BFS Capstone Project
CredX Credit Card Customers – Acquisition Analytics

BFS-12 Group Members

Sesha Sailendra Kona - DDA1720038

Krishnamani Ananthanarayanan – DDA1720224

Maheshwara Reddy Golla – DDA1720125

Sumit Arora – DDA1720037

Business objectives & Strategy

Objective

CredX intends to mitigate their credit risk during acquisition by ***Finding The Right Customers***. A Right customer would be the one who is not too risky to pay back their loan. In addition, right customer is the one who might miss one or two payment past due date but eventually closes the loan with interests and penalty.

Strategy

To Solve business problem we are using ***CRISP-DM*** **framework**

- Using past data of the bank's applicants identify the most important factors affecting credit risk
- Create strategies to mitigate the acquisition risk for new applications, by identifying right customers using predictive modelling to differentiate Good Vs Bad customer
- Assess the financial benefit of project.

Analytical Problem Solving Methodology

Data Understanding

Step 1 : Understand how the data are organized

Step 2 : Identify Rejected Applications

Data Preparation and Feature Engineering

Step 3 : Missing Value Imputation and outlier treatment

Step 4 : Feature Engineering

Step 5 : Feature selection based on EDA and WoE/IV

Model Building

Step 6 : Data Sampling to Balance Unbalanced Classes

Step 7 : Model Selection

Model Evaluation

Step 8 : Cross validation with Relevant Metrics

Step 9 : Metrics for Imbalanced and Credit score

Model Deployment

Step 10 : Application Scorecard preparation

Assessment of Financial Benefit

Step 11: Compute various measure to understand the financial benefit of the model.

Data Understanding, Preparation and Feature Engineering

The data collected is from two different sources **Demographic** and **Credit Bureau** and is high imbalanced with approximately ***95%*** applicants are ***non-defaults*** and only ***4%*** with ***defaulters***.

Data Assumptions

- **Rejected Applications** - The given data is for approved loans and hence the *NA* values of *Performance Tag* assuming those are rejected records.

Data Quality Checks

- Identification of categorical and continuous variables.
- Check for invalid, missing values sanity check, duplicate records & outliers

Data Cleaning

- **Remove Duplicates** records in both data sets i.e. rows with *Application* ID values *765011468, 653287861 and 671989187*
- ***Performance Tag* values as *NA*** - Separate these rejected applications data records with and create another data frame as *rejected_applications*. This is used later to evaluate the final model.
- ***Age*** – Removed records with incorrect values -3 and 0. Also, removed with ***Age <18 years***, as credit card not approved

Data Integration

- Merged both the demographic and Credit data

Outlier Treatment

- Studied Outliers treatment feasibility and decide not to remove them.

Missing Value Treatment

- The records with very low percentage of incorrect and/or missing values have very low significance impact and hence they are replaced basic imputation techniques as median / mode etc
- Variables with significant number of values missing, we leveraged ***Weight of Evidence (WoE)*** Analysis based ***Coarse Classing***

Data Preparation and Feature Engineering

Feature Engineering techniques applied

One-Hot Encoding
- Gender
- Marital Status
- Profession
- Typo of residence
- Presence of open home loan

Label Encoding
- Education

Binning / Buckets
- Age
- Income
- No of months in current residence
- No of months in current company
- Avgas CC Utilization in last 12 months
- Outstanding Balance

Weight of Evidence (Information Value) Analysis
- Use variables only with IV values range as >= 0.02 and <=0.5

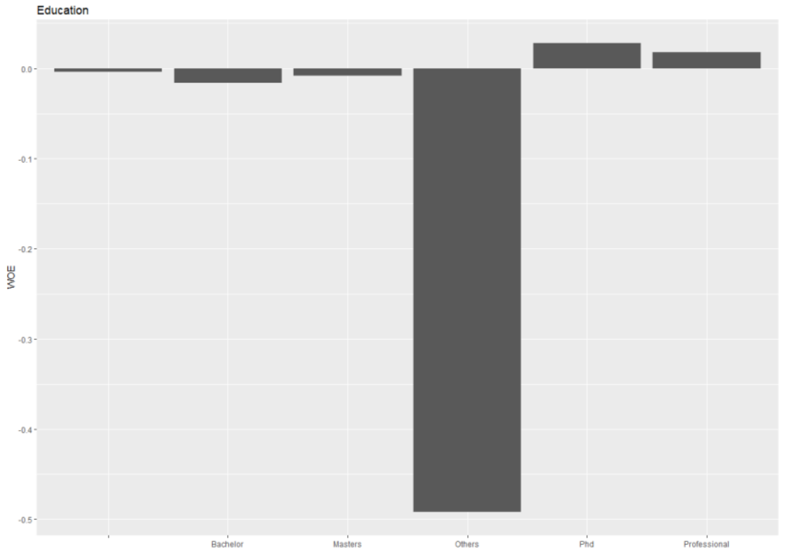
Feature Selection Based on EDA and WoE/IV Analysis
- Use WoE/IV Analysis together with EDA for final feature selection

