CredX Risk Analytics Case Study

Batch - Dec-2017 **BFS Capstone Project**

Piyush Gaur
Priya Gupta
Sahana K
Ria Nag

Objective & Approach

Problem Statement:

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 due to increase in defaults.

The CEO believes that the best strategy to mitigate credit risk is to acquire "the right customers".

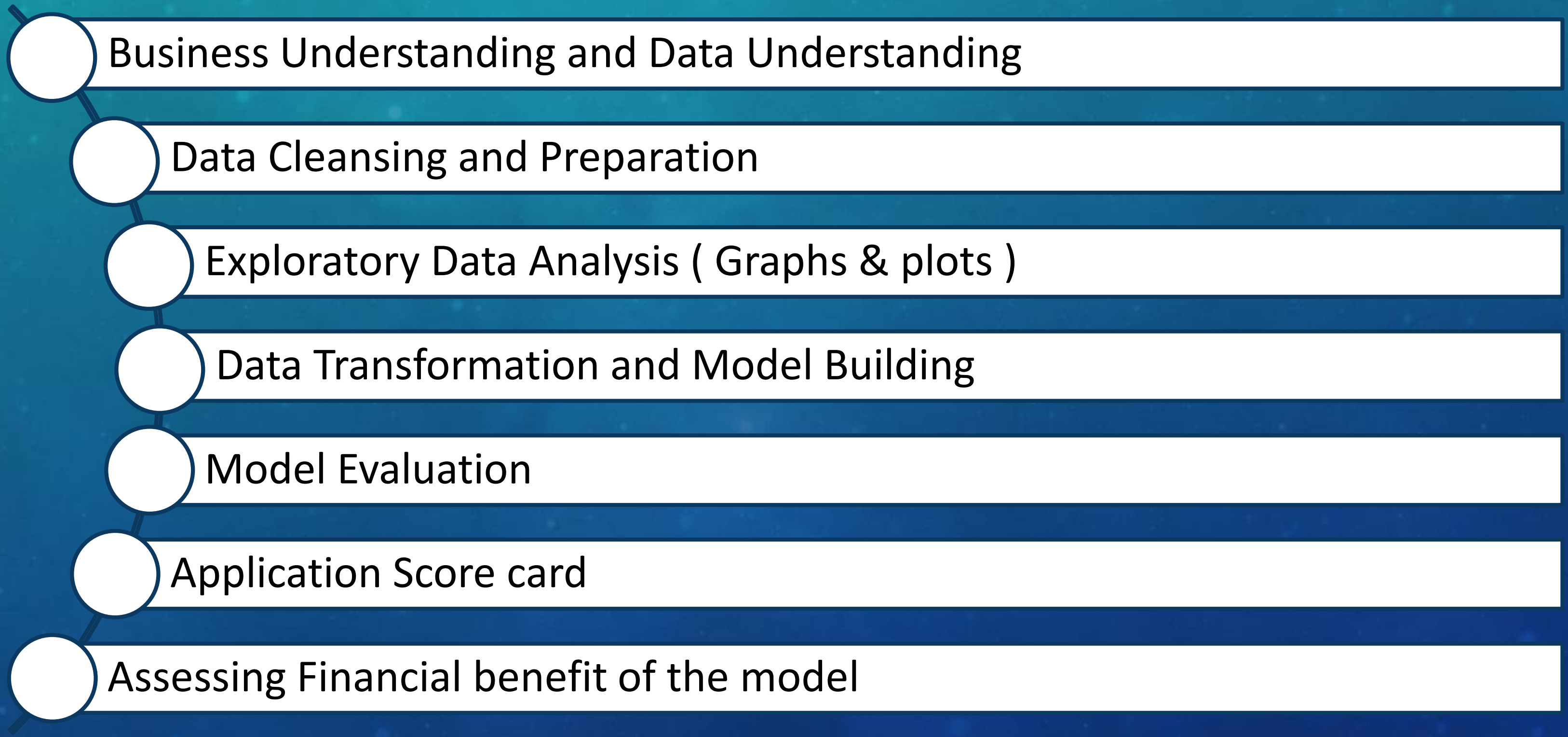
Objective :

The objective is to help CredX identify the right customers using predictive models. We need to determine the factors affecting credit risk and create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Build an application scorecard and identify the cut-off score below which one would not grant credit cards to applicants.

Solution Approach:

This is a binary supervised classification problem. We aim at building models such as Logistic regression, Random forest, SVM and Xgboost to identify the customers who are at a risk of defaulting if offered a credit card. We have followed CRISP–DM framework. It involves the following series of steps:



- Business Understanding and Data Understanding
- Data Cleansing and Preparation
- Exploratory Data Analysis (Graphs & plots)
- Data Transformation and Model Building
- Model Evaluation
- Application Score card
- Assessing Financial benefit of the model

Data Understanding

Demographic Data :

- This information is provided by the applicants at the time of credit card application.
- It contains customer-level information on age, gender, income, marital status, education, Profession and number of dependants.

Credit Bureau Data :

- It provides all details of past transaction.
- This information 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.

Nature of Data :

- The demographic data consists of 71295 observations with 12 variables.
- The credit bureau data consists of 71295 observations with 19 variables.
- Application ID is the common key between the two datasets for merging.
- Performance Tag is the target variable which says if customer is default or not.
- The values are 0(non-default) and 1(default).

Data Quality Issues

- The 1425 rows with no performance tag. Thus we can assume that the applicant is not given credit card, hence they are removed.
- Both occurrences of 3 duplicate Application ID records (765011468, 653287861, 671989187) has been excluded from the dataset.
- Since 18 is the minimum age to grant credit card, records with age <18 has been excluded from the dataset.
- The 1425 rejected records have been saved separately and would be used later to predict if they would default if they were given a credit card using a model made from non rejected records for making a model with better performance and application score card calculations.

Data Quality Issues

Variables	No. of missing values	Erroneous data
Application ID	-	3 Duplicate ID's are present
Age	-	65 records with age <18
Income	-	81 records have income <0
Gender	2	
Marital Status	6	
No of dependents	3	
Education	119	
Profession	14	
Type of residence	8	
Performance Tag	1425	

Variables	No. of missing values	Erroneous data
Application ID	-	3 Duplicate ID's are present
Avgas CC Utilization in last 12 months	1058	
No of trades opened in last 6 months	1	
Presence of open home loan	272	
Outstanding Balance	272	
Performance Tag	1425	

WOE and IV Analysis

- WOE and IV values are calculated for each of the attributes using information and woe.binning package.

- Attributes for which WOE values that were made by default using the Information package in R and were not monotonically changing across bins, coarser bins were made by decreasing the number of bins until monotonic behavior is observed across bins .

- For 9 variables with Missing values, the variable values were replaced by their corresponding WOE values.

- From the IV values we can conclude that parameters in the demographic data don’t play much significant role in prediction and most of the significant variables are from Credit Bureau data.

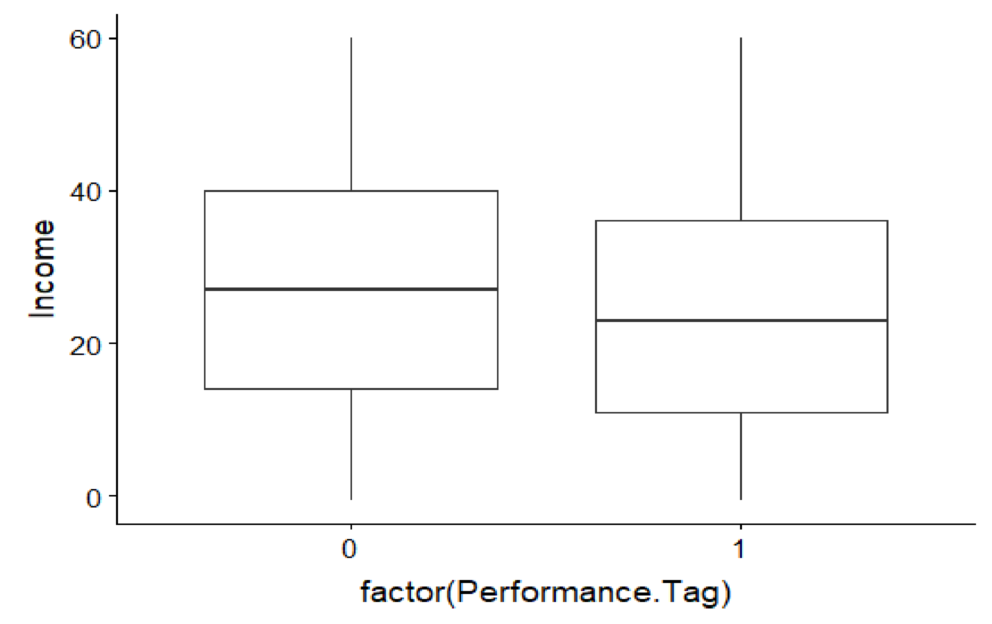
- Top 12 Variables with IV value of 0.1 to 0.3 has medium predictive power and are considered significant. There is no variable with strong predictive power.

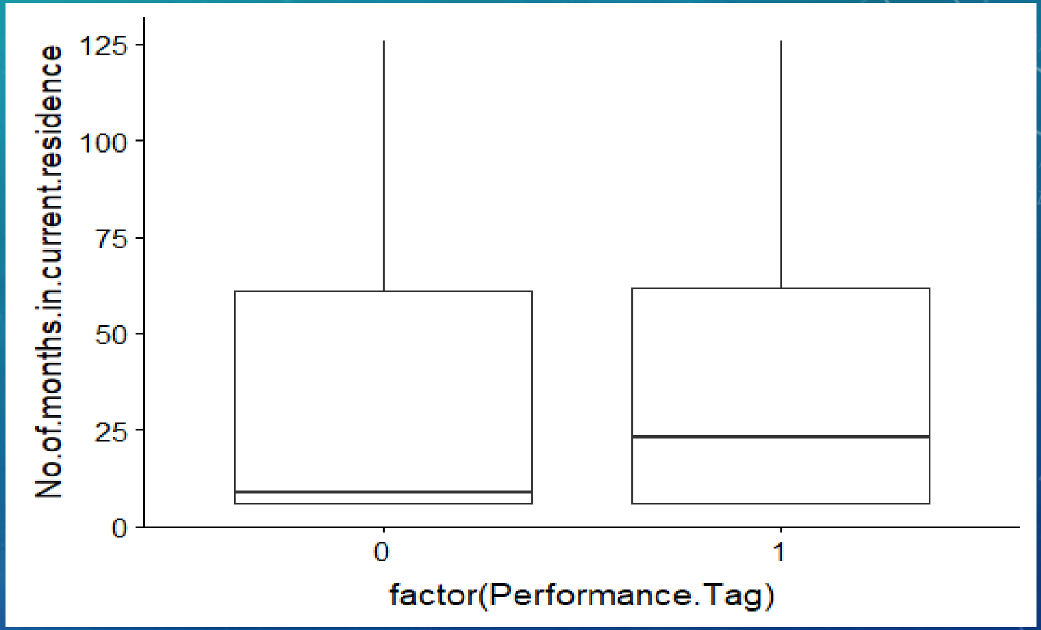
Top 12 variables with highest IV values

Variable	IV
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.2715
Avgas.CC.Utilization.in.last.12.months	0.2607
No.of.times.30.DPD.or.worse.in.last.6.months	0.2415
No.of.times.90.DPD.or.worse.in.last.12.months	0.2138
No.of.times.60.DPD.or.worse.in.last.6.months	0.2058
No.of.times.30.DPD.or.worse.in.last.12.months	0.1982
No.of.trades.opened.in.last.12.months	0.1943
No.of.times.60.DPD.or.worse.in.last.12.months	0.1854
Total.No.of.Trades	0.1822
No.of.PL.trades.opened.in.last.12.months	0.1766
No.of.trades.opened.in.last.6.months	0.1697
No.of.times.90.DPD.or.worse.in.last.6.months	0.1601

Exploratory Data Analysis

Insights derived by EDA : Both univariate and bivariate plots are made to get better insights of all variables of 2 datasets.



