Acquisition **A**nalytics for CredX

Submitted By: Rahul Dalvi
Ravi Gandhi
Sanjay Kushwah
Winnie Unnikrishnan

BUSINESS OBJECTIVE

CredX wants to determine the factors affecting credit risk. Thereafter create strategies to mitigate the risk and assess the financial benefit of the risk model.

Demographic Data

- Age
- Gender
- Marital Status
- Dependents
- Income
- Education
- Profession
- Residence Type

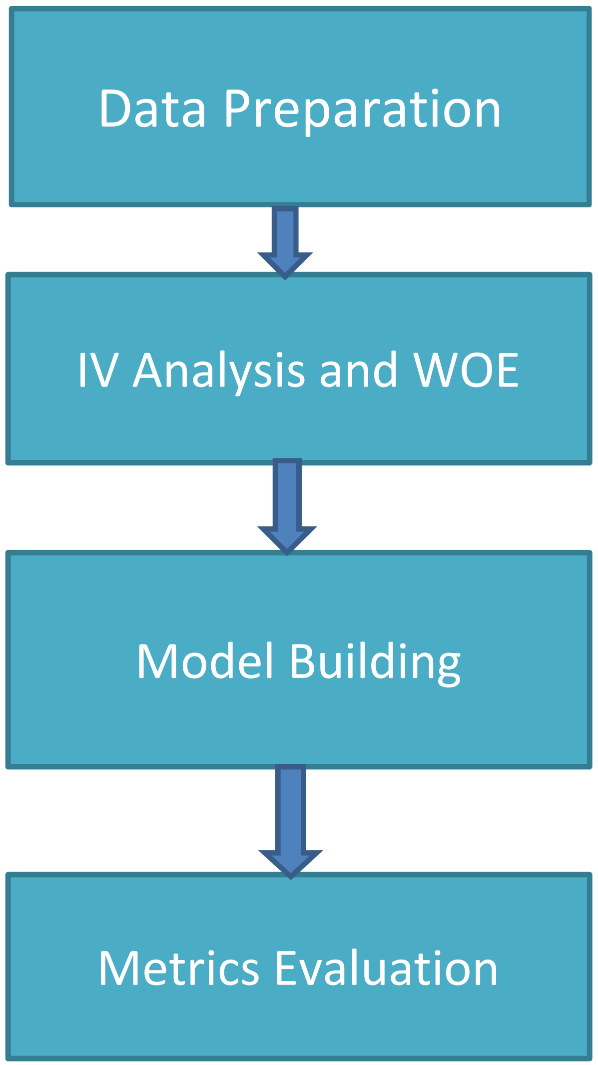
+

Credit Bureau Data

- Outstanding Balance
- DPDs
- Loan Inquiries
- Open Home Loan
- Credit Card Utilization
- No. of Trades

Methodology	CRISP-DM Framework
Task	Predict Performance / Credit Default
Data set	Demographic & Credit Bureau
Features	As per data set. Data Dictionary available
Models	Supervised Classification Models like Logistic Regression, Decision Tree & Random Forest

PROBLEM SOLVING METHODOLOGY



1. Merge Demographic and Credit Bureau Data on Applicant ID
2. Remove Duplicate Rows
3. Explore data by Univariate and Multivariate Analysis
4. Check NAs and NANs
5. Outlier Treatment

6. Compute Information Value
7. Populate features with their WOE values
8. Observe variability using Principal Component Analysis
9. Combining IV Analysis with Variable Clustering
10. Create data frames with and without WOE Values

11. Split data into train and test
12. Build models on Demographic Data and on All data separately
13. Obtain Performance Tag for 1425 rows having NAs in Performance Tag and merge with all data file

