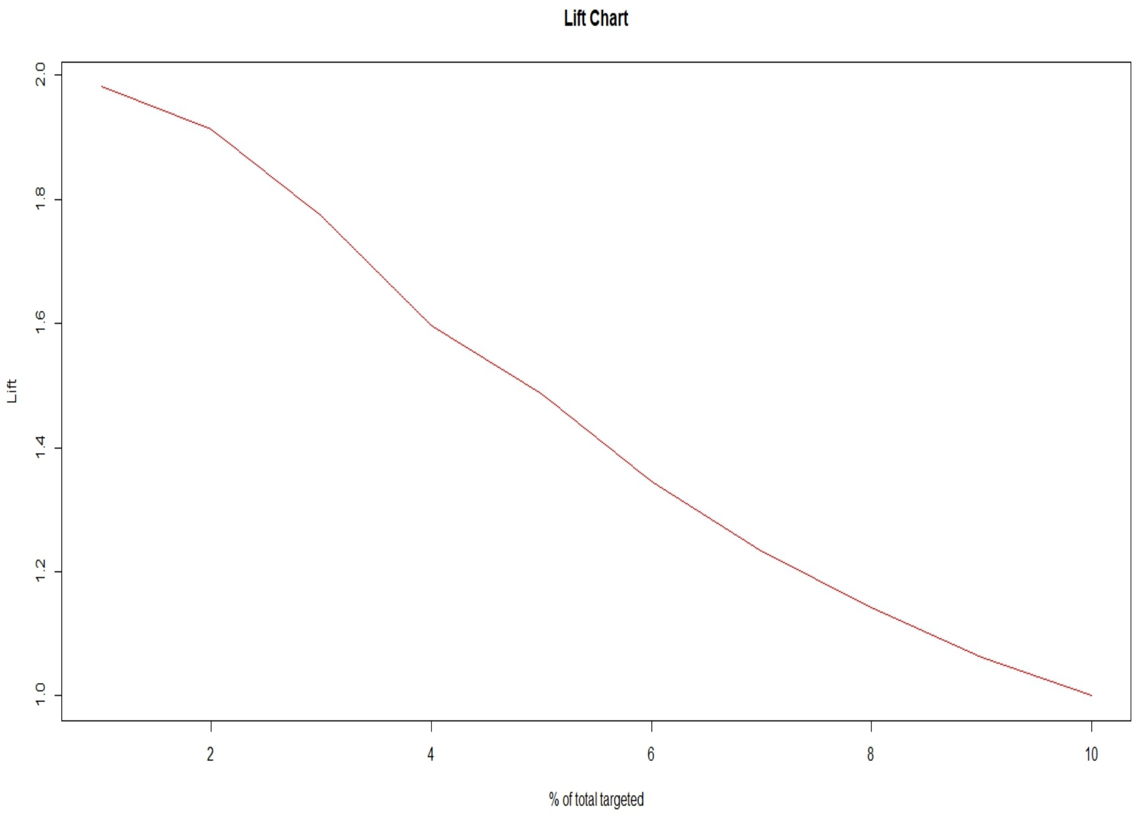


The following are the important predictor variables obtained using this model:

- No.of.times.30.DPD.or.worse.in.last.6.months
- No.of.times.60.DPD.or.worse.in.last.6.months
- No.of.times.30.DPD.or.worse.in.last.12.months
- No.of.times.90.DPD.or.worse.in.last.6.months
- No.of.times.90.DPD.or.worse.in.last.12.months
- No.of.times.60.DPD.or.worse.in.last.12.month
- Outstanding.Balance
- Avgas.CC.Utilization.in.last.12.months

Random Forest Model (using SMOT)





From the lift gain chart, it is clear that most of the values (80%) are accurately predicted by the 6th decile.

Model Evaluation

- Plot **the sensitivity, specificity and accuracy** at **various cut-off values**.
- Choose the best cut-off value where all the three parameters are very high.
- **Plot the confusion matrix for the best cut-off value**.
- Use any of the sensitivity, specificity and accuracy values as the appropriate evaluation metric.
- Use **the KS-Statistics and Lift-Gain chart** to find out at **which decile most of the values are accurately predicted**.
- The **area under ROC curve (auc)** is used as the **evaluation metric for model comparison and final model selection**. The higher the auc, the better the model.
- Hence train_smote_model_final (Logistic regression model using SMOT) is selected as the final model for scorecard building and financial analysis.