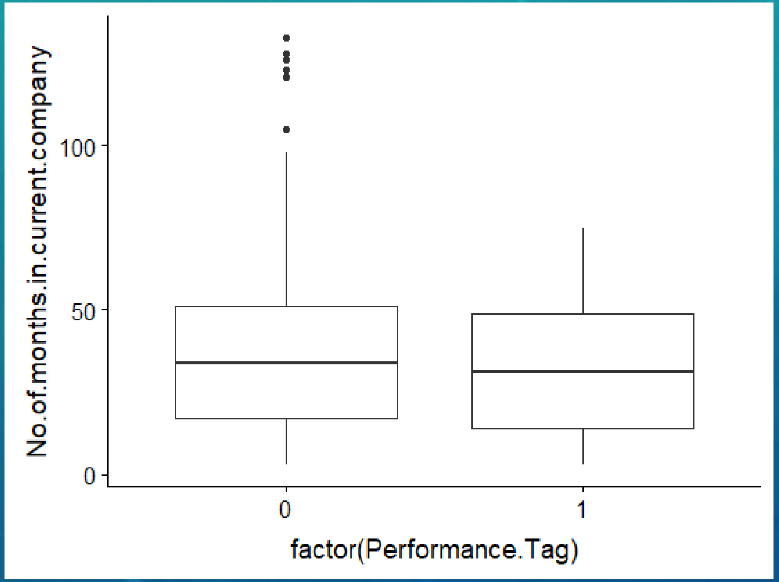


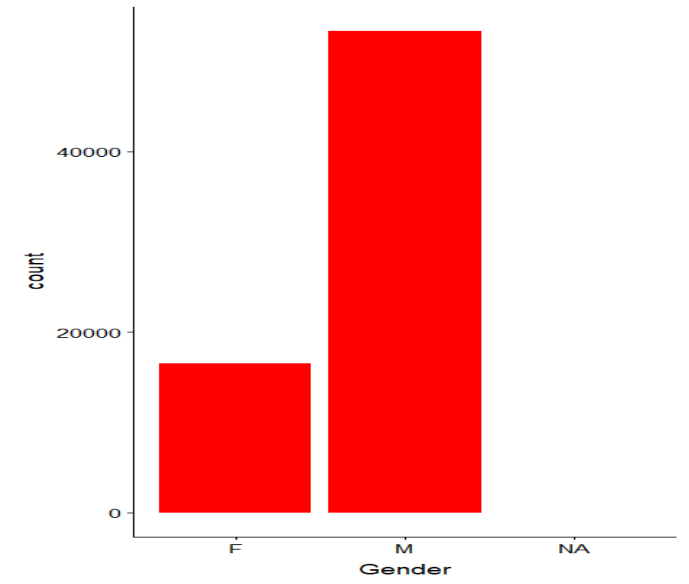
- The median values for income of defaulters are lower than that of non-defaulters

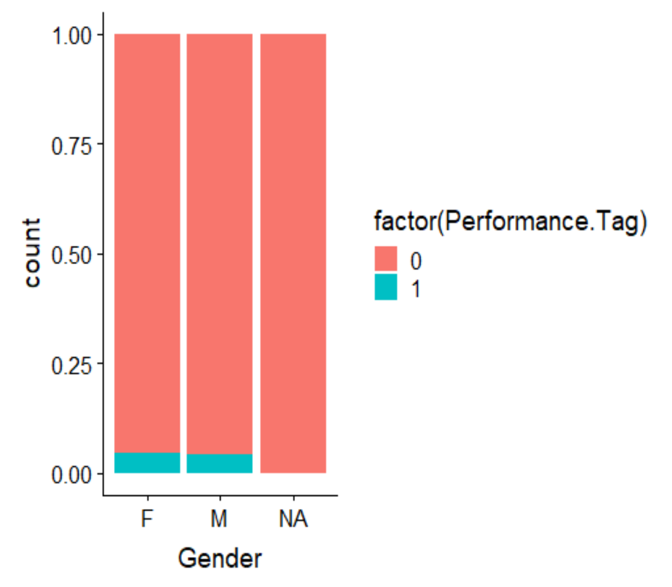
- The median No.of.months.in.current.residence of non-defaulters are lower than that of defaulters.

Insights derived by EDA

Counts in the y axis of Plots in bivariate analysis of categorical variables were normalized so that they have equal heights.



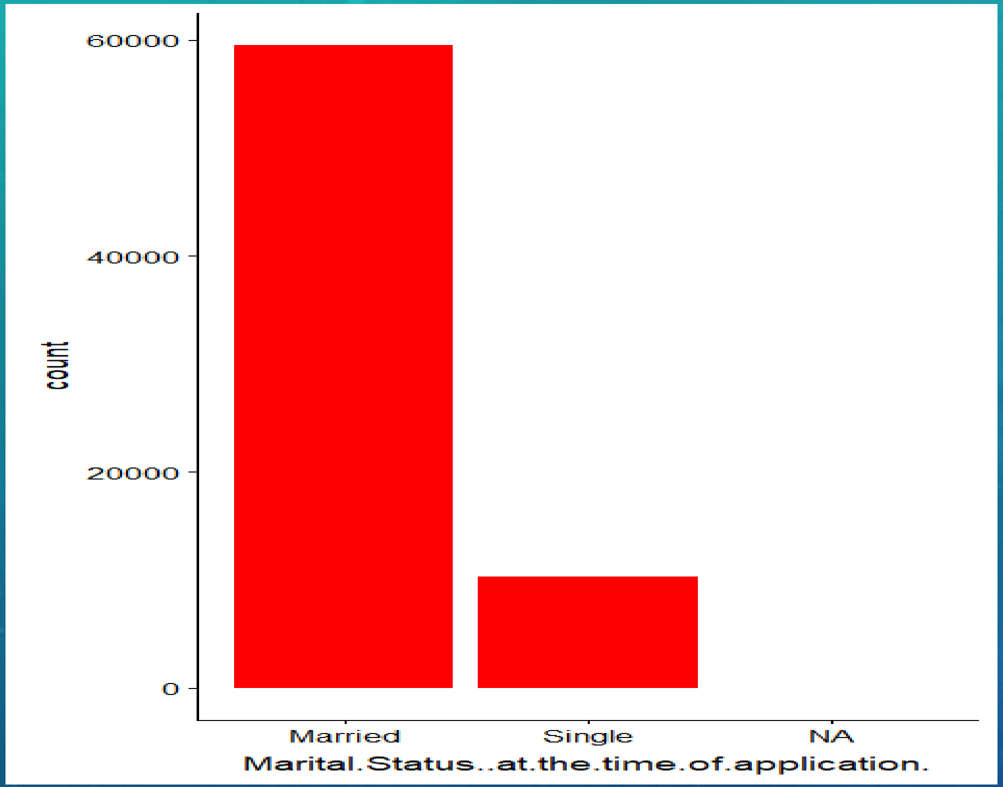


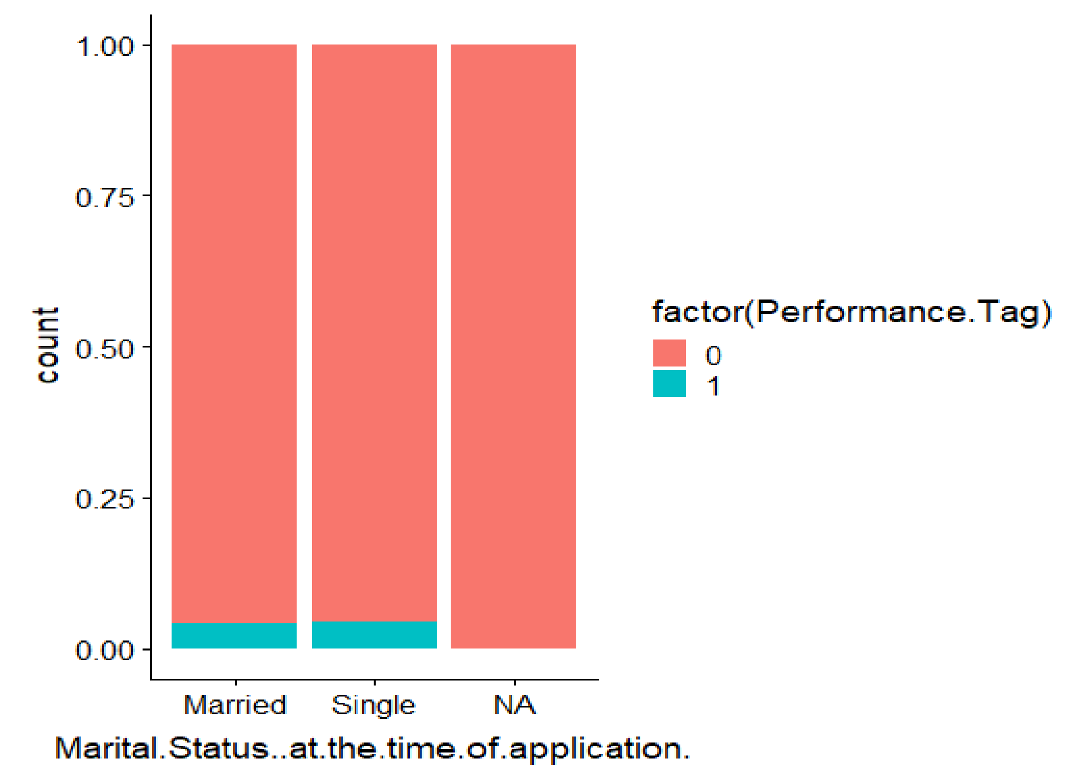


The median No.of.months. in.current.Company of non-defaulters is slightly lower than that of defaulters.

There are more male applicants than female applicants but there is no difference in default rates.