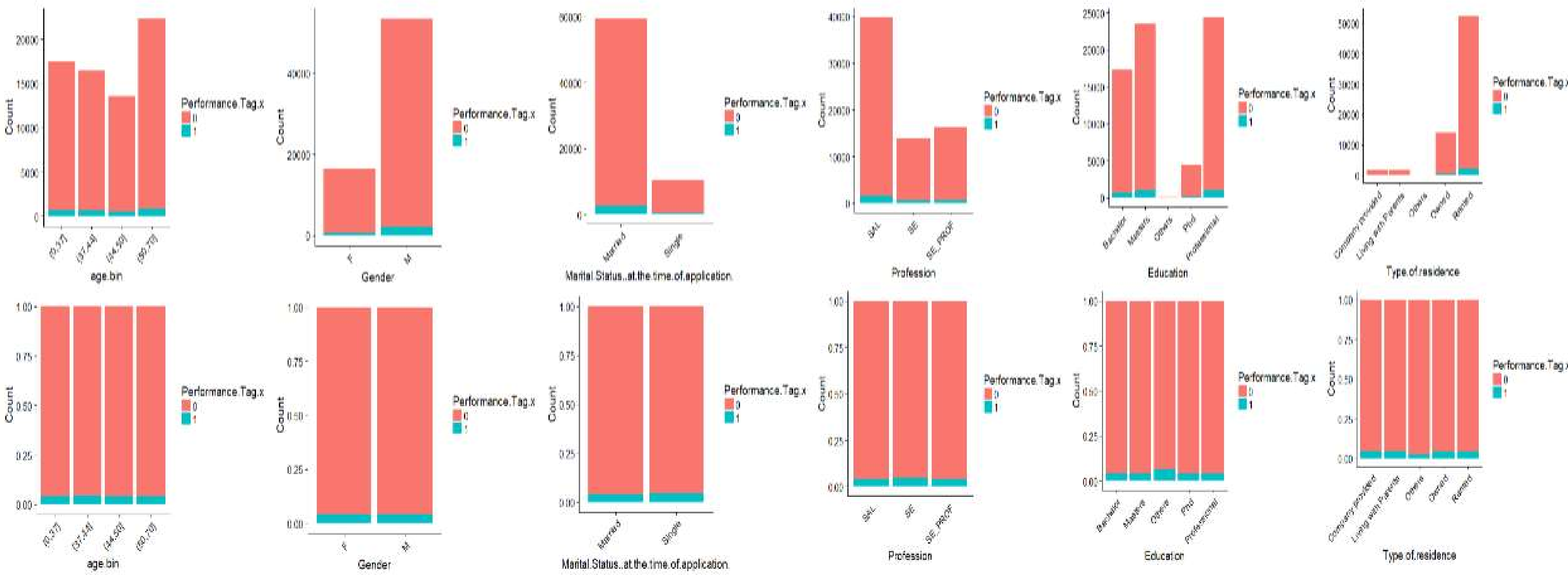
14 Rebuild Models using Logistic Regression, Decision Trees, Random Forest

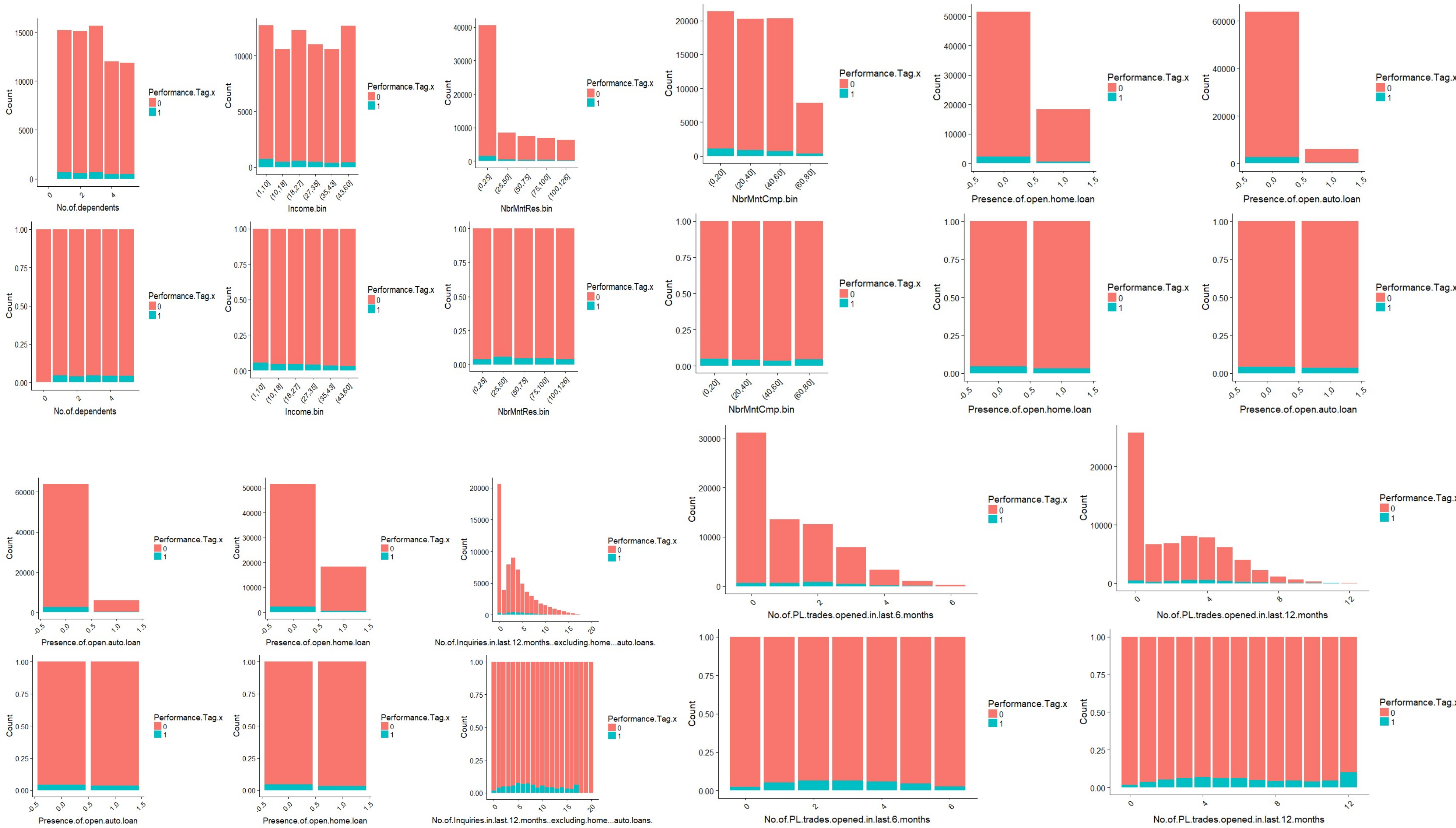
15. Check Model on Test Data
16. Evaluate Model Metrics
17. Build Application Score Card from Logistic Regression Model
18. Build Financial Strategies using the optimum model