Logistic Regression Model Comparison

Model	Cut-off	Accuracy	Sensitivity	Specificity	KS-Statistic	Area under ROC curve (auc)
Logistic regression model (without SMOT) (using only demographic data)	0.042	52.11%	57.69%	51.87%	9.554%	0.5478
Logistic regression model (without SMOT) (using both demographic and credit data)	0.05	67.22%	55.36%	67.75%	23.11%	0.6155

From the results given in the above table, it is clear that **credit factors** such as Outstanding.Balance No.of.times.30.DPD.or.worse.in.last.6.months, No.of.PL.trades.opened.in.last.6.months and Avgas.CC.Utilization.in.last.12.months **significantly impact the efficiency of the final model** to be used for analysis.

Logistic Regression Model Comparison

Model	Cut-off	Accuracy	Sensitivity	Specificity	KS-Statistic	Area under ROC curve (auc)
Logistic regression model (with SMOT) (using only demographic data)	0.0505	56.37%	53.06%	56.51%	9.57%	0.5479
Logistic regression model (with SMOT) (using both demographic and credit data)	0.0505	63.13%	63.65%	63.1%	26.75%	0.6337

From the results given in the above table, it is clear that **credit factors** such as Outstanding.Balance No.of.times.30.DPD.or.worse.in.last.6.months, No.of.PL.trades.opened.in.last.6.months and Avgas.CC.Utilization.in.last.12.months are **better predictors** of **default rate** as compared to demographic factors.

Comparison of the Models Obtained

Model	Cut-off	Accuracy	Sensitivity	Specificity	KS-Statistic	Area under ROC curve (auc)
Logistic regression model (without using SMOT)	0.05	67.22%	55.36%	67.75%	23.11%	0.6155
Logistic regression model (using SMOT)	0.0505	63.13%	63.65%	63.1%	26.75%	0.6337
Random Forest model (without using SMOT)	0.05	59.2%	64.59%	58.97%	23.56%	0.6178
Random Forest model (using SMOT)	0.16	62.86%	61.86%	62.91%	24.76%	0.6238

Model Evaluation Results for Logistic Regression (using Different Values of Function Parameters and Test Dataset)

Model	Cut-off	Accuracy	Sensitivity	Specificity	Seed
Logistic regression model (without using SMOT)	0.05	67.45%	54.4%	68.03%	1
Logistic regression model (without using SMOT)	0.05	67.68%	54.97%	68.23%	500
Logistic regression model (without using SMOT)	0.05	67.22%	55.36%	67.75%	100
Logistic regression model (using SMOT)	0.0505	63.28%	60.33%	63.41%	1
Logistic regression model (using SMOT)	0.0505	63.35%	61.03%	63.45%	50
Logistic regression model (using SMOT)	0.0505	63.13%	63.65%	63.1%	100

Comparison of the Models Obtained and Conclusions Drawn

- While **performing validation** on different test data sets, the **model results** are found do be **virtually similar** for a **given cutoff**.
- For random forests, **cross-validation can be done immediately after model building** rather than explicitly since the given problem is a **classification problem**.
- From the tables, it is clear that the **auc (area under the ROC curve) value is the highest for the logistic regression model using SMOT** (0.6337).
- Hence, **train_smote_model_final** is the one which is **used for financial analysis, prediction and scorecard building using its cutoff value 0.505**.
- Thus, this model is also used for **predicting the probability of default for the rejected candidates**.
- Thus, using the model it can be successfully predicted that **most of the rejected candidates are defaulters**.
- From the models obtained, it is clear that **credit factors like avg credit card utilisation and number of trades** have a **significant impact on prediction and financial analysis**.

Prediction for Rejected Population Using the Model

- **train_smote_model_final** is the one which is used for **prediction of rejected population and scorecard building** using its cutoff value 0.505.
- This is because **its auc is highest** among the created models (0.6337).
- After prediction, the results are as follows:

0 (default)	1 (non-default)
166	1259

- This shows that **most of the rejected population** is **correctly predicted as defaulters** using the model.

NOTE: The model evaluation and validation done is based on appropriate cutoff score using the graph of accuracy, sensitivity and specificity and also by using lift-gain charts.