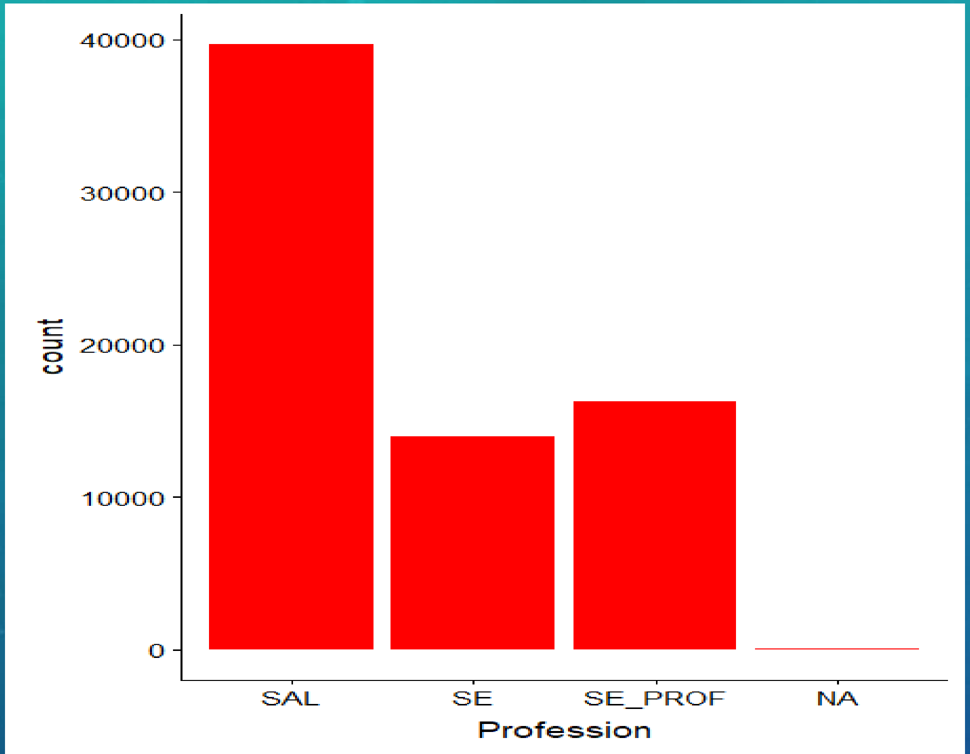
Insights derived by EDA

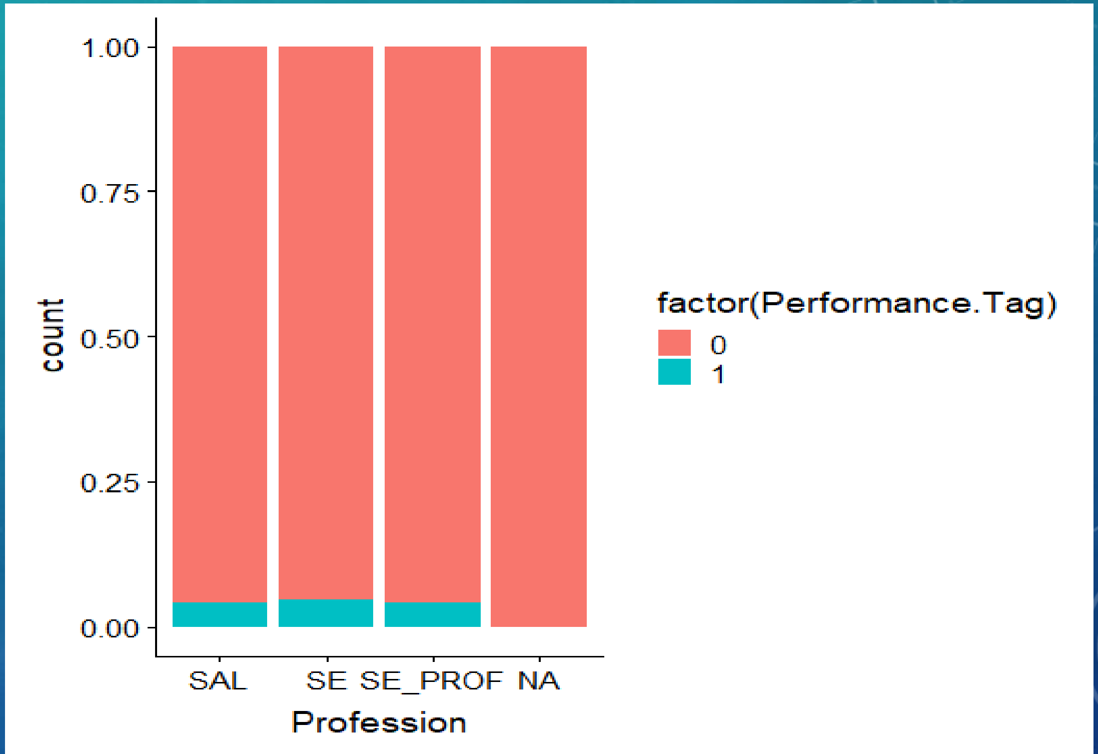




There are more married applicants than single applicants but there is no difference in default rates

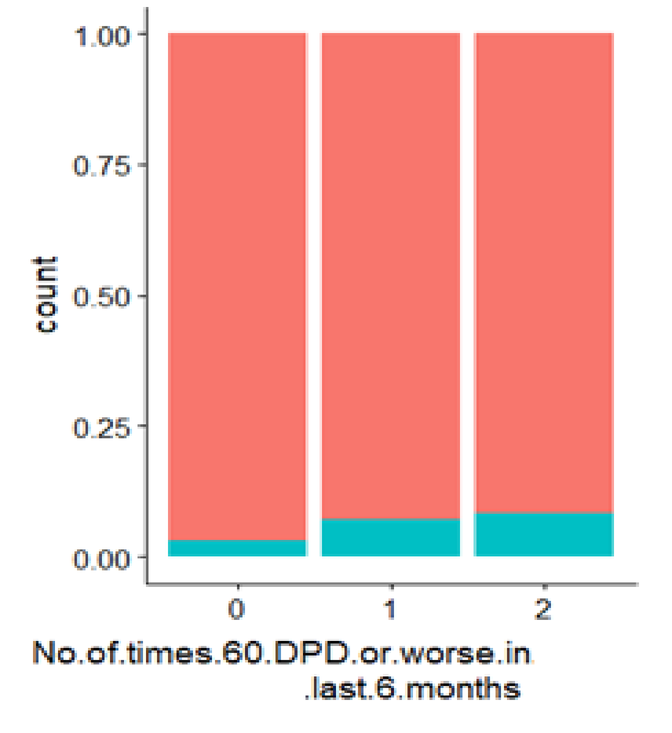
Insights derived by EDA

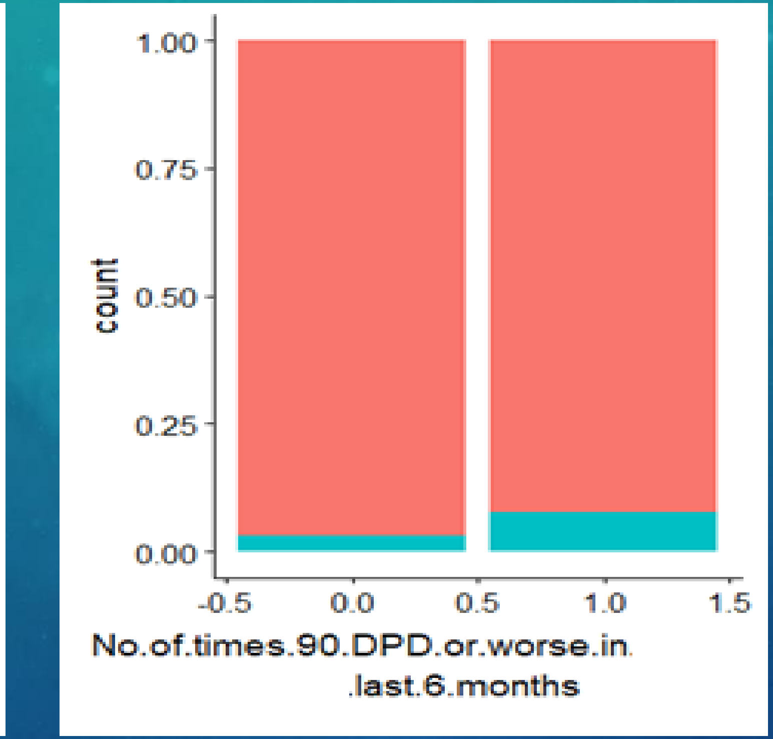


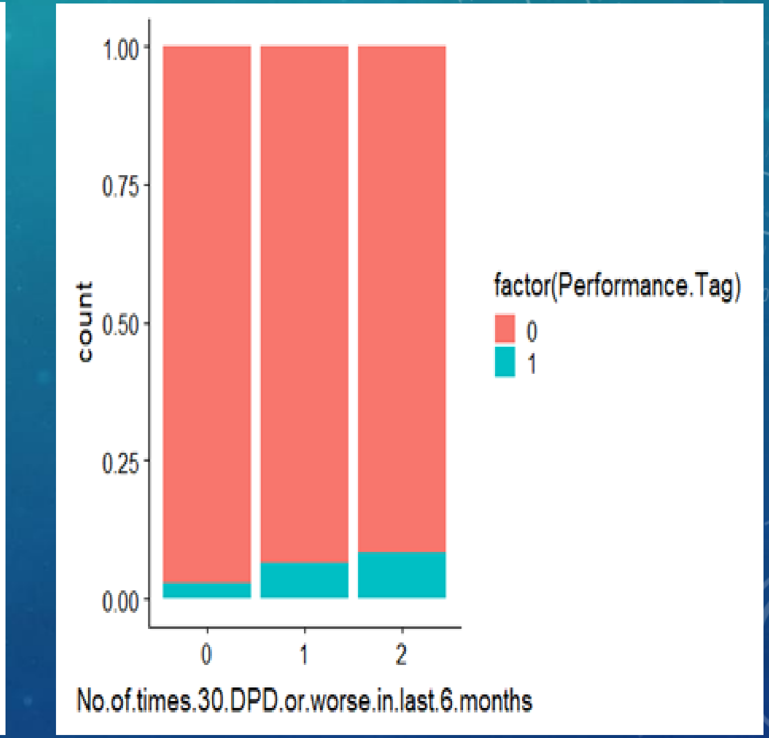


There are more applicants whose profession is SAL but there is no difference in default rates.