TOP VARIABLES WITH MOST INFORMATION VALUE

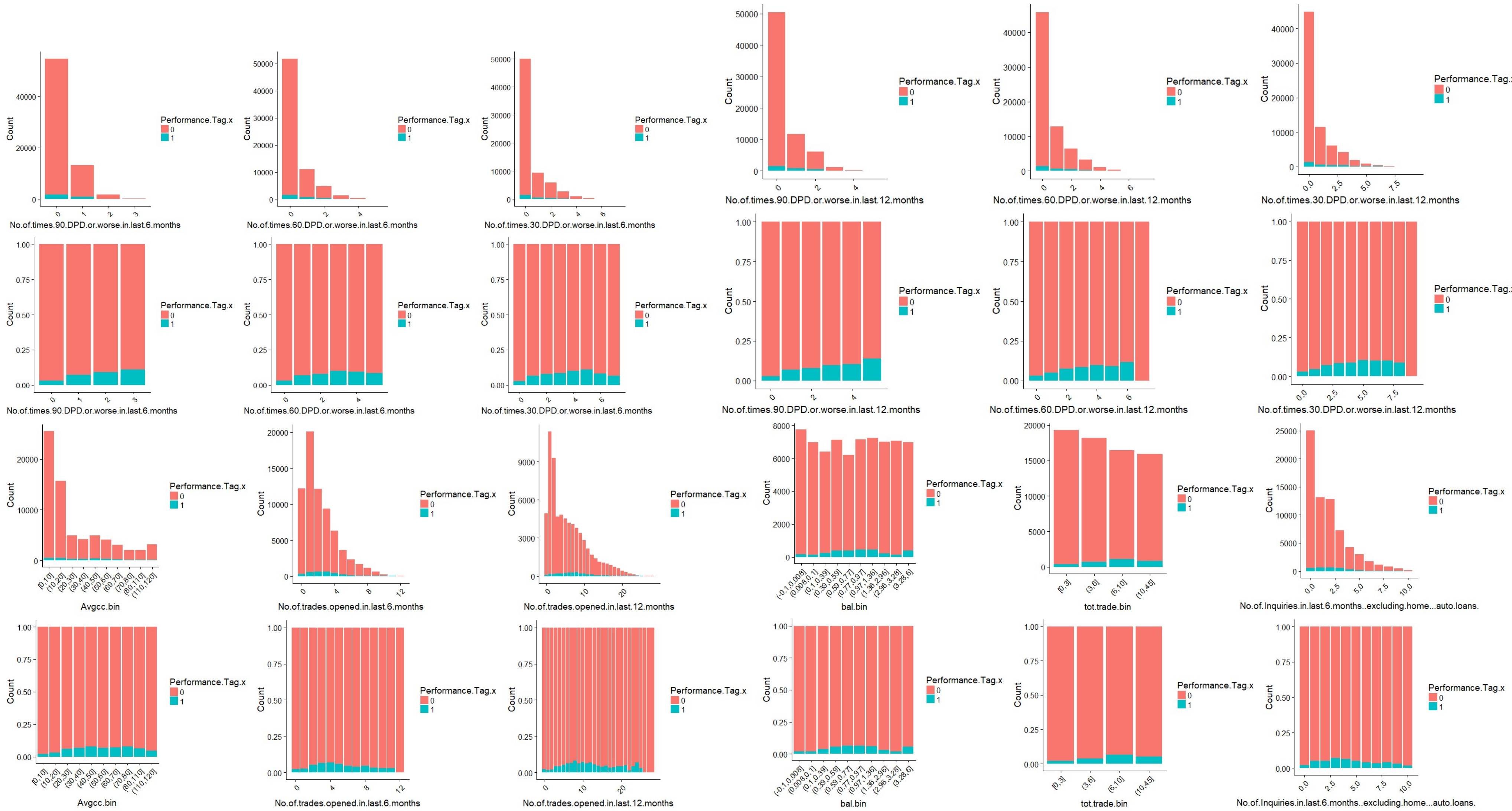
Variable	IV
Avgas.CC.Utilization.in.last.12.months	0.29952926
No.of.trades.opened.in.last.12.months	0.29812273
No.of.PL.trades.opened.in.last.12.months	0.29607072
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.29558254
Outstanding.Balance	0.24426983
No.of.times.30.DPD.or.worse.in.last.6.months	0.24174638
Total.No.of.Trades	0.23677017
No.of.PL.trades.opened.in.last.6.months	0.21980829
No.of.times.90.DPD.or.worse.in.last.12.months	0.21400973
No.of.times.60.DPD.or.worse.in.last.6.months	0.20599241
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans	0.20535587
No.of.times.30.DPD.or.worse.in.last.12.months	0.19841516
No.of.trades.opened.in.last.6.months	0.18614455
No.of.times.60.DPD.or.worse.in.last.12.months	0.18560548
No.of.times.90.DPD.or.worse.in.last.6.months	0.16022924
No.of.months.in.current.residence	0.0789817
Income	0.04354888
No.of.months.in.current.company	0.02176916