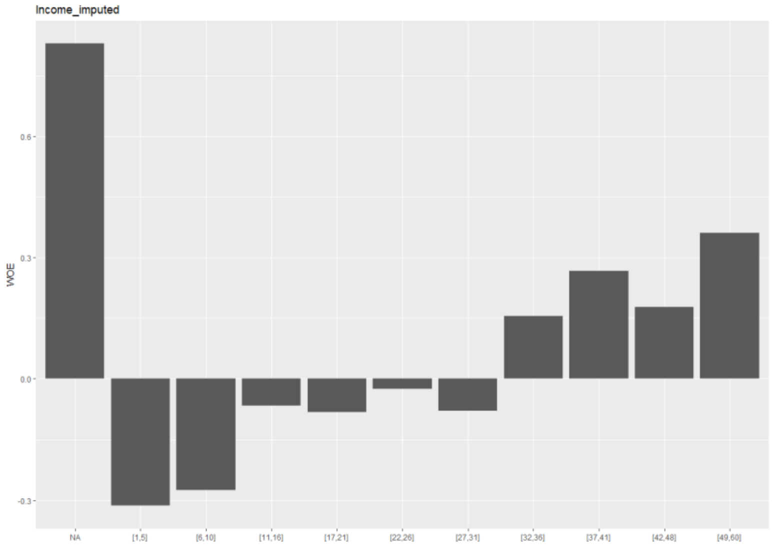
Feature Selection based on IV >=0.02

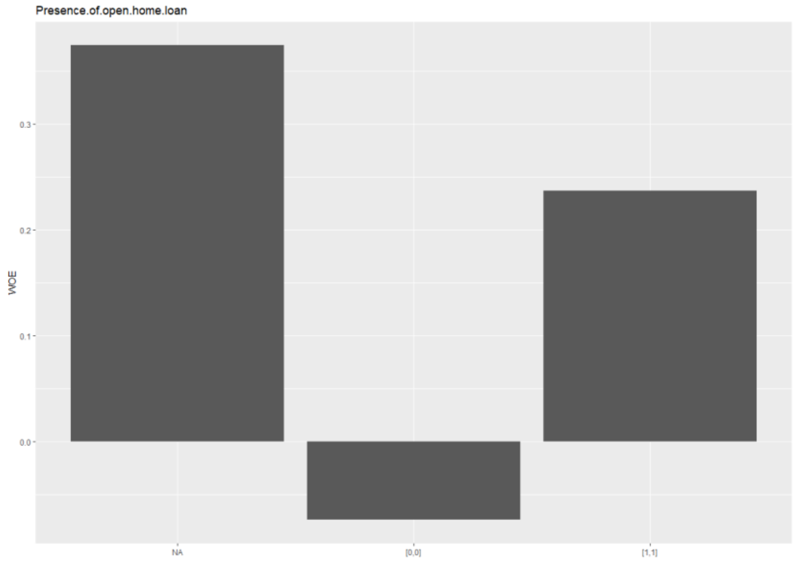
	Variable	IV
1	Avgas.CC.Utilization.in.last.12.months	0.31015860
2	No.of.trades.opened.in.last.12.months	0.29827712
3	No.of.PL.trades.opened.in.last.12.months	0.29604052
4	No.of.Inquiries.in.last.12.months..excluding.home....	0.29560438
5	Outstanding.Balance	0.24604217
6	No.of.times.30.DPD.or.worse.in.last.6.months	0.24167952
7	Total.No.of.Trades	0.23713984
8	No.of.PL.trades.opened.in.last.6.months	0.21973098
9	No.of.times.90.DPD.or.worse.in.last.12.months	0.21393775
10	No.of.times.60.DPD.or.worse.in.last.6.months	0.20592613
11	No.of.Inquiries.in.last.6.months..excluding.home......	0.20523987
12	No.of.times.30.DPD.or.worse.in.last.12.months	0.19848248
13	No.of.trades.opened.in.last.6.months	0.18597050
14	No.of.times.60.DPD.or.worse.in.last.12.months	0.18563797
15	No.of.times.90.DPD.or.worse.in.last.6.months	0.16016368
16	No.of.months.in.current.residence	0.07913990
17	current_residence_bin	0.06080075
18	Income_imputed	0.04393392
19	Income	0.04255551
20	Income_bin	0.04007508
21	No.of.months.in.current.company	0.02175181

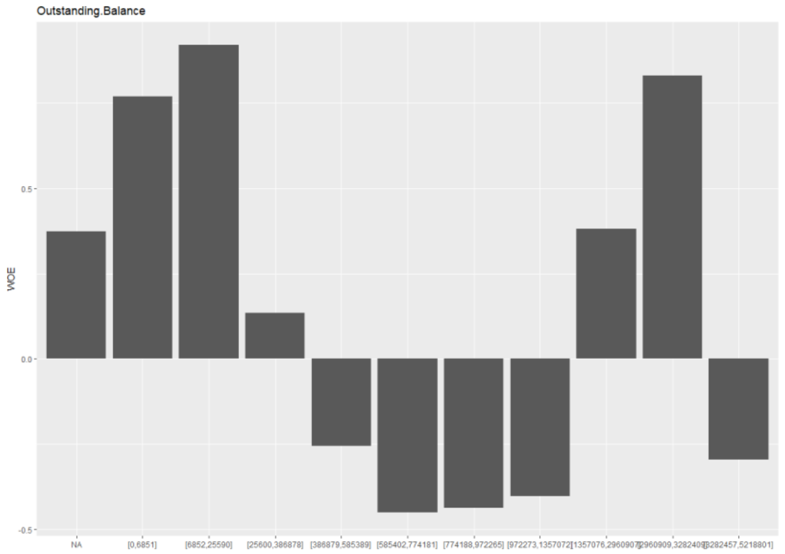
Exploratory Data Analysis – Weight Of Evidence (WoE)