Insights derived by EDA

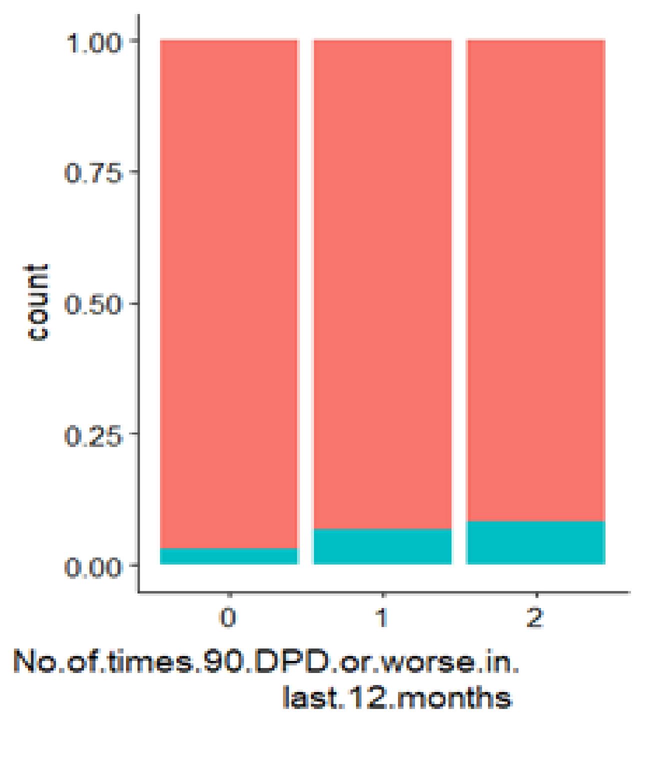


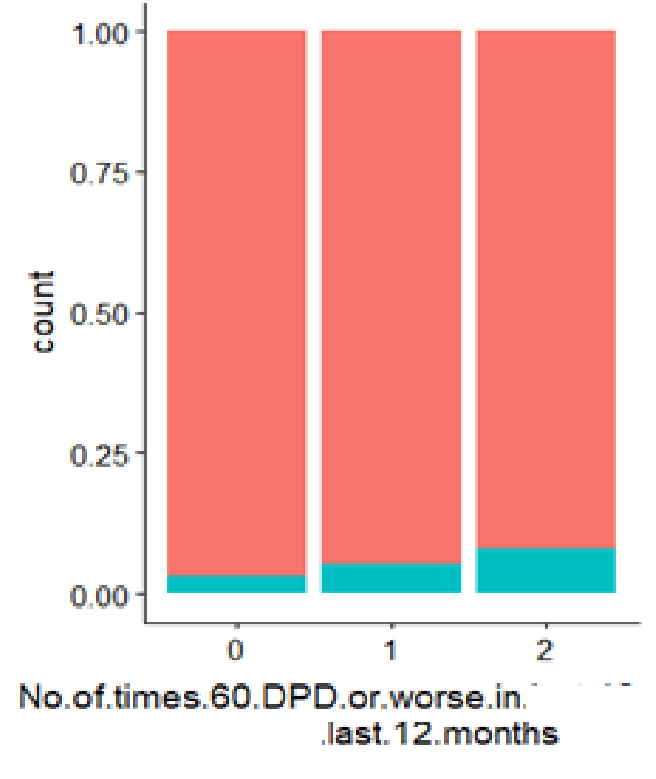


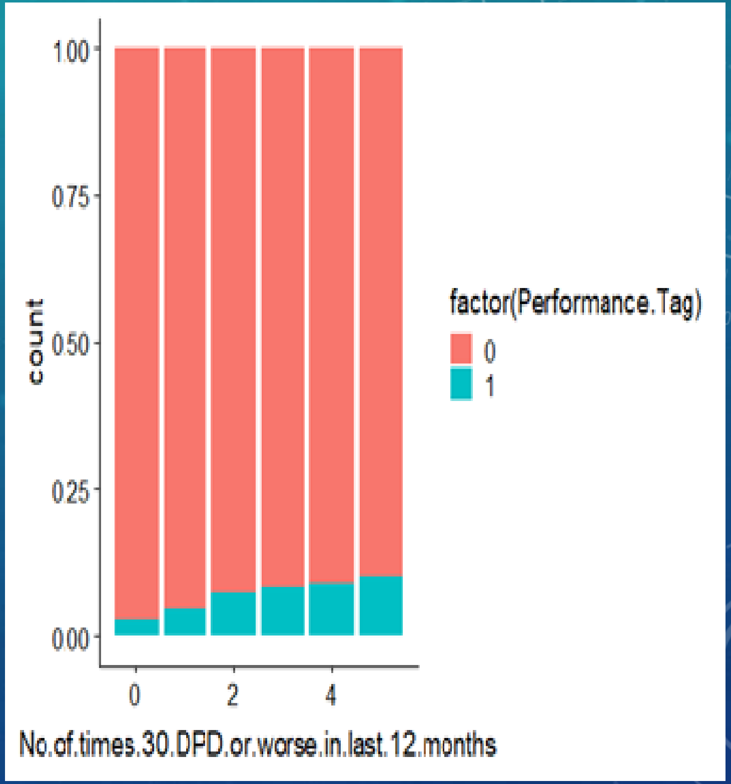


- Percentage of defaulters are increasing with increase in Number of 30/60/90 DPD or worse in last 6 months variable values. Hence these variables can be important predictors.

Insights derived by EDA

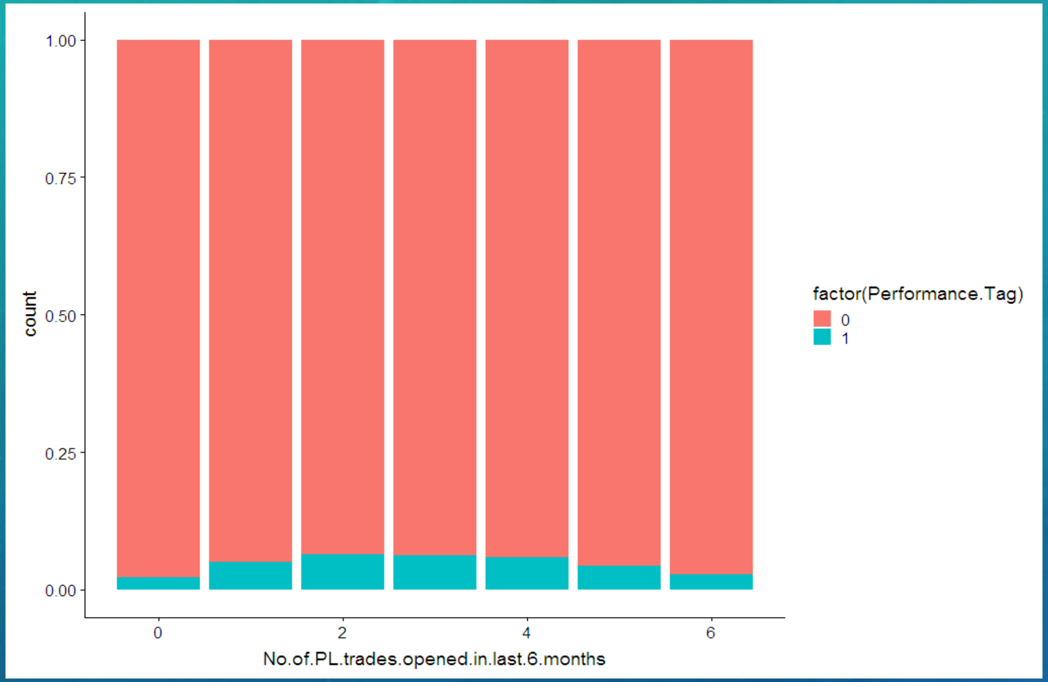


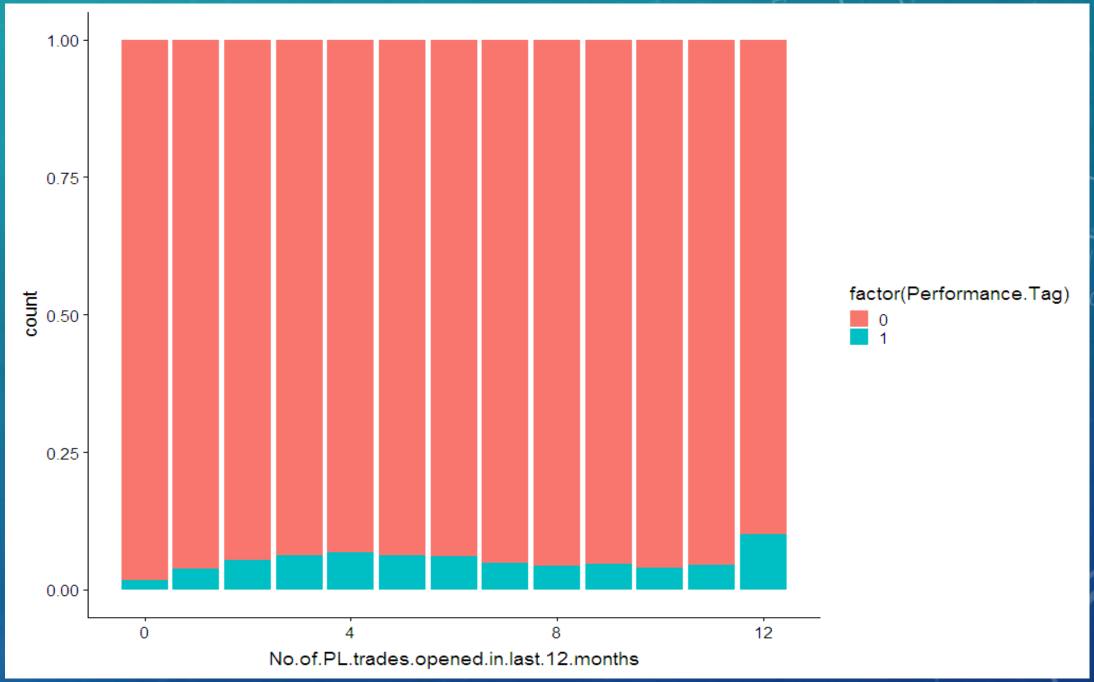




Percentage of defaulters is increasing with increase in Number of 30/60/90 DPD or worse in last 12 months variable values. Hence these variables can be important predictors.

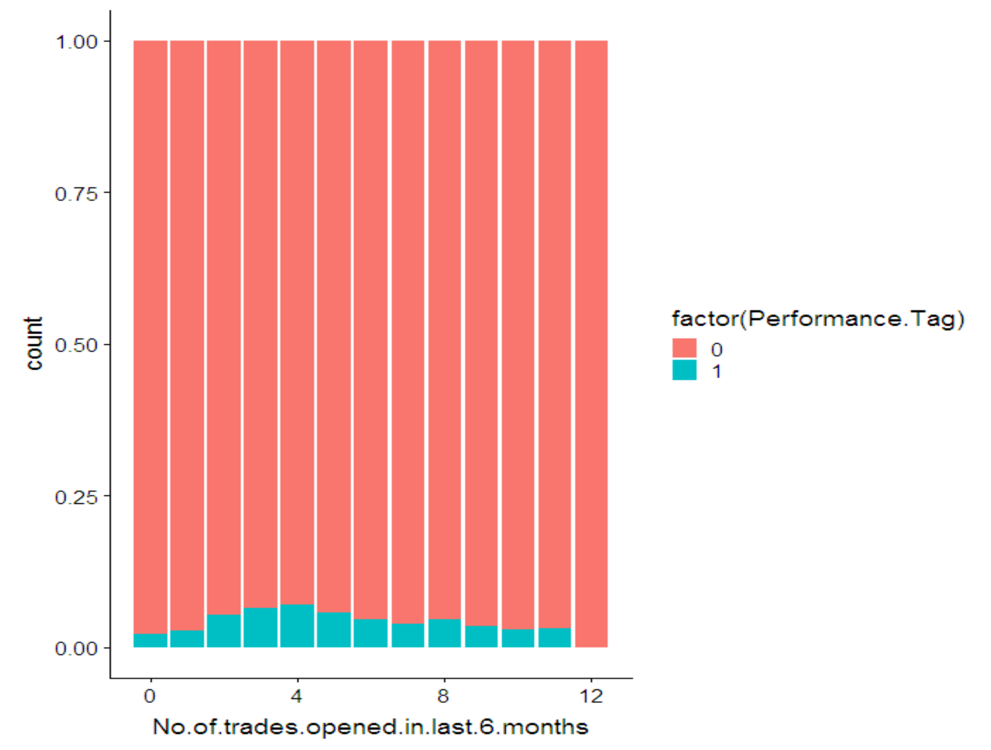
Insights derived by EDA

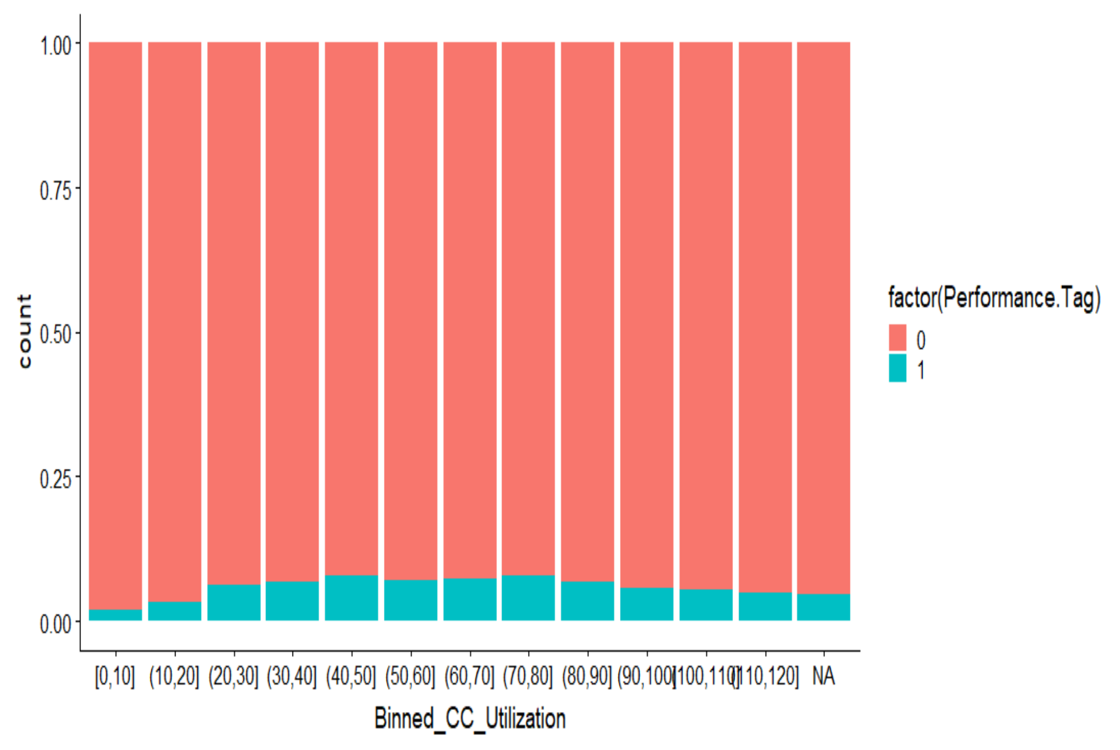




- Percentage of defaulters increases with increase in number of PL trades opened in last 6 months till the 4th month and then decreases
- Percentage of defaulters is highest amongst applicants who opened 12 PL Trades in last 12 months.