EXPLORATORY DATA ANALYSIS

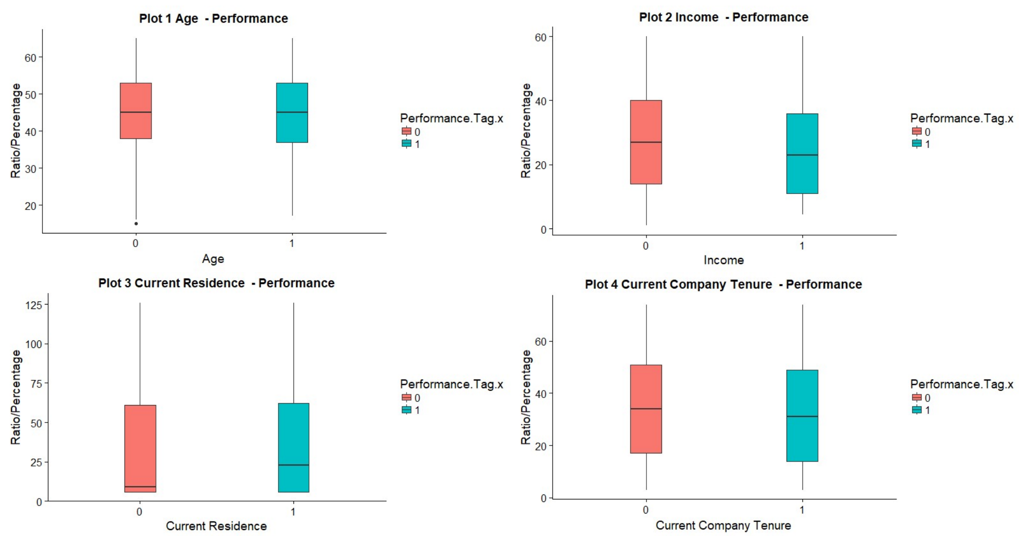


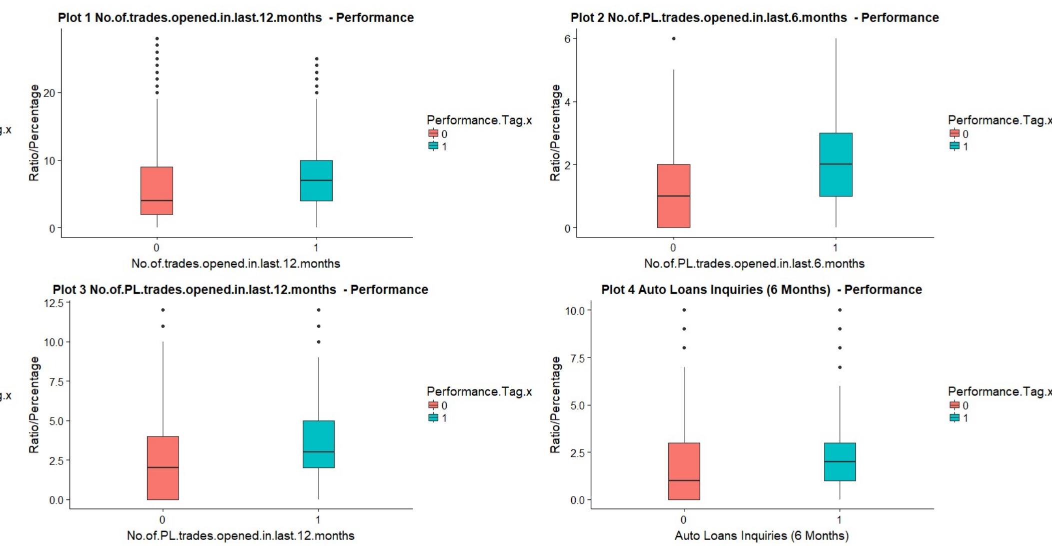


Performance.Tag.x
0
1
Count
No.of.dependents
Income.bin
NbrMntRes.bin
NbrMntCmp.bin
Presence.of.open.home.loan
Presence.of.open.auto.loan
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
No.of.PL.trades.opened.in.last.6.months
No.of.PL.trades.opened.in.last.12.months