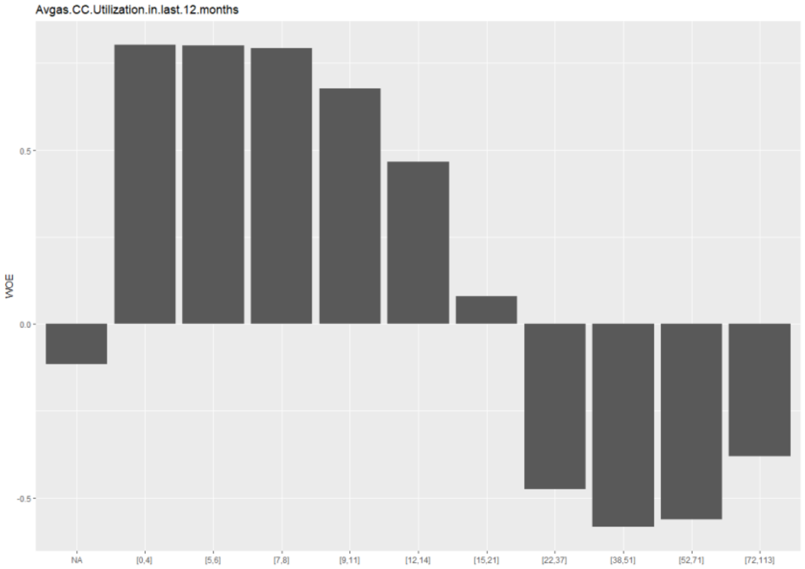
WoE Plots









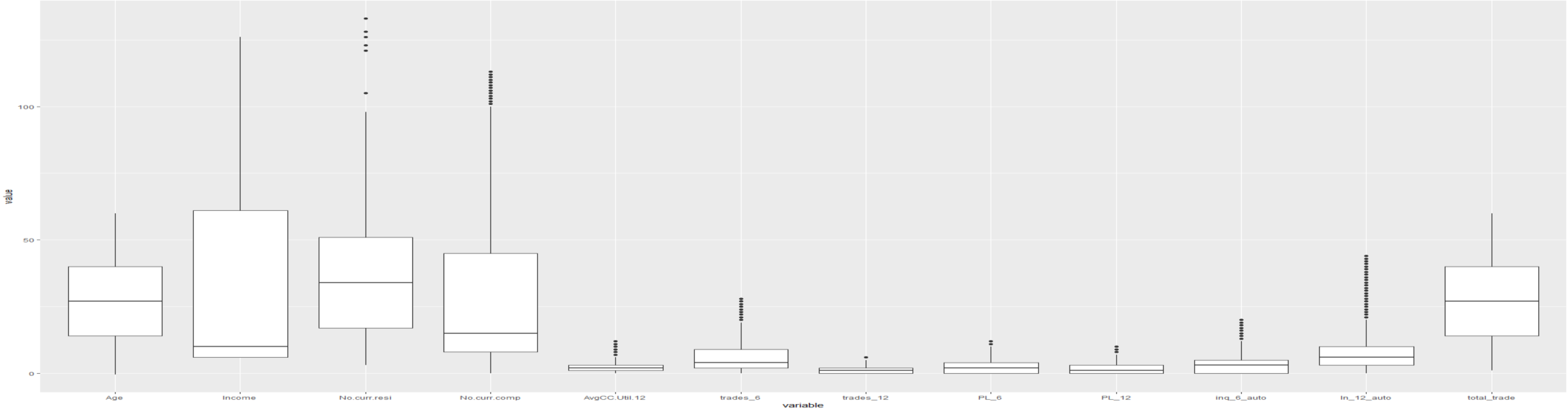


WoE / IV Method used for,

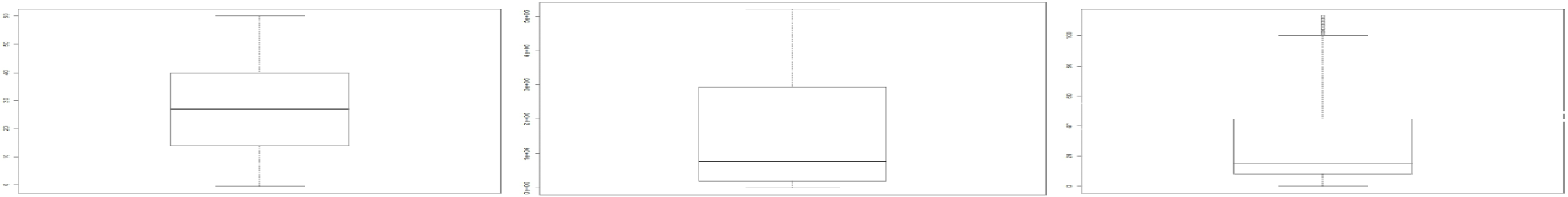
- Feature Importance
- Data Imputation with ***Coarse Classing*** for both *Master Data* and also *Rejected Applications* data

Exploratory Data Analysis - Univariate

Univariate Analysis for Continuous Variables



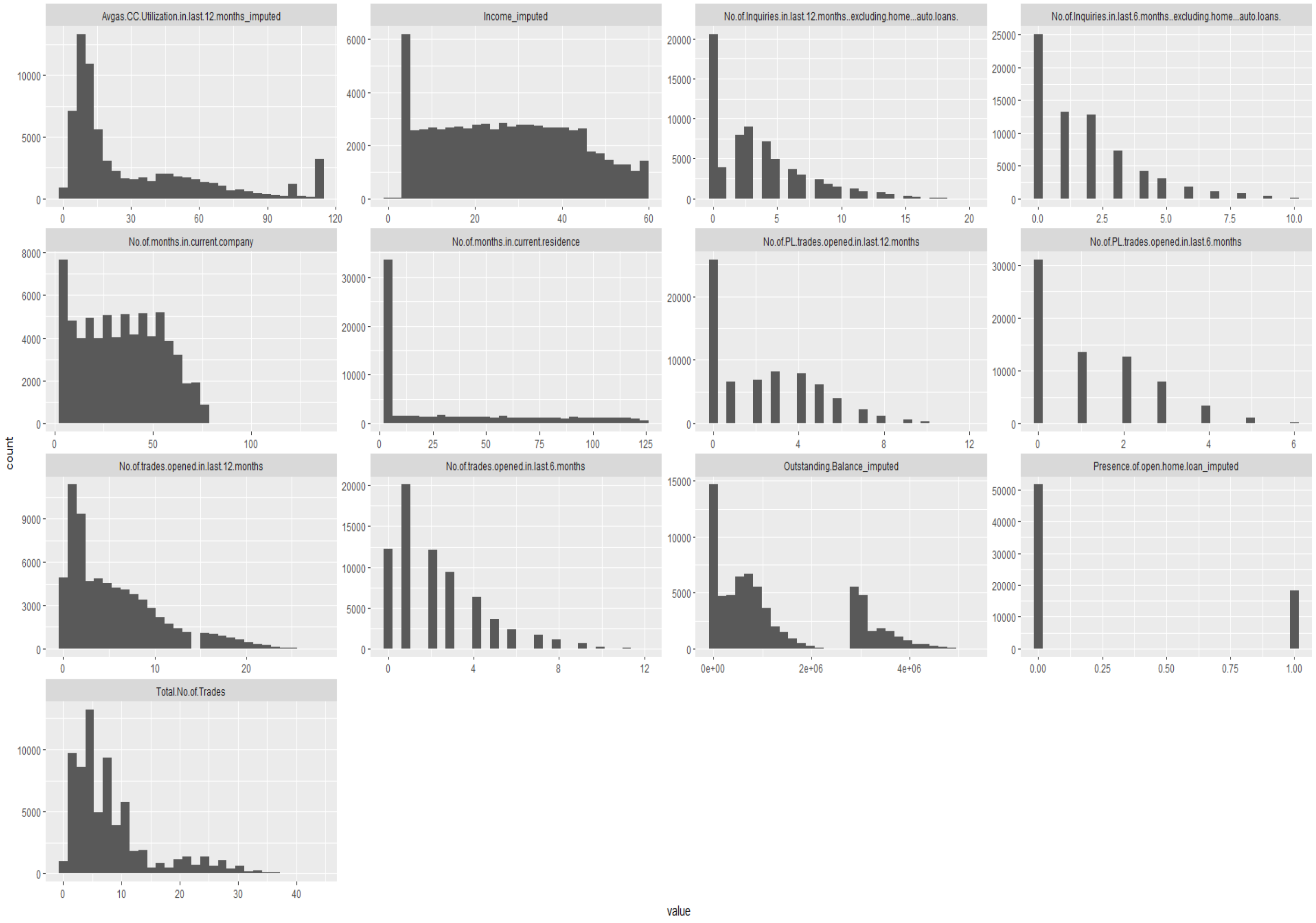
Income, Outstanding Balance and Average Credit Card Utilization