Building the Application Scorecard And Obtaining the Cutoff Score

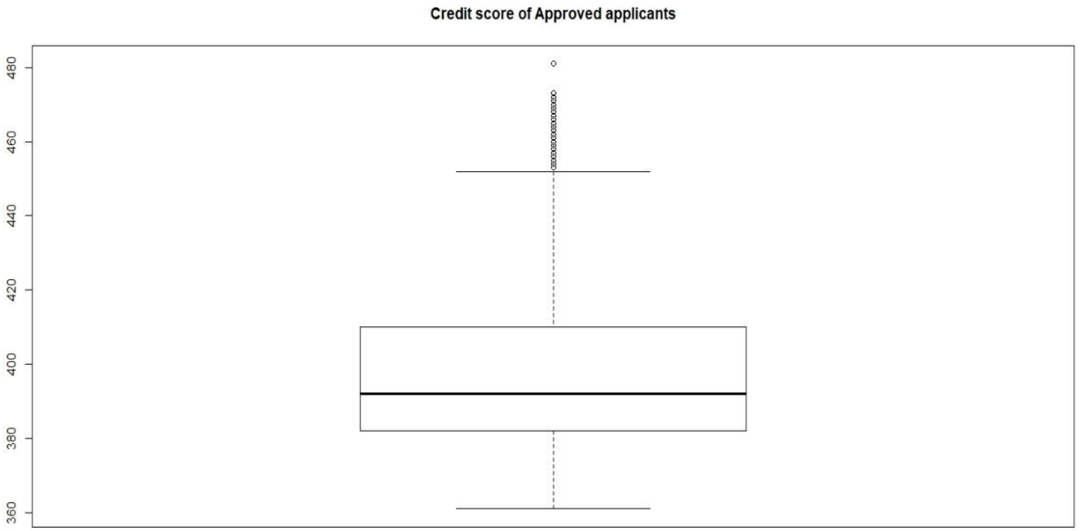
- An application scorecard is built with the good to bad odds of 10 to 1 at a score of

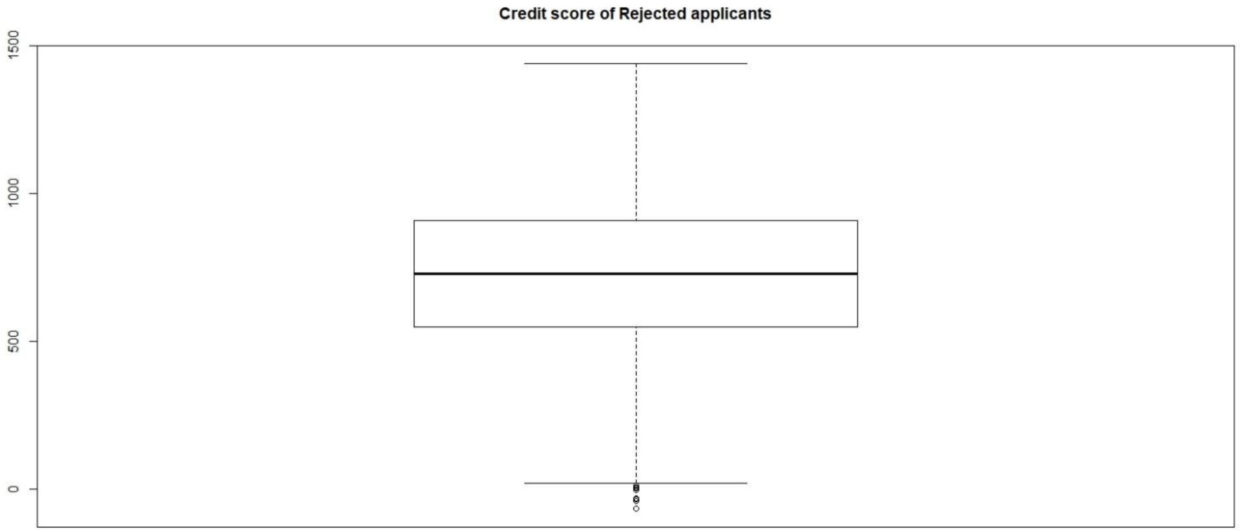
  400 doubling every 20 points.
- This is done as follows:
  - For the rejected population, **calculate the application scores** and **assess the results**. Compare the scores of the rejected population with the approved candidates and comment on the observations.
  - On the basis of the scorecard, **identify the cut-off score** below which you would not grant credit cards to applicants.
- The cutoff score for both approved and rejected population is found using the formula below:

  cutoff<- 400 + (Factor * (log((1-model_cutoff)/model_cutoff) - log(10)))

  Where, Factor = PDO/log(2) =20/log(2)
- Thus, the cutoff score obtained using the model scorecard is 333.
- Thus, using the above predictions and formulae, the application scorecard is built.

Comparison of Approved and Rejected Population Using the Credit Score





From the above boxplots, it is clear **that mean and median score of rejected customers is much higher** than those of approved customers.
Based on the cutoff score, 106 people from the rejected population would not be issued the credit card since all of them have credit scores below the cutoff score (333).