Count
Performance.Tag.x
0
1
No.of.times.90.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.6.months
No.of.times.90.DPD.or.worse.in.last.12.months
No.of.times.60.DPD.or.worse.in.last.12.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.60.DPD.or.worse.in.last.6.months
No.of.times.30.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.months
No.of.times.30.DPD.or.worse.in.last.12.months
Avgcc.bin
No.of.trades.opened.in.last.6.months
No.of.trades.opened.in.last.12.months
bal.bin
tot.trade.bin
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
Avgcc.bin
No.of.trades.opened.in.last.6.months
No.of.trades.opened.in.last.12.months
bal.bin
tot.trade.bin
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

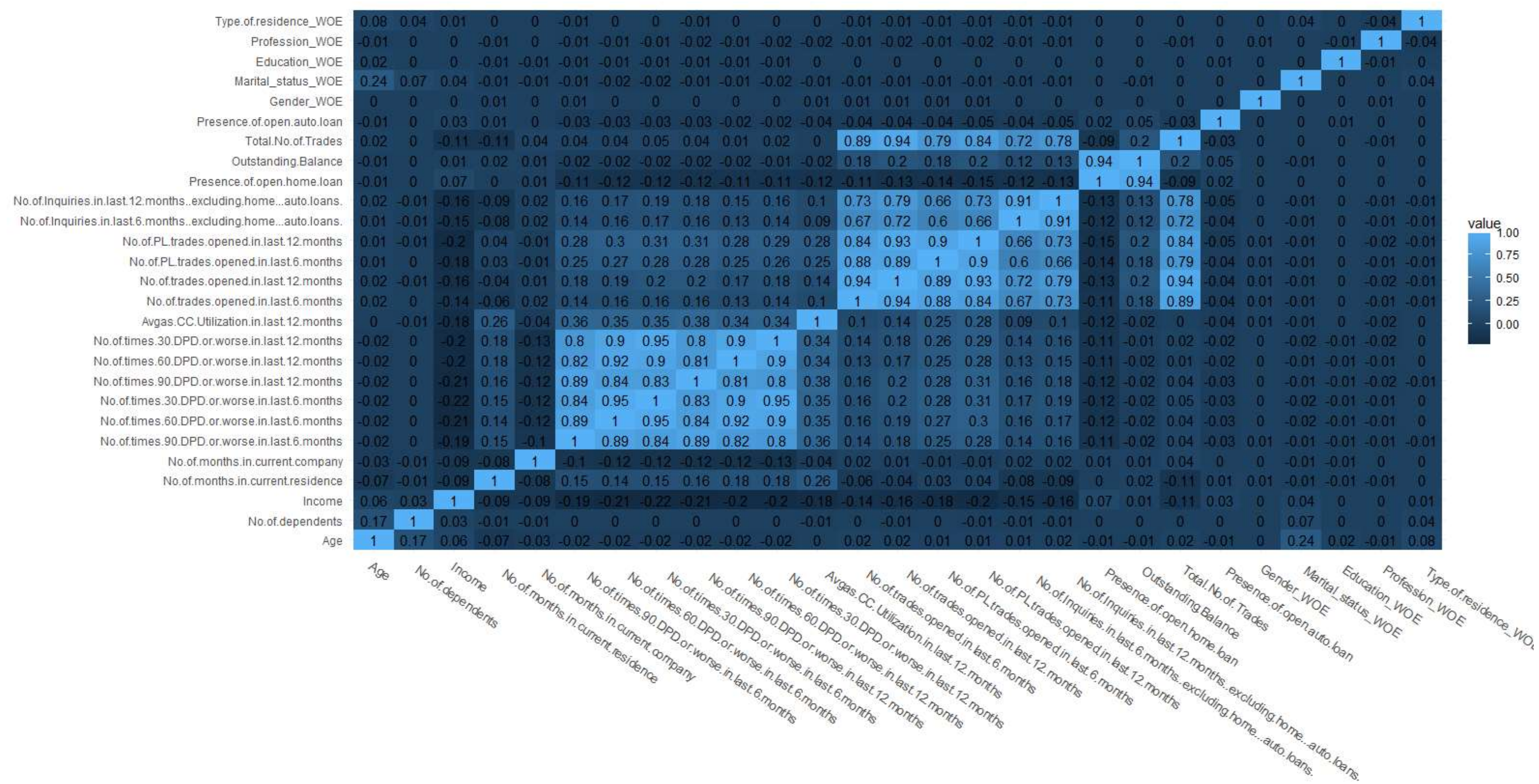




Insights from EDA

- With increase in outstanding balance default rate increases
- DPD 6 months and 12 months for 30/60/90 days is a significant variable. With the increase in No. of DPD the Default rate increases
- Default rate increases for low income group.
- As the No. of enquiries increases in 6/12 months default rate increases
- Type of residence - Rented is the most significant with high applicants
- With the increase in the No. of PL Trade opened in 6/12 months the default rate increases
- With increase in Avg. CC utilization the default rate increases
- No. of open trades shows the positive correlation to the default rate
- ***Majorly it's looks like the credit bureau data is driving the default rate***

EXPLORATORY DATA ANALYSIS – CORRELATION MATRIX



Insights from EDA

- DPD variables for 30/60/90 days in 6/12 months seems highly correlated with each other
- Presence of open home loan and Outstanding balance is highly correlated
- No. of inquiries made and total number of trades is highly correlated. This implies that People with high trading habits are inquiring more. This is obvious and insightful.