Outliers exist in Outstanding.Balance, Income, Avgas.CC.Utilization.in.last.12.months and No.of.trades.opened.in.last.12.months etc

Exploratory Data Analysis - Univariate

Univariate Analysis for continuous variables

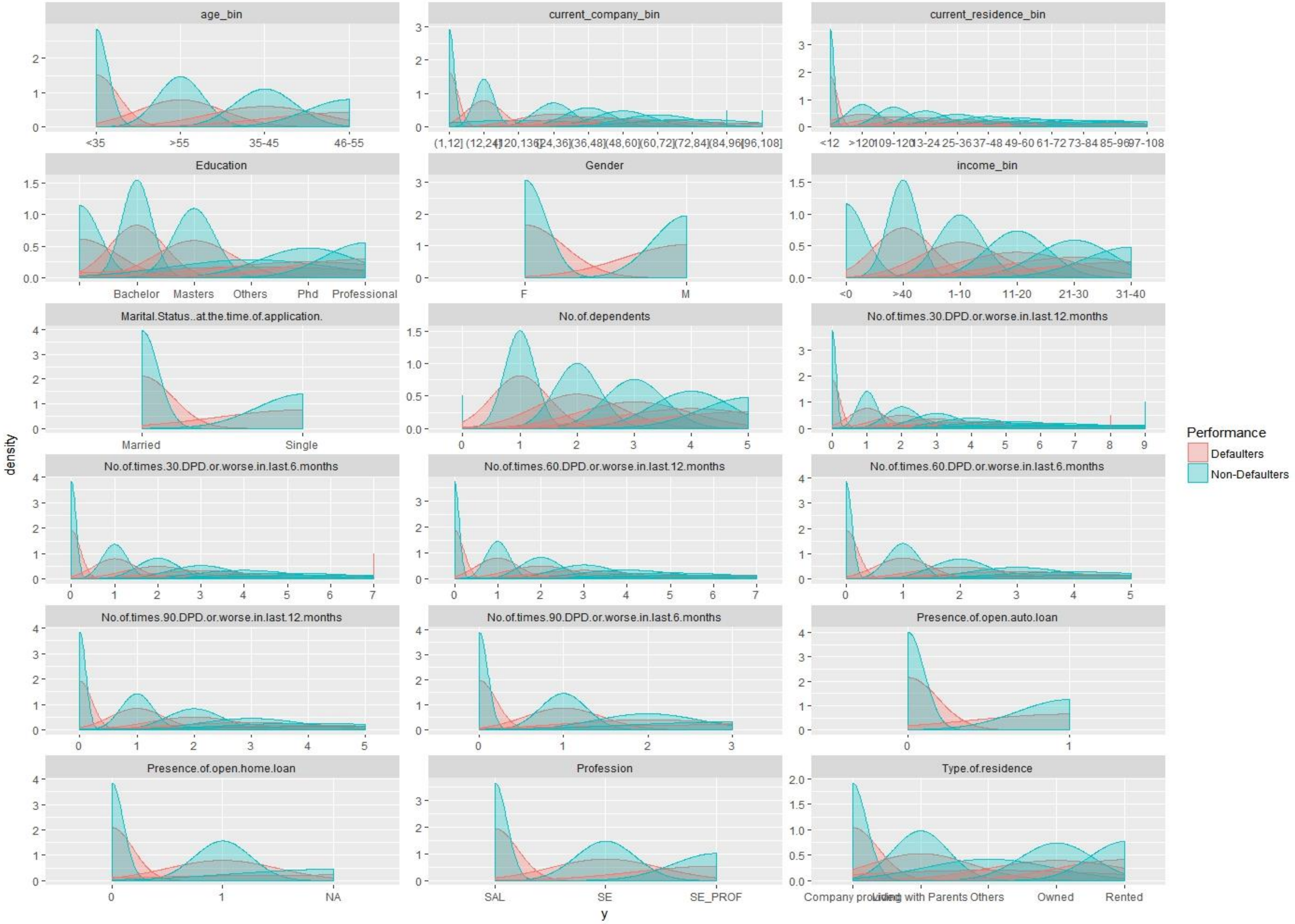


Inferences

- *Avgas CC Utilization in last 12 months* - Most of the defaults are with ***Utilization are <20***
- *OutStanding Balance* is significant but higher the value doesn't reflect the cause of high default

Exploratory Data Analysis – Bi-Variate

Bi Variate Analysis against Default Vs Non-Default

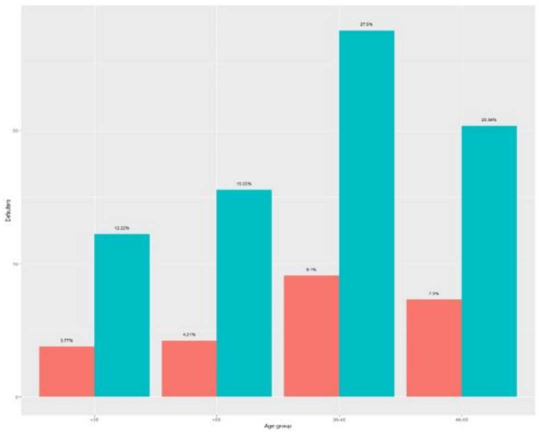


Inferences

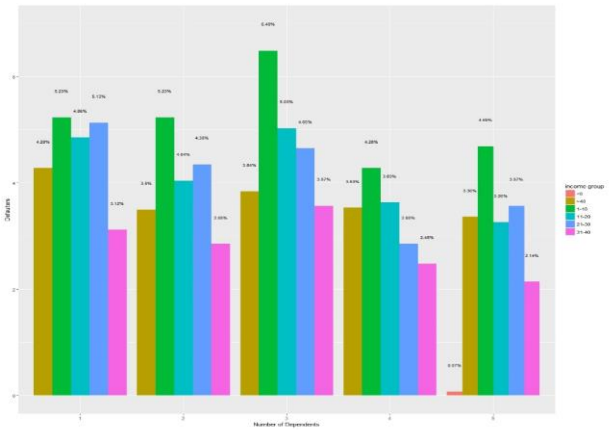
- *Age* - group of ***35-55*** is significant
- *Marital Status* - ***Married*** Significant
- *Profession* - ***Salaried*** is significant with high frequency
- *Type of residence* - *Rented* is the most significant with high
- *No of months in current residence –* ***<12 Months*** is high frequency
- *No of months in current company –* ***<24 Months*** has significant default
- All DPDs - Higher the number has default effect