Financial Analysis Using The Model

- The **potential financial benefit** of the project is explained in P&L terms in the slides that follow.
- The main objective is to report the following:
  - The implications of using the model for auto approval or rejection, i.e. how many applicants on an average would the model automatically approve or reject
  - The **potential revenue and credit loss** avoided with the help of the model
  - Assumptions based on which the model has been built

Financial Analysis Using The Model

- The **main assumption** used here is that we are considering **the average loss of 10000 when each non-defaulter's application is rejected and an average loss of 100000 when each accepted applicant defaults**.
- Loss_without_model = 374500000
- Loss_with_model = 215560000
- Thus, the **financial loss incurred is reduced by 42.44%** due to the final model.
- The following is the confusion matrix obtained after prediction.

Reference Matrix		
Prediction	No	Yes
No	41413	1191
Yes	25507	1756

- Thus, the number of non-defaulters and defaulters are 66920 (95.8%) and 2947 (4.2%), respectively.

Financial Analysis using the Final Model

From the model predictions and confusion matrix obtained in the previous slide, the following conclusions are drawn:

- **25507 customers** (**38.12%** of the total non-defaulting customers**) cause revenue loss to CredX**. They were incorrectly identified as non-defaulting customers by the model.
- **1191 approved customers** (**1.705%** of the total population**) did not default on Credit card payments**. They were incorrectly identified as defaulting customers by the model.
- Thus, **the credit loss incurred by CredX is reduced by 40.28%** due to the final model.
- Hence, using correct model and cutoff predictions we **can reduce the revenue loss, credit loss and financial loss incurred by CredX**.

Assumptions Made in Financial Analysis

The following assumptions have been made while assessing the model in financial terms:

- The **cost of acquisition** for every customer has been **considered the same.**
- The **expected business/revenue/profit for every customer** has been **considered the same.**
- The figures (acquisition cost/revenue per person) are taken **for representational purpose only**. Actual values may vary.

Important Assumptions Used in the Capstone Project

- The **categorical and independent variables** are converted into **factors for EDA** and **dummy variables** are created using the same for ease of analysis.
- The **NA/missing value imputation is done for numeric variables with WOE/IV values** for ease of analysis.
- The **missing value imputation is done using mode for categorical variables** for ease of analysis.
- But **NA values in performance tag** are considered as the **rejected population**. The rows containing such values are **not ignored** as they are **used for application scorecard and financial analysis** at the later stage.
- Some of the **outliers** can be taken care of using **WOE/IV values** and **binning of variables.**
- **Warning messages** arising from function calls **can be ignored** as they **do not affect** the running of the code.

Important Assumptions Used in the Capstone Project

- For some EDA, only **performance tag 1 is considered** as we want to **analyse the influence of various factors on number of defaulters**.
- **WOE/IV binning and values** are not **explicitly shown in model building** as they are **taken care of in earlier stages of the EDA**.
- **SVM is not used** since the **amount of data** involved and to be modelled **is very large**.
- **Random Forests** are used for modelling rather than Decision Trees since a **diverse number of trees** based on **different features and parameters** are constructed, thus **reducing the problem of overfitting** .
- For **random forests**, **cross-validation can be done immediately** after model building rather than explicitly since the **given problem is a classification problem**.
- In **random forests**, **outliers are not ignored** as **decision rules at the leaves of the trees constructed are not sensitive to outliers** and **hence model accuracy is not affected**.
- This is because **lot of trees are built at the same time** using random forest.
- **Outliers** are **not used in logistic regression** as **it affects model accuracy**.

Important Factors Taken Care of During Model Building

- In **logistic regression models**, variable removal is done on the **basis of VIF and then based on p-values**.
- Thus, in final logistic regression models, only variables having **VIF < 2 and p-values <0.05** are considered.
- For demographic data, **only logistic regression model** is used since it acts as **the baseline model** for comparing with models done based on both credit and demographic data.
- The **final model** selected is involved to **predict the default probabilities for rejected population** and hence, it is used for **financial analysis and scorecard building**.
- The **model evaluation and validation** done is based on **appropriate cutoff score** using the graph of **accuracy, sensitivity and specificity** and also by using **lift-gain charts**.
- In case of **logistic regression model**, **cross-validation** is also carried out.

Inferences and Conclusions

- Thus, the EDA and models used identify the important variables that predict the probability of default.
- The **auc (area under the ROC curve)** value is used as the **evaluation metric for model comparison and selection**.
- Thus, the **logistic regression model obtained after using SMOT** (train_smote_model_final) is used for financial analysis, prediction and scorecard building.
- **Credit factors** like avg credit card utilization and number of **trades have greater impact on the model efficiency and analysis (EDA)** and **model efficiency** as compared to demographic factors.
- The **cutoff score obtained was 333**. The population having a score **below that cutoff** would not be **issued the credit card**.
- Using the model, **the financial, revenue and credit loss incurred by CredX** was **reduced substantially**.