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

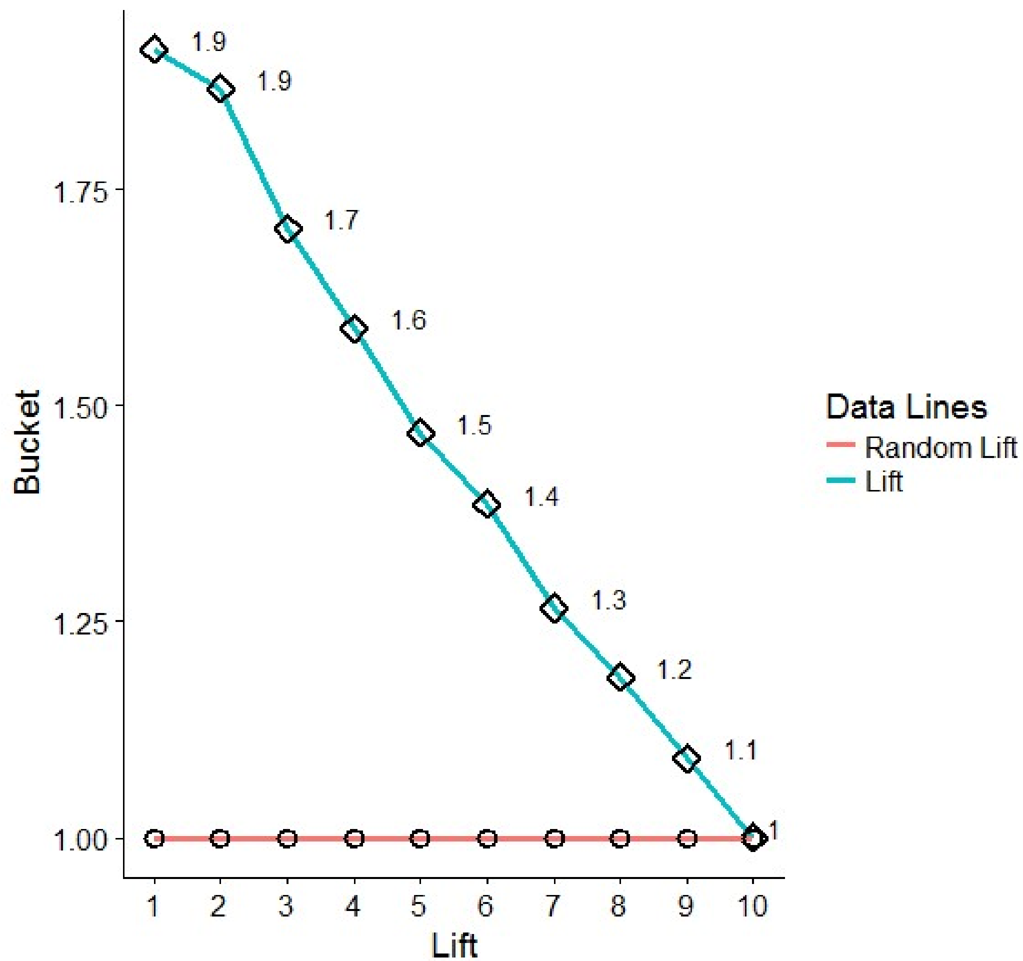
- Original Unbalanced Data
- Under Sampling
- Over Sampling
- SMOTE Sampling