Exploratory Data Analysis – Multi Variate

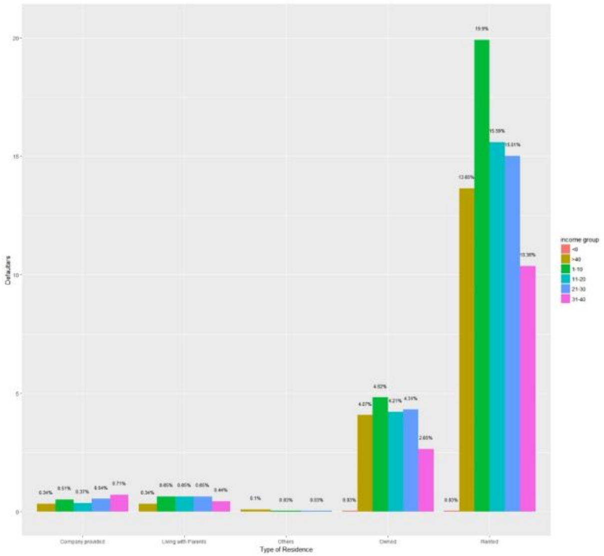
Age Group Vs Gender Vs Performance



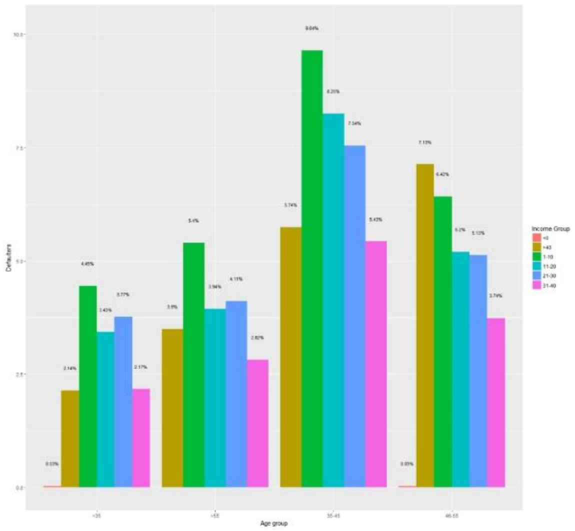
No. of. Dependents Vs Income Group Vs Performance



Type of Residence Vs Income Group Vs Performance



Age Group Vs Income Group Vs Performance

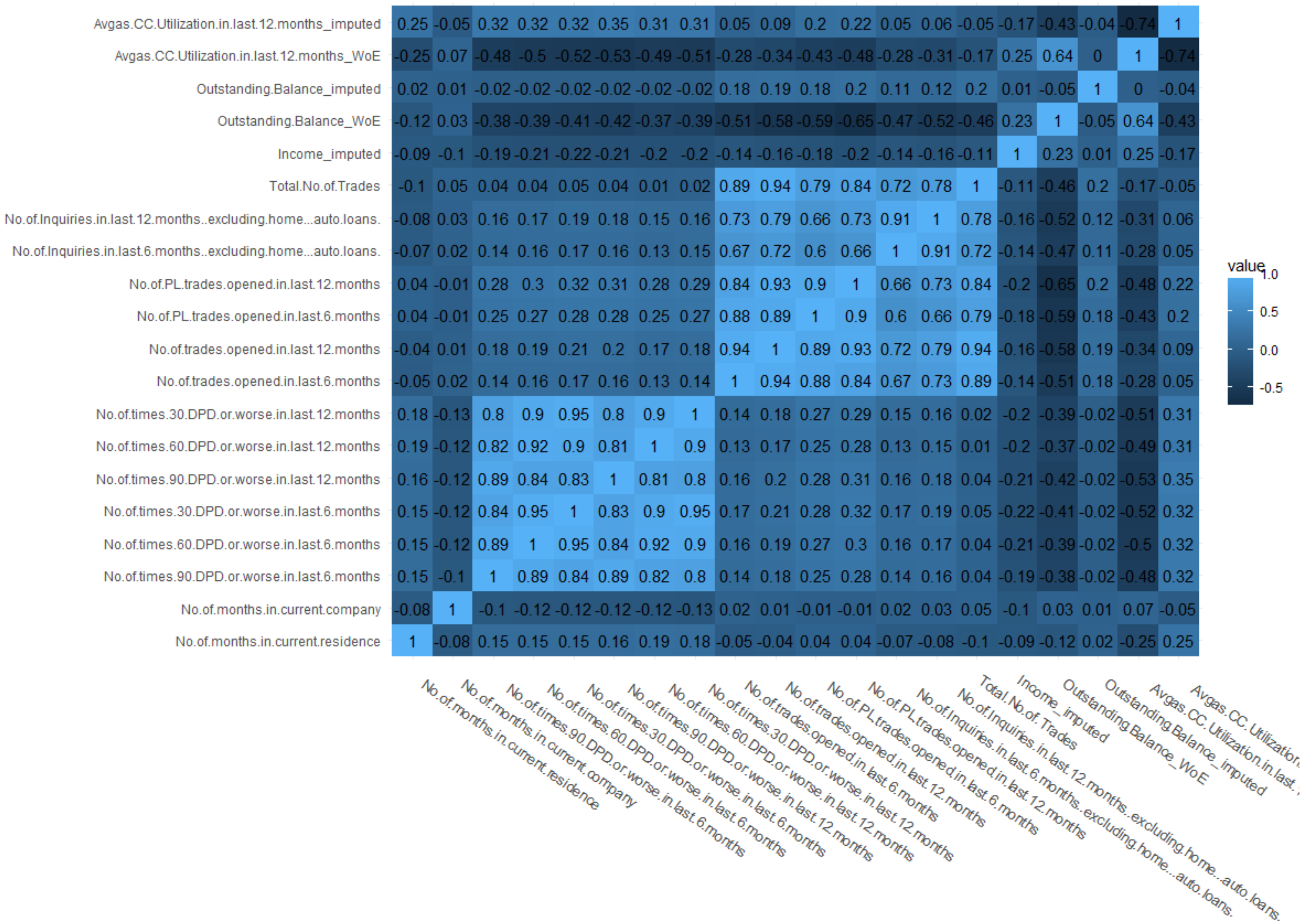


Inferences

- *Residence Type* ***Rented*** across all Income Group has higher rate of default
- *Income Group* **[1-10]** is highest across all Age Groups. The Lower the *Income* high the *default rate*
- *Gender* has no significance across *Age Groups* for defaults
- *No. of Dependants* has no significance across *Age Groups* for defaults