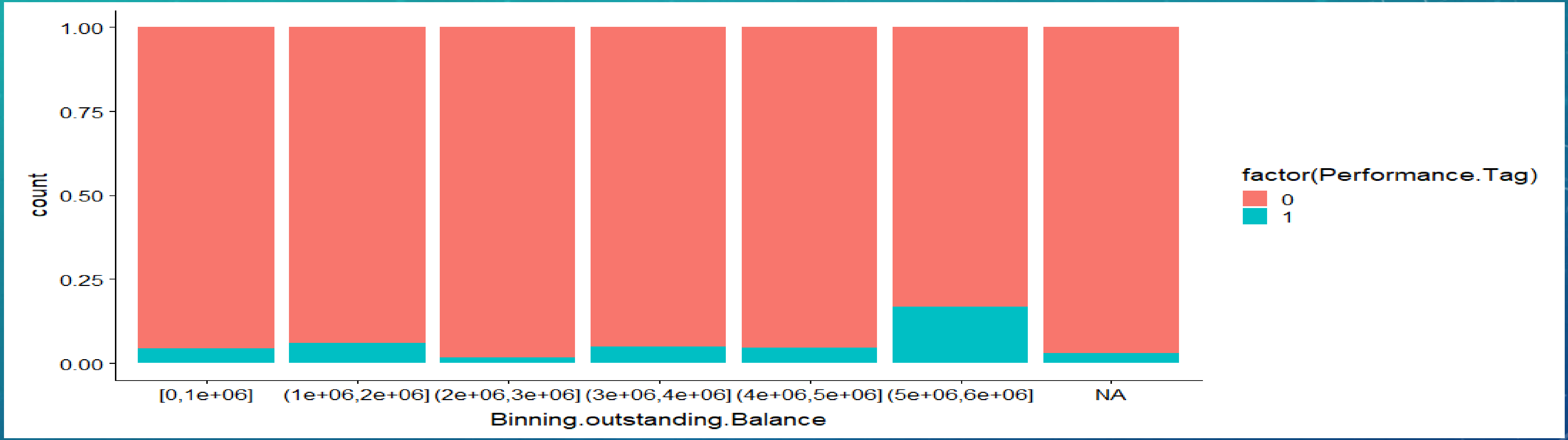
Insights derived by EDA





- Percentage of defaulters is lower among customers with Average Credit card utilization between 0 to 20.
- Applicants who opened trades four times in the last six months tend to default more.

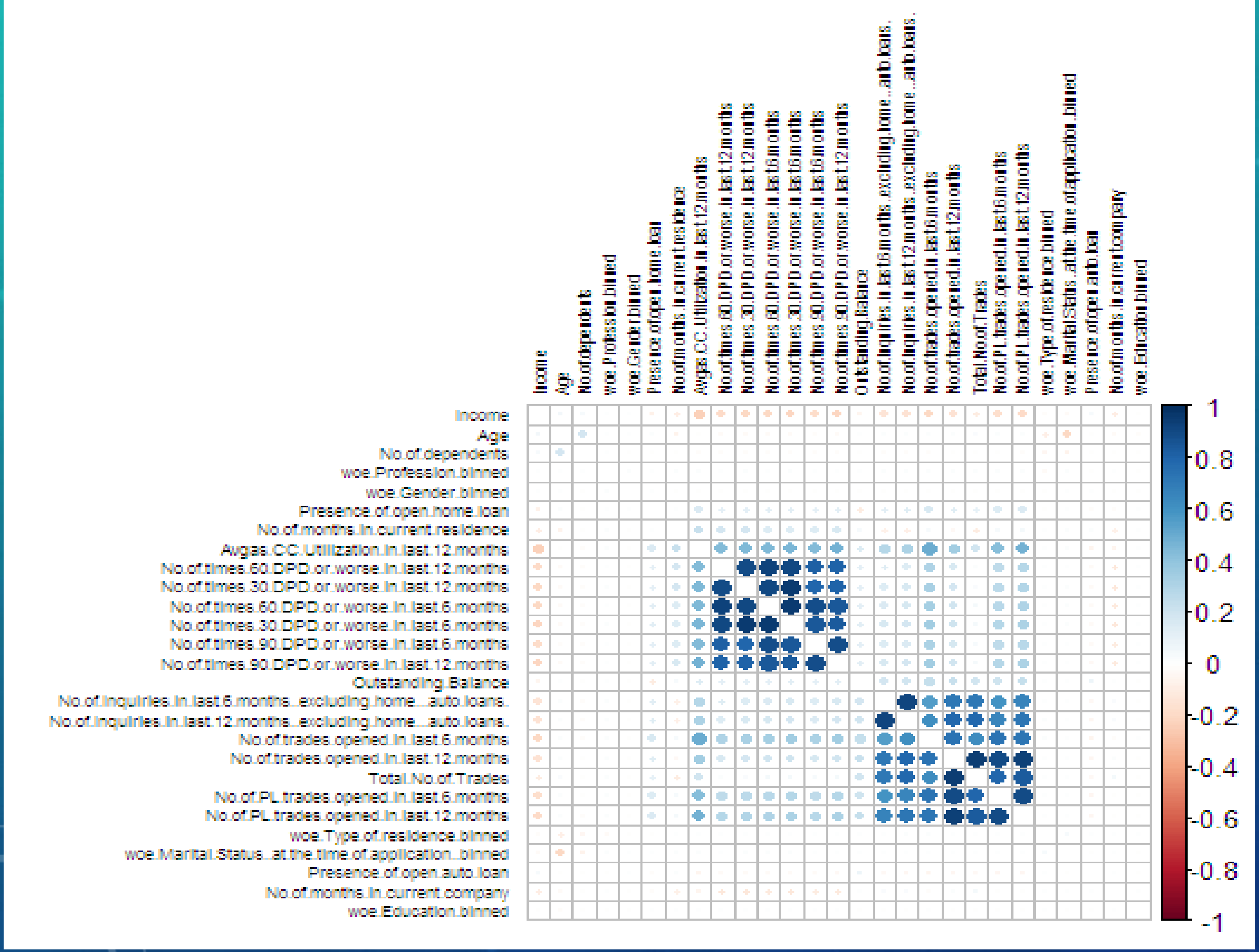
Insights derived by EDA



- **Plot for bivariate analysis of Outstanding balance binned in segments of Rs.10 Lac**
- Outstanding balance field shows higher percentage of defaulters in 50L-60L bin compared to lower outstanding balance bins. This can be an important predictor of default.

Conclusion : Variables of credit bureau dataset showed better insights than demographic variables.

Correlation plot



We see two groups of variables being correlated with variables within the group

Group1

- 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.30.DPD.or.worse.in.last.12.months
- No.of.times.60.DPD.or.worse.in.last.12.months
- Avgas.CC.Utilization.in.last.12.months

Group2

- No.of.trades.opened.in.last.6.months
- No.of.PL.trades.opened.in.last.6.months
- No.of.PL.trades.opened.in.last.12.months
- No.of.trades.opened.in.last.12.months
- Total.No.of Trades
- No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
- No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

MODEL BUILDING APPROACH

- **OUTLIER TREATMENT:** Outlier detection is done using boxplot on continuous variables and quantiles function and the variables with outliers has been corrected by capping the outliers to the nearest non-outlier values.

- **DATA SCALING:** Scaling is performed for all variables except Application ID and performance tag to standardize the data into common scale.

- **DATA SPLIT:** The final dataset is split into Train and Test in 70:30 ratio for model building.
  - All models are trained on training datasets and regularization was done by tuning of hyper parameters with cross validation on validation datasets.
  - All the models are tested on test datasets that were kept separate from training and validation datasets.

- **DATA SAMPLING:** The given data is highly imbalanced. We have sampled data using ROSE package for balancing the training data sets.

- The cutoff value for the probability of default was chosen such that model evaluation metrics like accuracy ,sensitivity and specificity were almost equal to each other.

- Logistic Regression was built by iteratively removing using these two algorithms

1. Stepwise variable selection based on AIC[using stepAIC()]
2. Backward variable selection based on VIF and p value

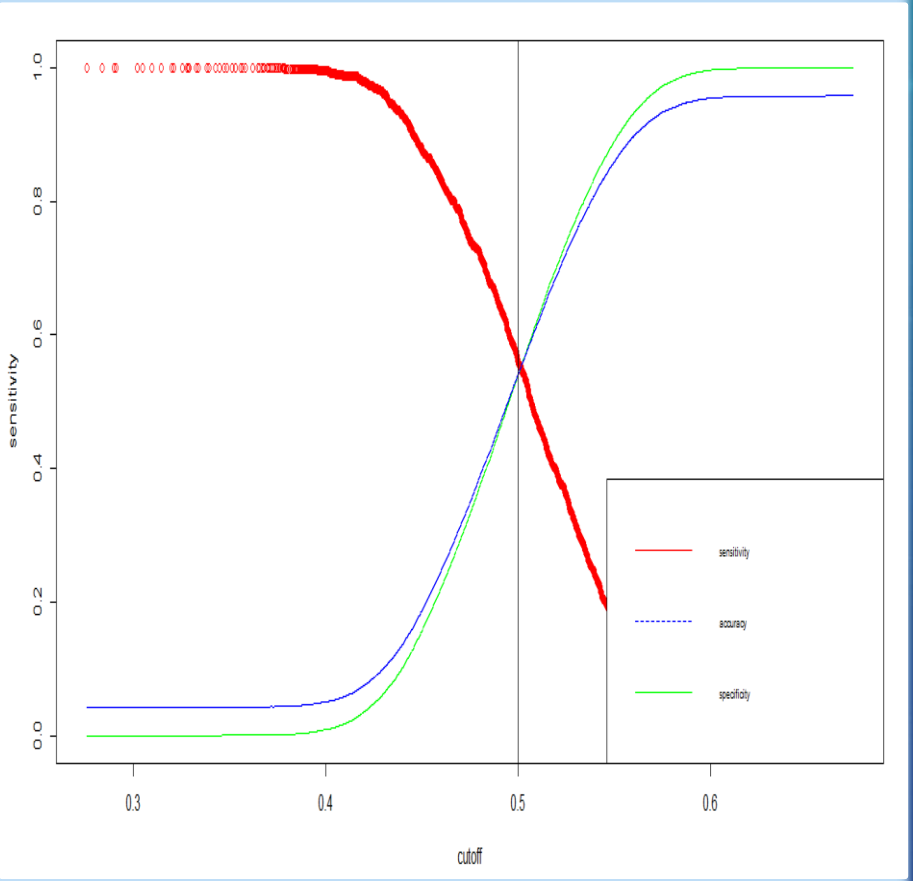
LOGISTIC REGRESSION MODEL ON DEMOGRAPHIC DATA

Important predictors : All variables have extremely low p values , Hence keeping all variables on that criteria.