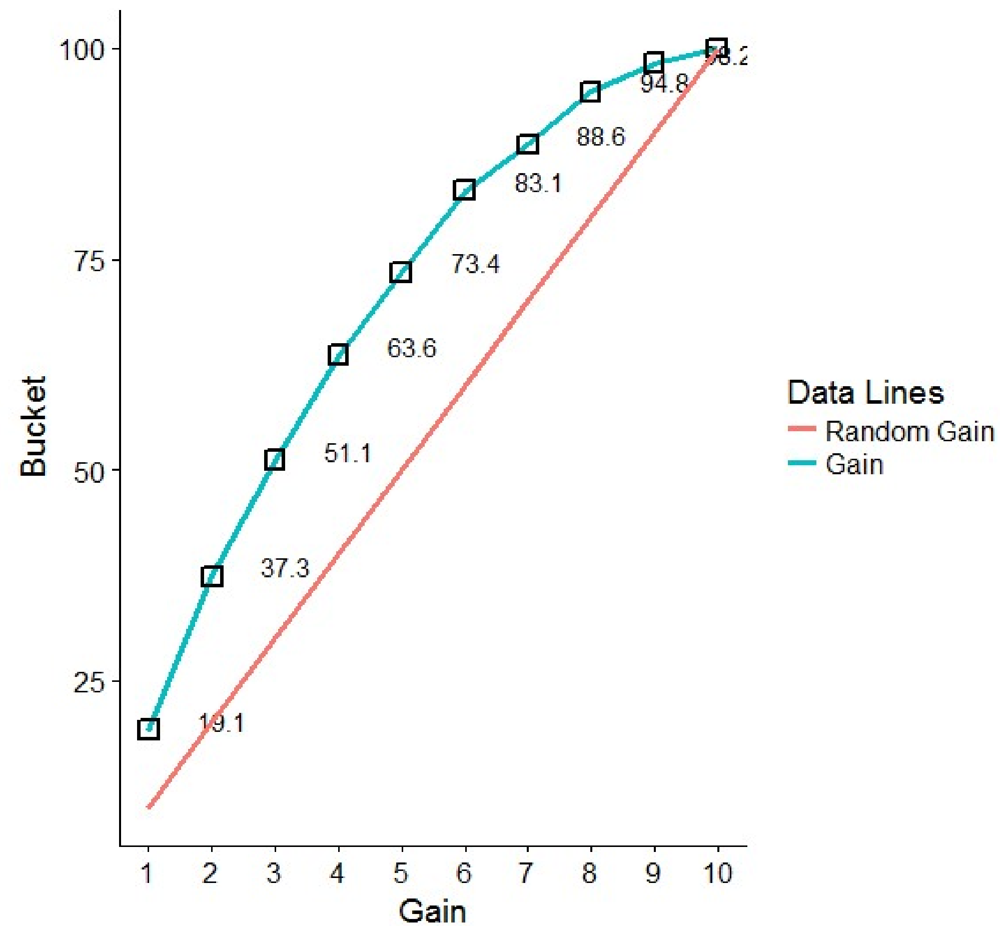
MODEL SELECTION

Model Performance Metrics with Lift and Gain Charts

Model	Accuracy	Sensitivity	Specificity	AUC	False_Positive_Rate	True_Positive_Rate
Logistic Regression Model	0.6201021	0.6176471	0.6202103	0.618929	0.3797897	0.6176471
GLM Down Model	0.3825705	0.3766968	0.3828292	0.620237	0.3766968	0.6171708
GLM Up Model	0.3791343	0.3800905	0.3790921	0.620409	0.3800905	0.6209079
GLM SMOTE Model	0.3823796	0.3812217	0.3824306	0.618174	0.3812217	0.6175694
Decision Tree Model	0.4116833	0.4355204	0.4106333	0.576923	0.4355204	0.5893667
Random Forest SMOTE Model	0.414165	0.4061086	0.4145199	0.589686	0.4061086	0.5854801

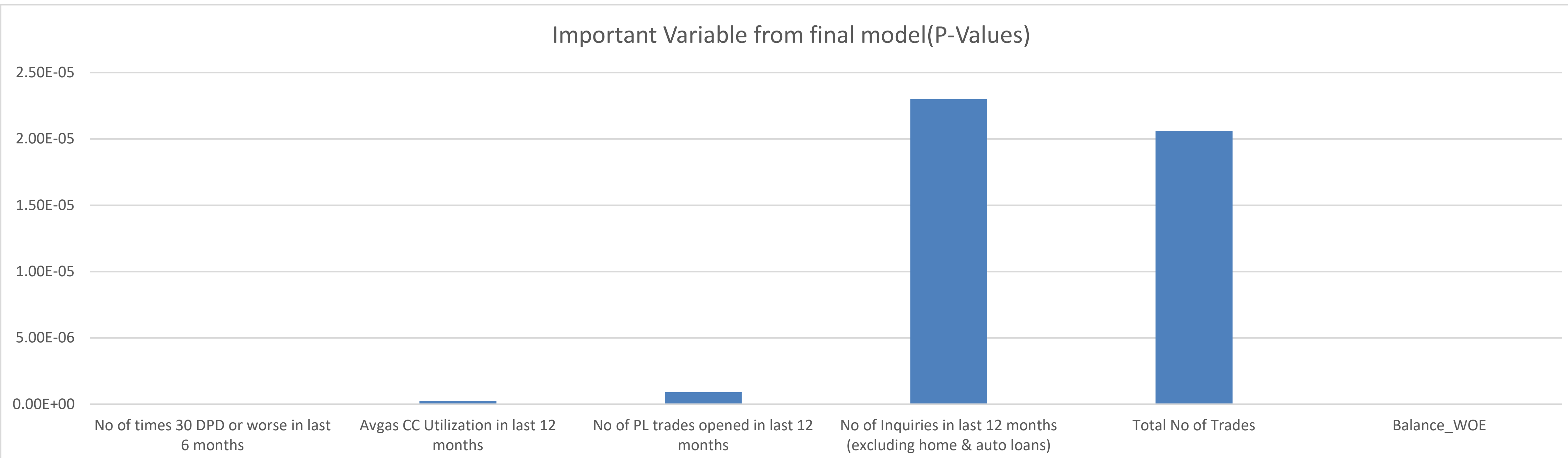
Normal logistic regression model performs the best with highest accuracy and specificity