Exploratory Data Analysis – Correlation Analysis

Correlation Matrix for Continuous variables based IV>=0.02



Inferences for Feature Selection

- All ***DPD*** columns are highly correlated (0.8 to 0.95), choosing one variable with high ***IV*** value range (0.02 to 0.5) i.e. ***No.of.times.30.DPD.or.worse.in.last.6.months***
- All ***Trades*** related features are highly correlated 0.6 to 0.94, choosing only with high ***WoE*** value i.e. ***No.of.trades.opened.in.last.12.months***
- **Feature *Outstanding.Balance_WoE*** is correlated (0.65) with ***Avgas.CC.Utilization.in.last.12.months_WoE*** but considered based on business intuition. Also considered ***Outstanding.Balance_imputed***
- Discarding both ***Presence.of.open.home.loan_WoE*** and ***Presence.of.open.home.loan_imputed*** because they both are highly correlated (0.93 and 0.94 respectively) with ***Outstanding.Balance_imputed***

Model Building and Evaluation Methodology

Model Building
This problem belongs to supervised and binary classification problem with ***Performance Tag*** as the target variable.

Model Selection
As it is a binary classification problem we started used following 3 modelling techniques
- *Logistic Regression*
- *Decision Tree*
- *Random Forest*

Data Sampling with Stratified Partitioning of Train/Test datasets
The data is high imbalanced with approximately 96% applicants are non-defaults and only 4% with defaulters. Following varieties of data sets are used for building models

- Original *Unbalanced* Data
- *Under* Sampling
- *Over* Sampling
- *SMOTE* Sampling

Multiple Models build with Cross Validation
Multiple models are developed for choosing best one
- *Logistic* - Demographic - Unbalanced Data
- *Logistic* - Demographic & Credit Data – *Unbalanced, Under, Over* & *SMOTE*
- *Decision Tree & Random Forest* - Demographic & Credit Data – *SMOTE*

Model Evaluation
Used cross validation extensively for the model evaluation and also the following metrics

Standard Metrics
- *Accuracy*
- *Sensitivity*
- *Specificity*
- *KS Statistic*
- *ROC Curve*

Metrics for Imbalanced
- *F1 Score*
- *AUC*

Model Selection

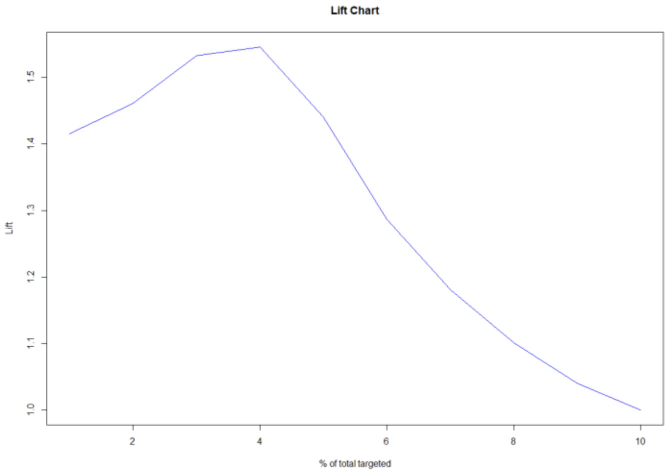
Model Performance Metrics - Lift and Gain calculated for 5th Decile

	Accuracy	Sensitivity	Specificity	F_score	Threshold	AUC	False_positive_Rate	True_positive_Rate	KSStatistic	Lift	Gain
FullData - GLM - Unbalanced	0.6333636	0.6172140	0.6340746	0.04046375	0.049	0.6256443	0.3659254	0.6172140	0.26664225	1.49	74.97
FullData - GLM - Over-Sampling	0.6223793	0.6387316	0.6216594	0.04046375	0.532	0.6301955	0.3783406	0.6387316	0.25689175	1.44	72.23
FullData - GLM - Under-Sampling	0.6344620	0.6228766	0.6349721	0.04046375	0.541	0.6289243	0.3650279	0.6228766	0.25322336	1.43	71.80
FullData - GLM - SMOTE-Sampling	0.6271551	0.6387316	0.6266454	0.04046375	0.458	0.6326885	0.3733546	0.6387316	0.21530067	1.27	63.75
DemographicData - GLM - Unbalanced	0.5397583	0.5537939	0.5391404	0.04046375	0.042	0.5464671	0.4608596	0.5537939	0.10734253	1.10	59.91
FullData - RPART - SMOTE-Sampling	0.5839343	0.5900340	0.5836657	0.04046375	0.191	0.5868499	0.4163343	0.5900340	0.11278189	1.08	54.36
FullData - RF - SMOTE-Sampling	0.6070490	0.6013590	0.6072996	0.04046375	0.279	0.6043293	0.3927004	0.6013590	0.06415095	1.03	51.64

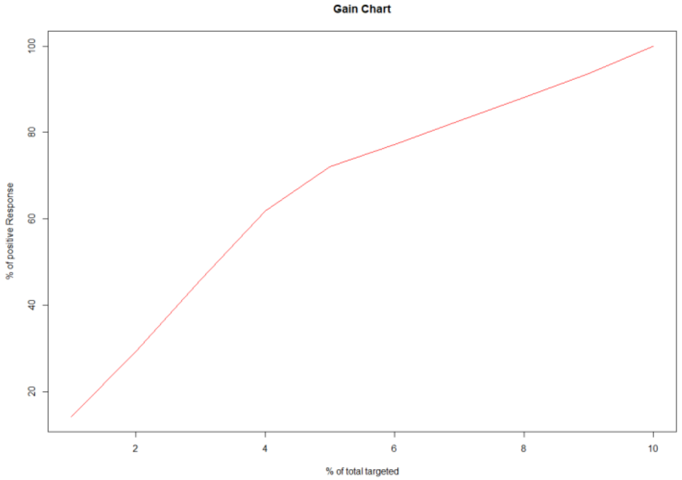
Best Model Selected

Logistic Regression model with unbalanced data performed better than other models

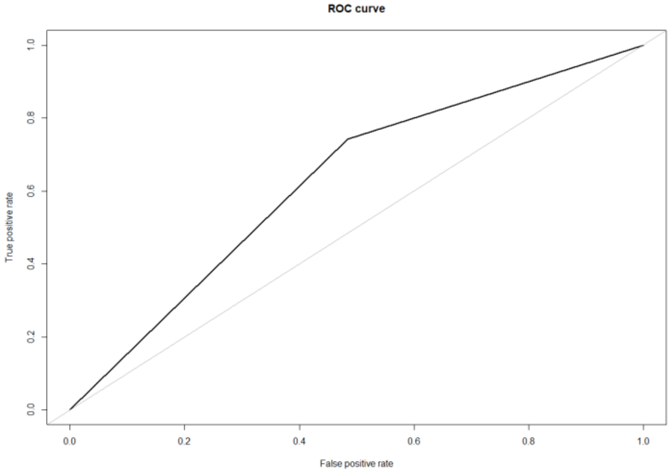
Lift Chart– 1.49



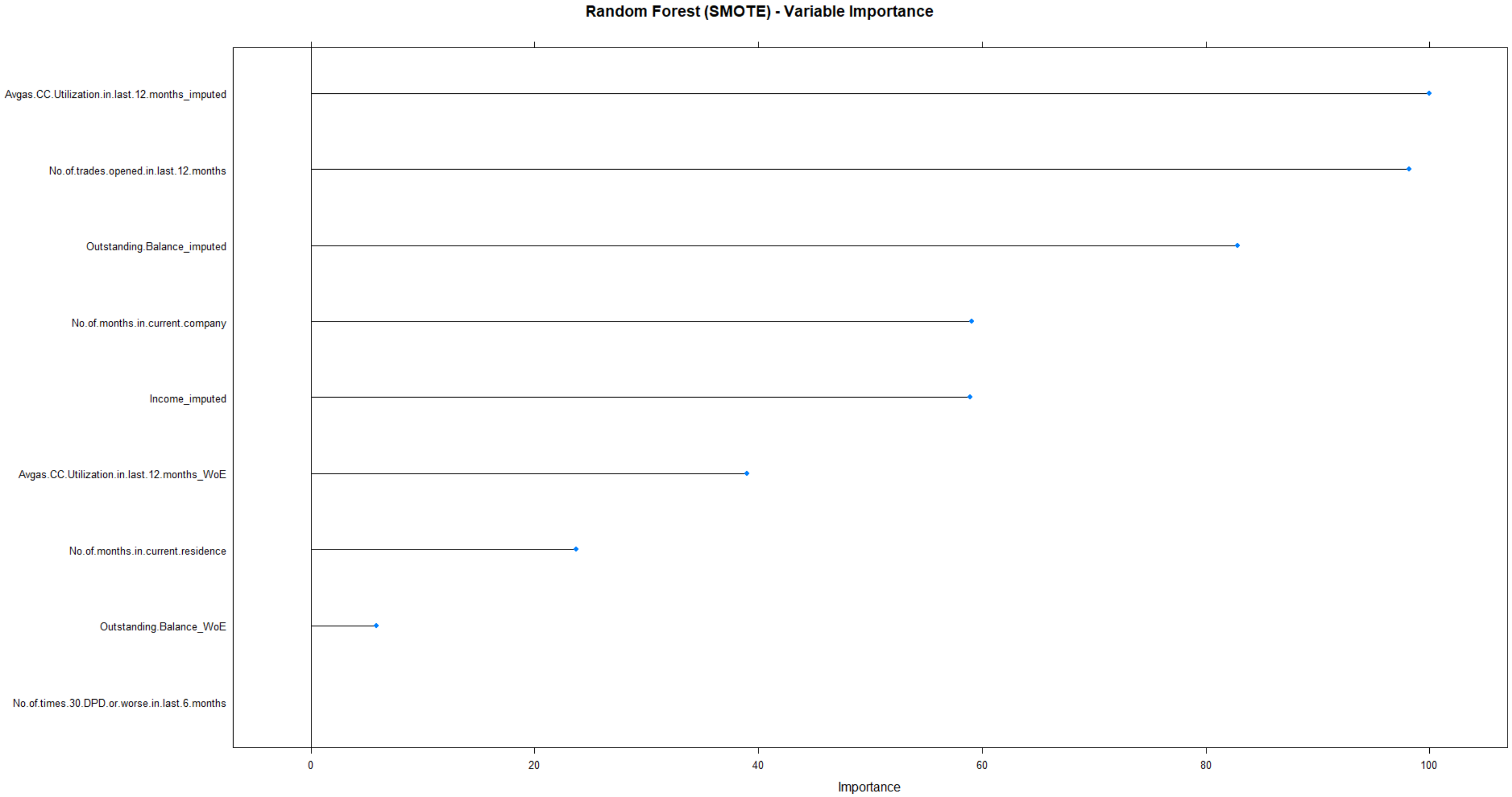
Gain Chart – 74.97



ROC Chart



Variable Importance Plot Random Forest



Plot is for reference w.r.t influencing default likelihood

Model Deployment Using Application Scorecard

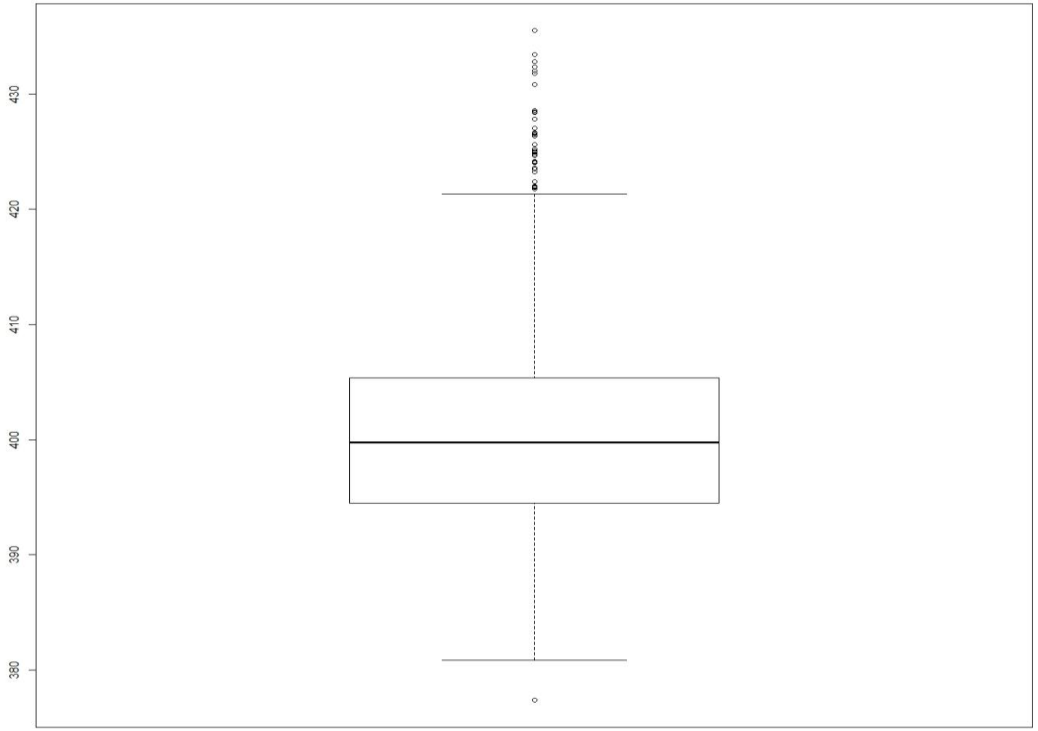
The Application Scorecard built with the ***Good to Bad*** odds of 10 to 1 at a score of 400 doubling every 20 points

Observations from the application scorecard

- **Cutoff score used for separating good vs bad customer is 419.**
- **Auto approval rate is 62.64%**
- **Wrong prediction with the model 36.40%**

Sample Application score on Rejected data

Application.ID	Bad	Good	Odds	LogOdds	Score
207075	0.09	0.91	10.51	2.35	401.43
4498953	0.07	0.93	12.42	2.52	406.24
5976236	0.09	0.91	9.72	2.27	399.19
6353025	0.09	0.91	10.63	2.36	401.76
6663850	0.07	0.93	14.00	2.64	409.72
7295535	0.11	0.89	7.75	2.05	392.63
8121487	0.06	0.94	15.32	2.73	412.31
8365035	0.10	0.90	9.52	2.25	398.58
8367698	0.09	0.91	9.82	2.28	399.46
8893148	0.09	0.91	10.67	2.37	401.87
9012681	0.12	0.88	7.69	2.04	392.40
9061459	0.11	0.89	8.43	2.13	395.09
9339648	0.09	0.91	10.49	2.35	401.39
10551219	0.08	0.92	12.22	2.50	405.77



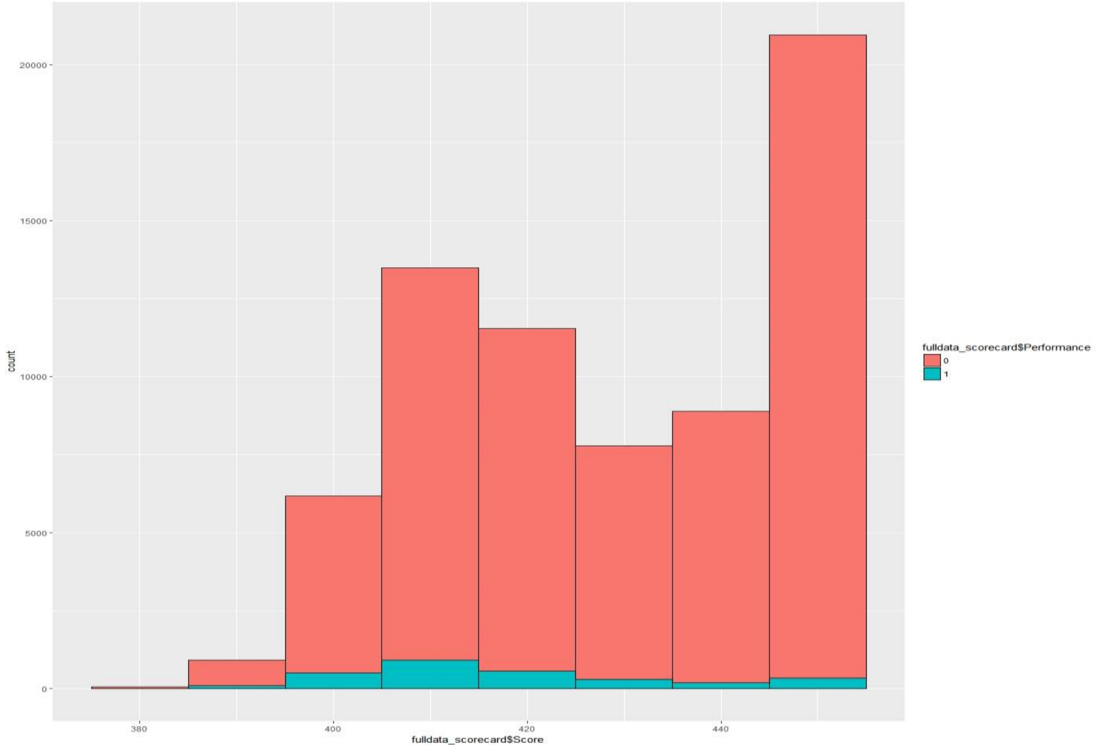
Plot - The score on the Rejected application data shows that there are some good customers (non-defaulters) in the rejected data

Assessment of Financial Benefits

Insights from Application Scorecard

- The histograms plots indicates that the number of defaulters decreases after ***Cut-off Score of 419***
- Even though ***419*** is boundary value with *Good and Bad* Customers, we can suggest that the boundary range of customers fall between *Good and Bad*.

This can be interpreted from the box plot of Application Score.



Potential Credit Loss

Calculated on Full data

Total prospect loss = 2634047450

(Prob of bad * Exposure at default * Loss given default)

Expected loss by default customer from model 147718048

- The loss amount of 147718048 can be straight away avoided by not giving loan to default customer prospects
- However, by looking into the application score card, some customers of default category can be consider at medium risk because they fall in the boundary range.
- This potential credit loss can be minimized by target those customer, which Credit Score falls within Good and Intermediate.
- The verification / acquisition cost of Bad Customer can be minimized by this Model

Rejected data

Total prospect loss = 96026810

(Loss because of the full rejected data)

Loss due of Rejection of *Good* customers is 43876837

The amount of ***43876837*** would have been gained on using the model because it was the loss by rejection the good customers