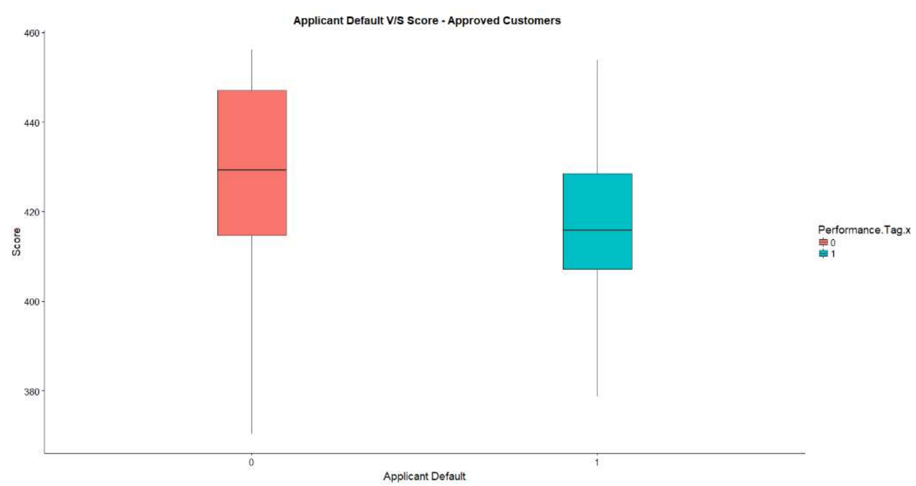
IMPORTANT VARIABLES FROM FINAL MODEL

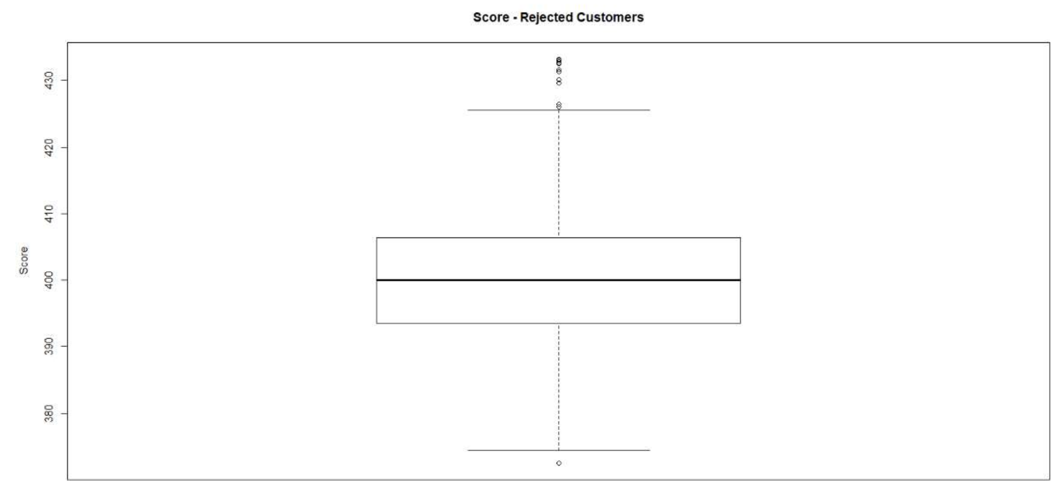
#Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Importance
(Intercept)	-3.59709	0.061972	-58.044	2.00E-16	***
No of times 30 DPD or worse in last 6 months	0.198048	0.02025	9.78	2.00E-16	***
Avgas CC Utilization in last 12 months	0.004506	0.000872	5.166	2.39E-07	***
No of PL trades opened in last 12 months	0.104814	0.021349	4.91	9.12E-07	***
No of Inquiries in last 12 months (excluding home & auto loans)	0.041597	0.009825	4.234	2.30E-05	***
Total No of Trades	-0.03257	0.00765	-4.258	2.06E-05	***
Balance_WOE	-0.41914	0.070186	-5.972	2.35E-09	***



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 421.
- Auto approval rate is 61.33%
- Misclassification using this model is 37.6%
- The 2% of the rejected population are good customers.
- The median score of approved customers is 428 as compared to 400 for rejected applicants. The 3rd quantile of approved customers is 446 as compared to rejected population with 3rd quartile as 406 i.e. way below cutoff score. As seen in the charts above, the application score card clearly helps us in identifying the default customers (from the chart of approved customers).

ASSESSMENT OF FINANCIAL BENEFITS

Current approval system

69844 Approved application

- 4.22% Defaulted
- 3728.39 Million Outstanding balance of defaulters

1425 Rejected applications

Risk Model based approval system

42861 approved applications

- 2.55% default
- Outstanding balance of defaulters reduced to 1386.76 Million

28408 rejected applications

- 2341.62 Million Outstanding Balance of possible defaulters saved

Highlights

- Potential loss can be minimized by using the application score card to classify applicants into Good and Bad

- Rejecting possible bad customers would result in reduced operational cost to avoidance of related verification and application processing cost.