- AGE
- INCOME
- NO.OF.MONTHS.IN.CURRENT.RESIDENCE
- NO.OF.MONTHS.IN.CURRENT.COMPANY
- WOE.PROFESSION.BINNED
- WOE.TYPE.OF.RESIDENCE.BINNED
- WOE.EDUCATION.BINNED

Thus a logistic regression model based only on demographic data seems to have low performance.

Statistics	Values
Cut-off	0.55
Accuracy	55%
Sensitivity	55%
Specificity	55%

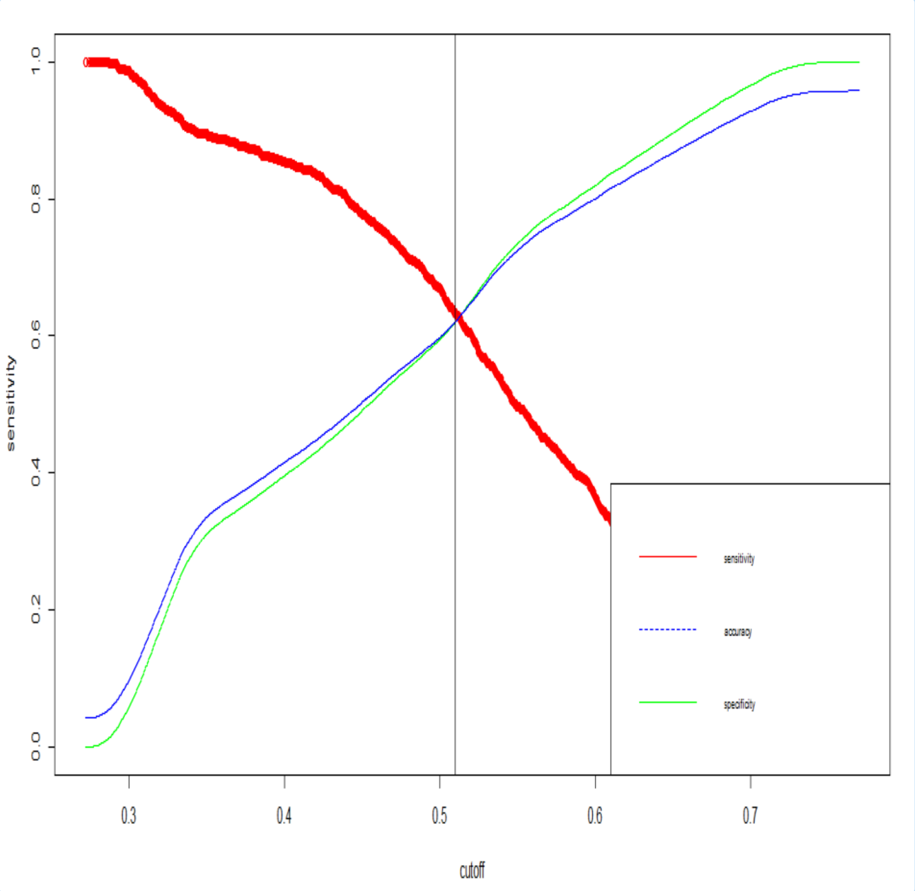


LOGISTIC REGRESSION MODEL ON MERGED CREDIT BUREAU AND DEMOGRAPHIC DATASET WITHOUT REJECTED 1425 RECORDS

Predictors in logistic regression model trained on a part of merged credit bureau and demographic dataset (merged on the application id column) without rejected 1425 records which does not have performance tags are as follows :

- INCOME
- NO.OF.MONTHS.IN.CURRENT.RESIDENCE
- NO.OF.MONTHS.IN.CURRENT.COMPANY
- WOE.EDUCATION.BINNED
- AVGAS.CC.UTILIZATION.IN.LAST.12.MONTHS
- NO.OF.TRADES.OPENED.IN.LAST.6.MONTHS
- NO.OF.PL.TRADES.OPENED.IN.LAST.6.MONTHS
- NO.OF.PL.TRADES.OPENED.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.30.DPD.OR.WORSE.IN.LAST.12.MONTHS
- NO.OF.INQUIRIES.IN.LAST.12.MONTHS..EXCLUDING.HOME...AUTO.LOANS.
- PRESENCE.OF.OPEN.HOME.LOAN
- OUTSTANDING.BALANCE
- TOTAL.NO.OF.TRADES

Statistics	Values
Cut-off	0.51
Accuracy	63%
Sensitivity	63%
Specificity	63%



ALL VARIABLES HAVE EXTREMELY LOW P VALUES AND VIF LESS THAN OR ALMOST EQUAL TO 2, HENCE KEEPING ALL VARIABLES ON THAT CRITERIA

LOGISTIC REGRESSION MODEL ON MERGED DATASET WITHOUT REJECTED 1425 RECORDS

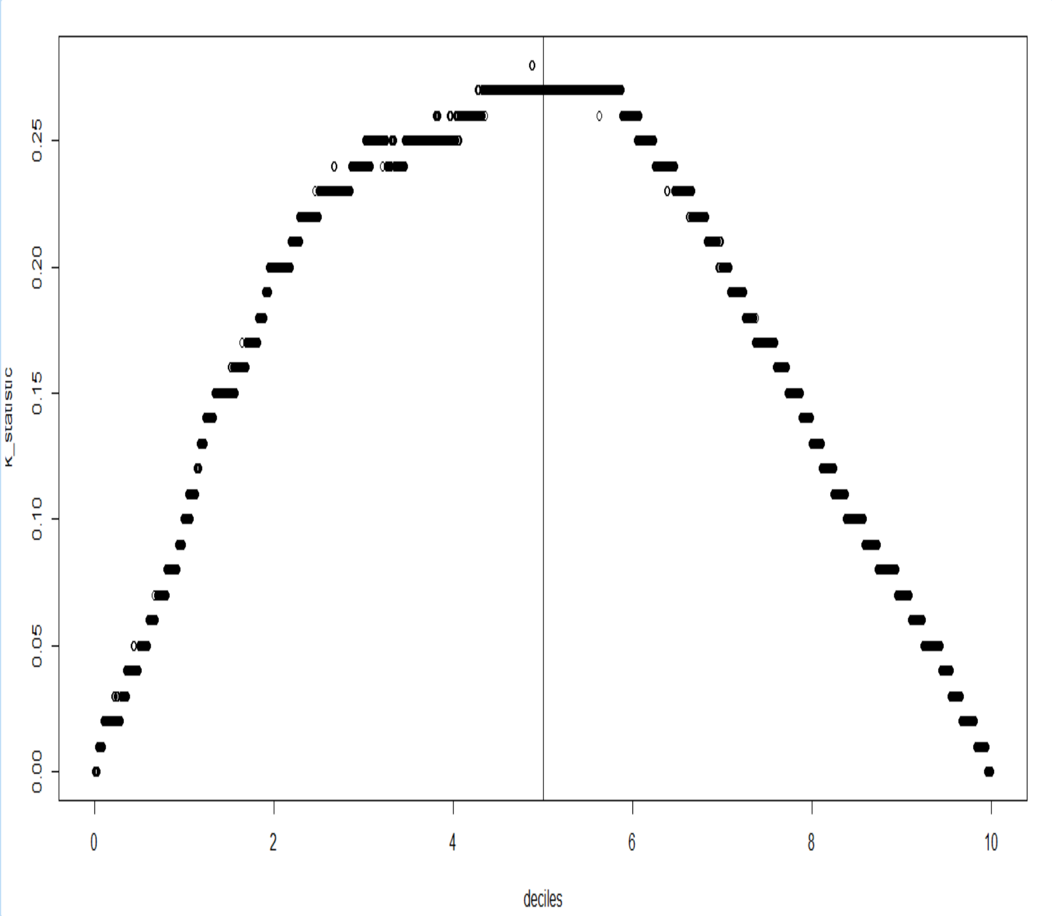
CONFUSION MATRIX AND KS CHART:

Prediction	0	1
0	12436	324
1	7620	560

Statistics	Values
Accuracy	0.6206
Sensitivity	0.62006
Specificity	0.63348

KS STATISTIC FOR THIS MODEL IS 0.27 AND LIES WITHIN IN FIRST 5 DECILES

KS CHART



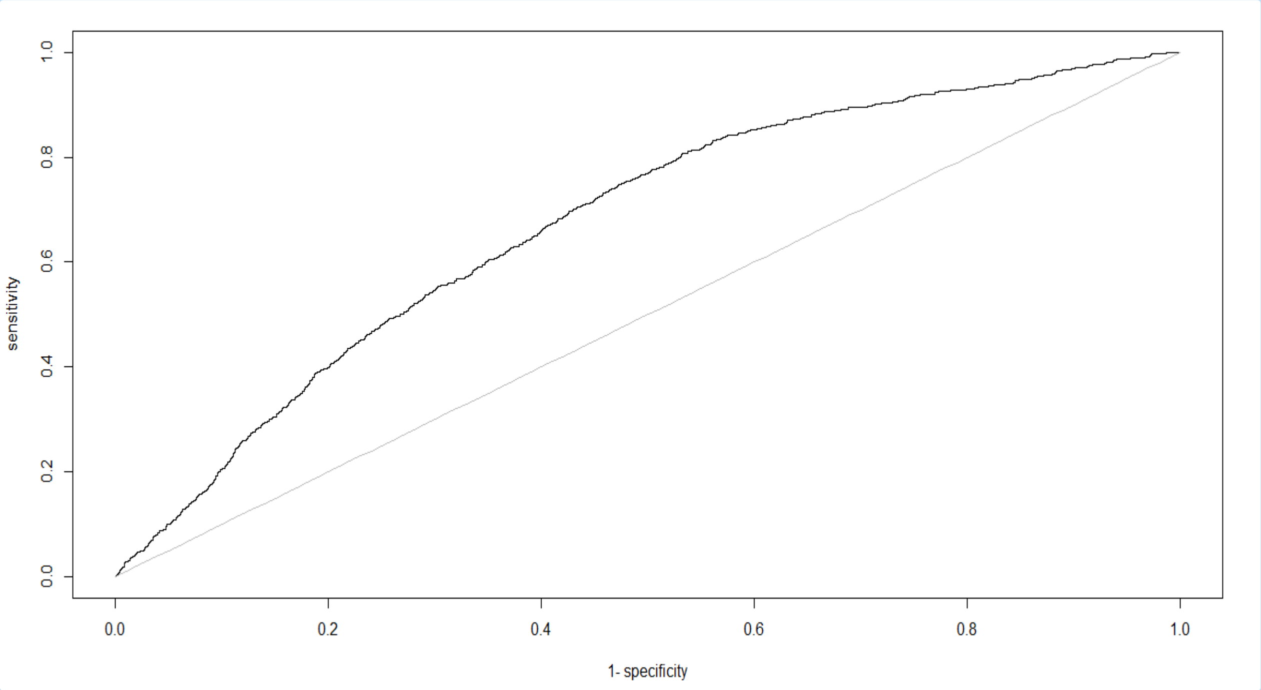
LOGISTIC REGRESSION MODEL ON MERGED DATASET WITHOUT REJECTED 1425 RECORDS

CROSS VALIDATION ON TEST DATA SETS:

On Test Data set 1 :
- sensitivity=62%
- specificity=60%
- accuracy=62%

On Test data set 2:
- sensitivity=62%
- specificity=63%
- accuracy=62%

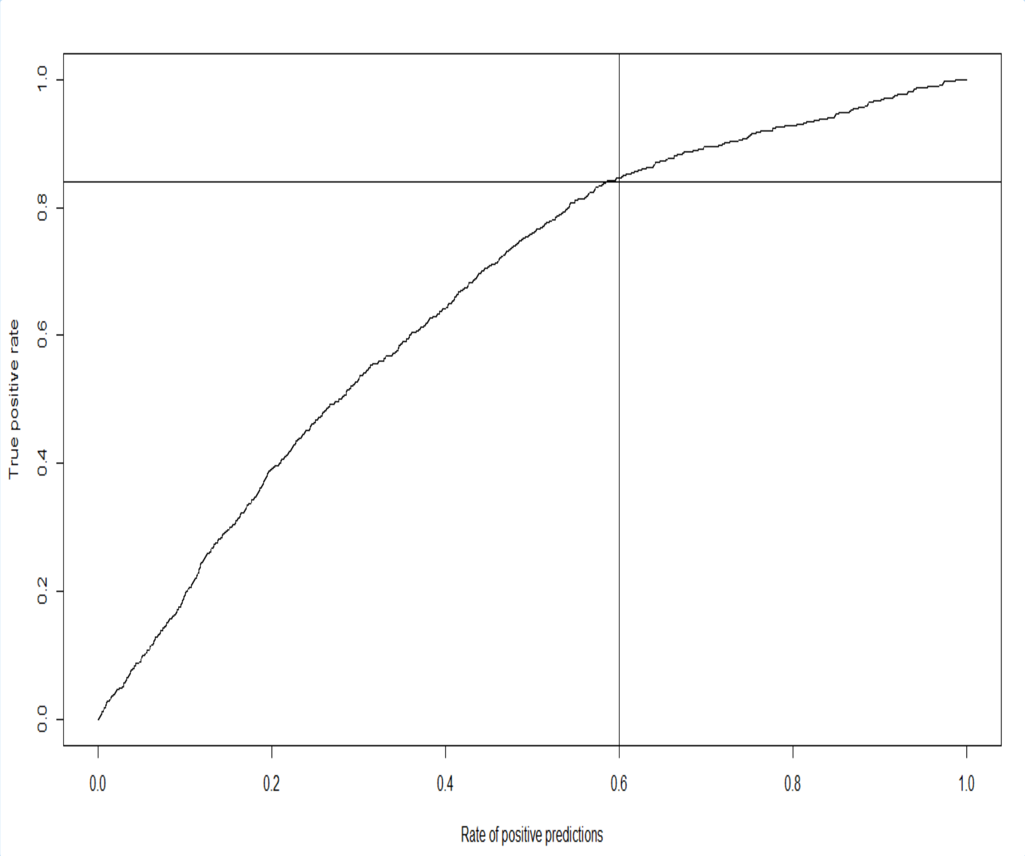
AREA UNDER THE CURVE :



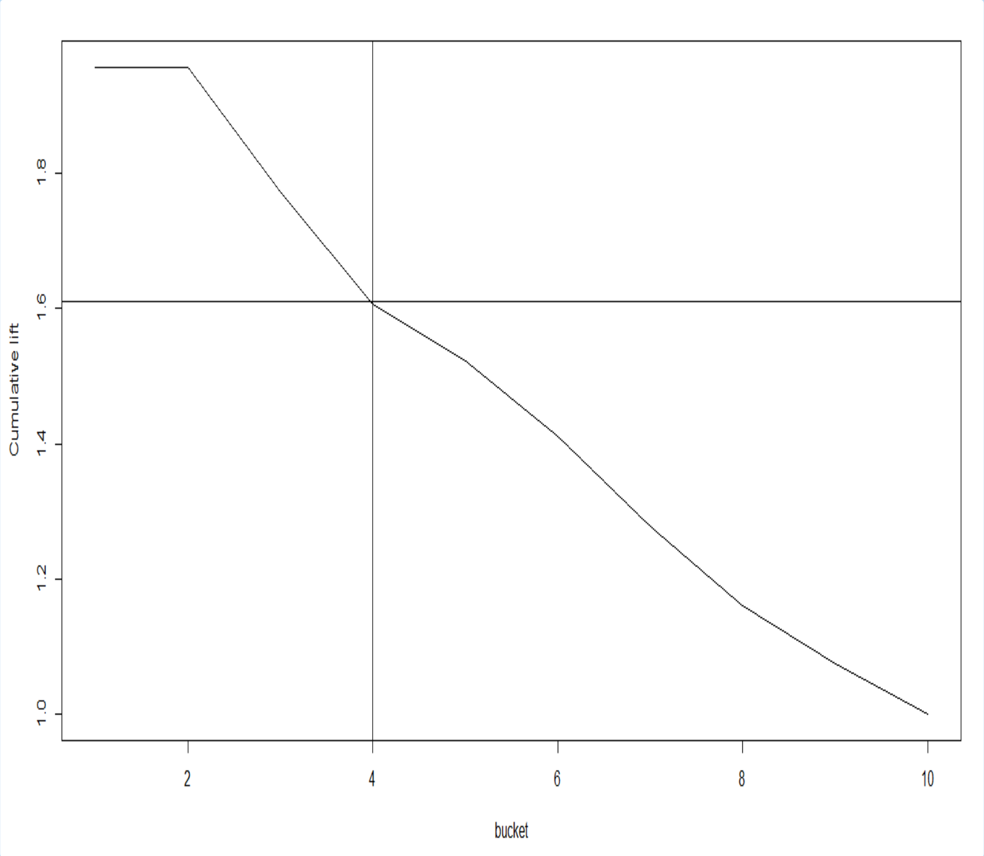
AREA UNDER ROC CURVE = 0.67

LOGISTIC REGRESSION MODEL ON MERGED DATASET WITHOUT REJECTED 1425 RECORDS

GAIN AND LIFT CHARTS :



GAIN CHART LIES WITHIN FIRST 6 DECILES AS PER THE MODEL WE ARE ABLE TO PREDICT MORE THAN 80% OF DEFAULTERS CORRECTLY



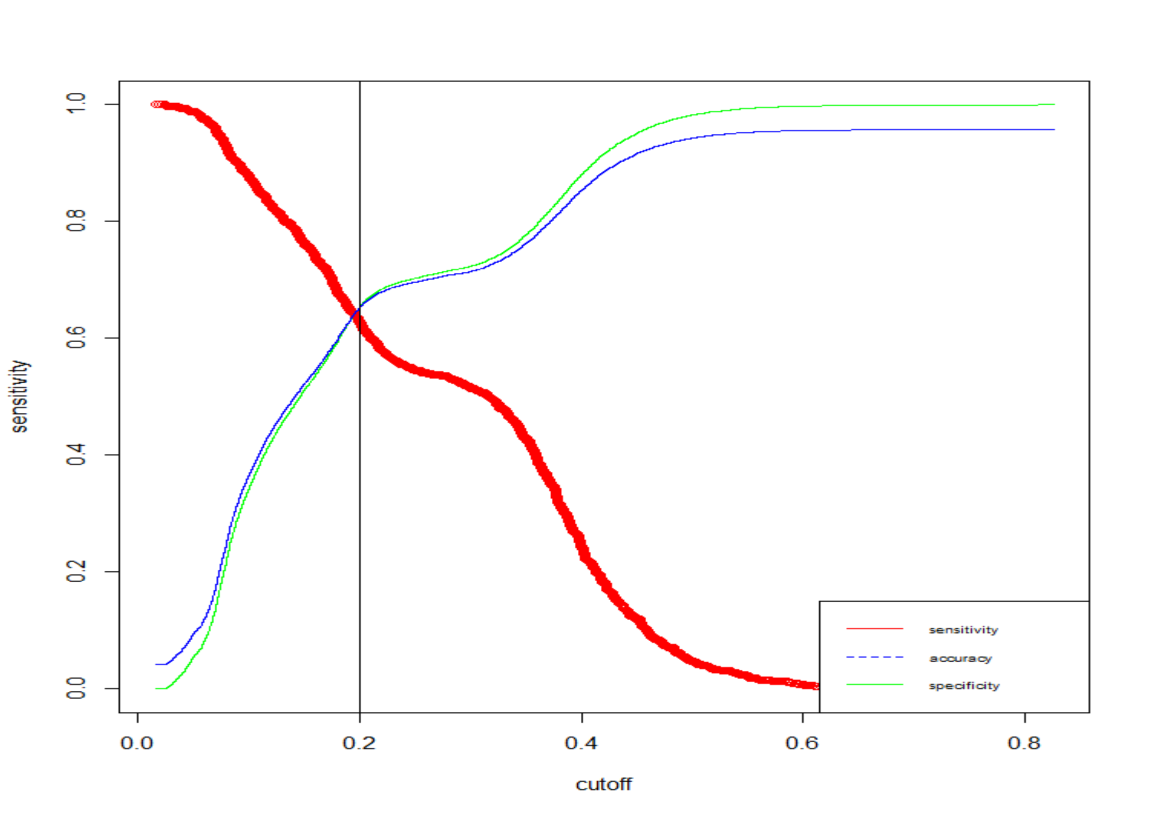
A LIFT OF 1.6 TIMES IS ACHIEVED WITH THE MODEL WITHIN FIRST 4 DECILES COMPARED TO RANDOM MODEL

XGboost model based on Merged Dataset

Top 5 important variables in terms of gain(fractional contribution of each feature to the model)

1: Avgas.CC.Utilization.in.last.12.months
2: No.of.times.30.DPD.or.worse.in.last.12.months
3: No.of.times.90.DPD.or.worse.in.last.6.months
4: No.of.times.60.DPD.or.worse.in.last.6.months
5: No.of.times.30.DPD.or.worse.in.last.6.months

Statistics	Values
Cut-off after 5 fold cross validation	0.195
Accuracy	64.13%
Sensitivity	64.13%
Specificity	64.14%

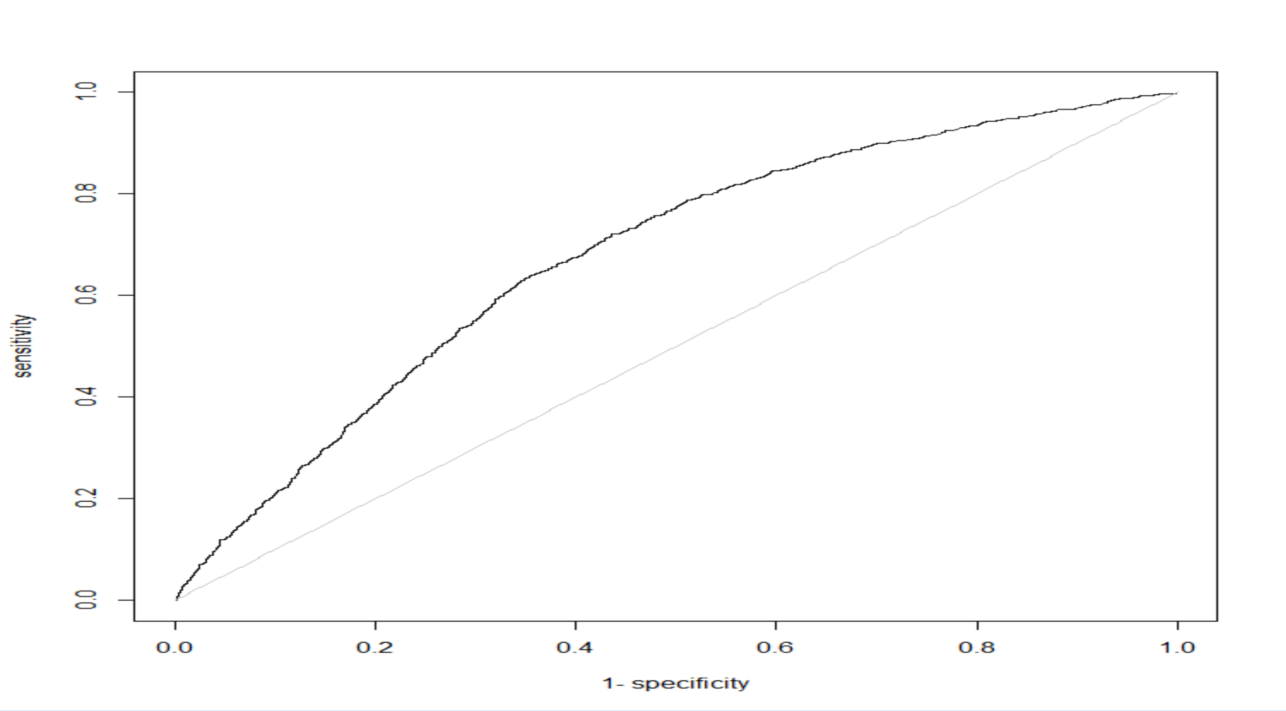


XGboost model based on Merged Dataset

Confusion matrix:

	Reference	
Prediction	**0**	**1**
0	12862	317
1	7194	567

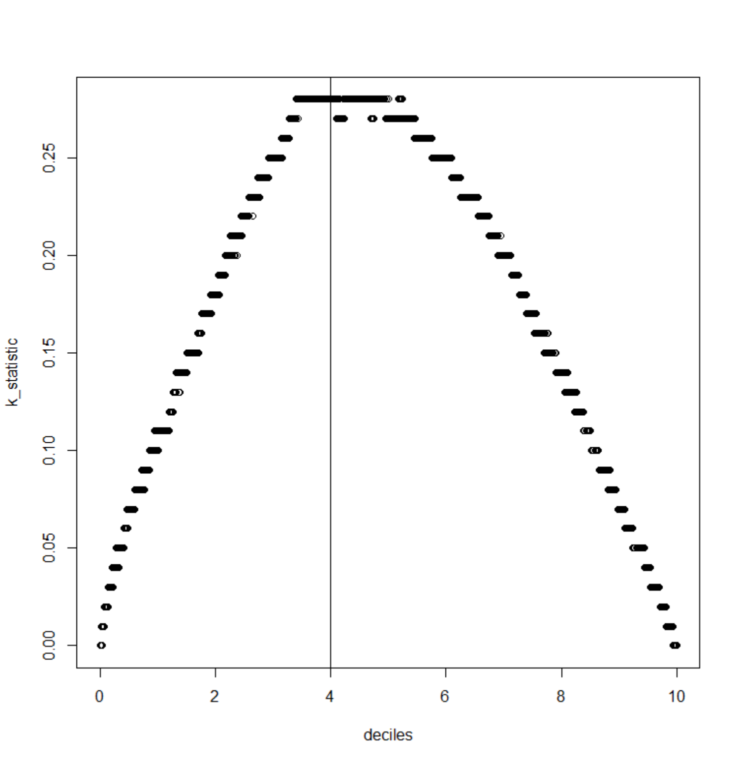
AREA UNDER THE CURVE



AUC under ROC curve for this model is 0.67

XGboost model based on Merged Dataset

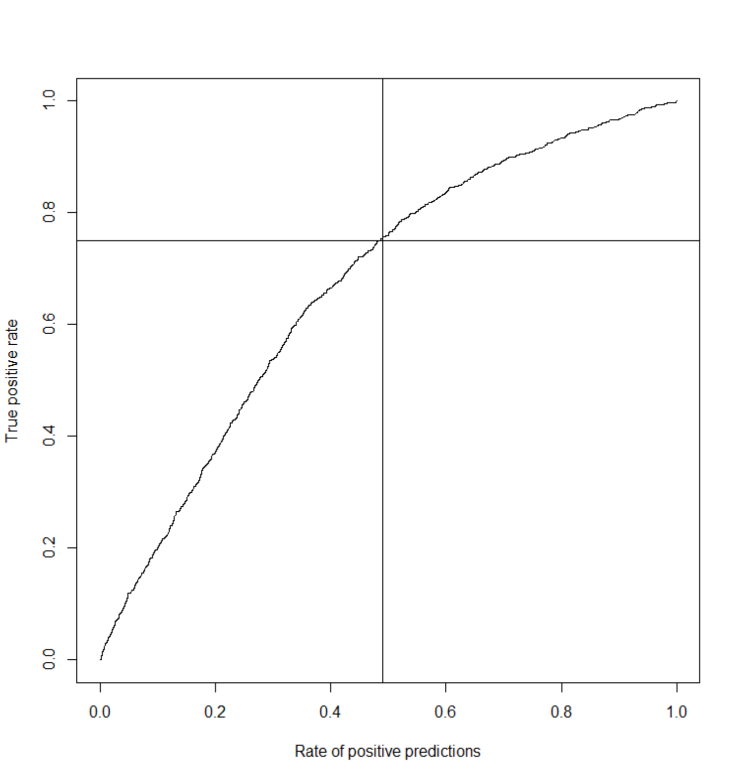
KS chart



KS statistic for this model is 0.28.

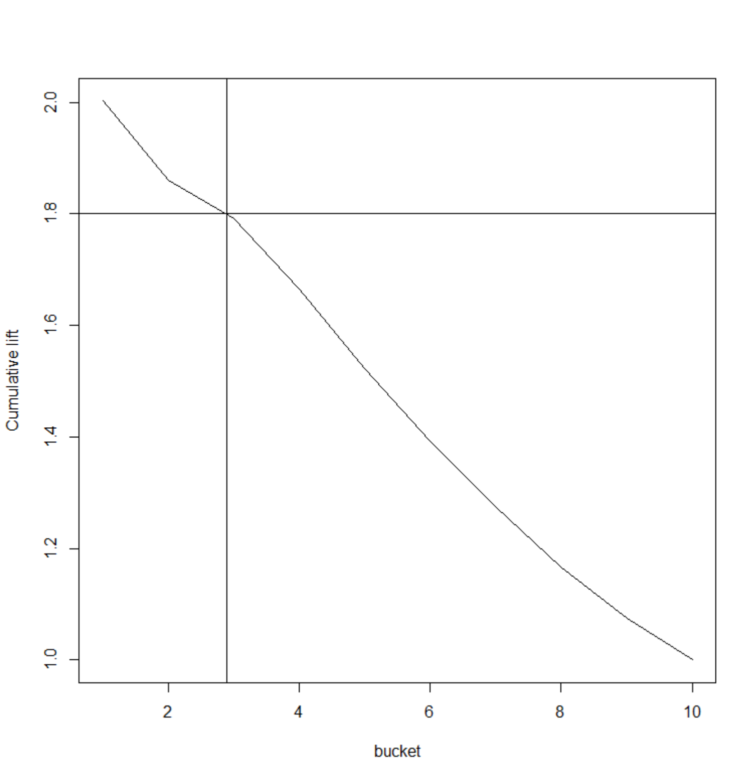
KS statistic lies within first 4 deciles.

Gain Chart



Gain lies within first 5 deciles as per the model we are able to capture 75% of defaulters correctly. Here positive predictions are predictions of default

Lift Chart



Lift little below than 1.8 times is achieved with the model compared to a random model within first 3 deciles

DIFFERENT MODEL’S ACCURACY, SENSITIVITY & SPECIFICITY

Models	Accuracy	Sensitivity	Specificity
Logistic Regression on Demographic dataset	55%	55%	55%
Random Forest on Demographic dataset	54.49 %	54.56%	52.94%
Logistic Regression on Merged dataset (without rejected 1425 records)	63%	63%	63%
SVM model with linear Kernel on Merged dataset (without rejected 1425 records)	53.44%	52.59%	72.51%
SVM model with RBF Kernel on Merged dataset (without rejected 1425 records)	67.42%	67.86%	56.35%
Random Forest on Merged dataset (without rejected 1425 records)	63.4%	63.4%	63.6%
XG Boost on Merged dataset (without rejected 1425 records)	64.13%	64.13%	64.14%

Conclusions made from models made on dataset from which records with missing performance tags were removed

- All the performance evaluation metrics for all models made on this dataset were quite low with maximum accuracy being nearly 64%.

- Thus we used the logistic regression model made on the merged (credit bureau and demographic dataset) to predict the performance tag values for 1425 applicants with missing performance tags .

- 9 out of 1425 applicants were predicted to be non defaulters, while the rest 1416 applicants were predicted to be defaulters.

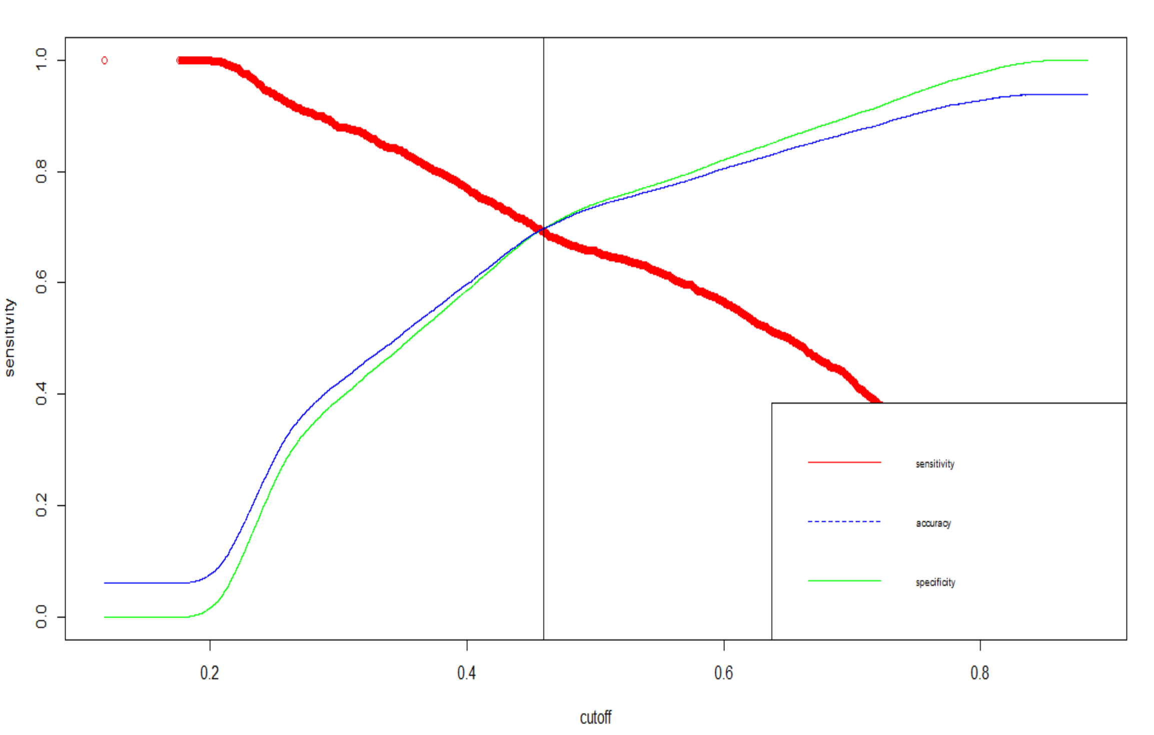
- These predictions for 1425 missing performance tags was included in the dataset and then the entire dataset was used further to make models to see if these models give us better performance.

LOGISTIC REGRESSION MODEL ON MERGED DATASET INCLUDING REJECTED 1425 RECORDS (PREDICTED)

Predictors in logistic regression model on merged dataset including rejected 1425 records with predictions for missing performance tag.

- Age
- Income
- No.Of.Months.In.Current.Residence
- No.Of.Months.In.Current.Company
- Woe.Profession.Binned
- Woe.Education.Binned
- 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
- Avgas.Cc.Utilization.In.Last.12.Months
- No.Of.Trades.Opened.In.Last.6.Months
- No.Of.Pl.Trades.Opened.In.Last.6.Months
- No.Of.Pl.Trades.Opened.In.Last.12.Months

Statistics	Values
Cut-off	0.46
Accuracy	70%
Sensitivity	70%
Specificity	70%



LOGISTIC REGRESSION MODEL ON MERGED DATASET WITH REJECTED 1425 RECORDS (PREDICTED)

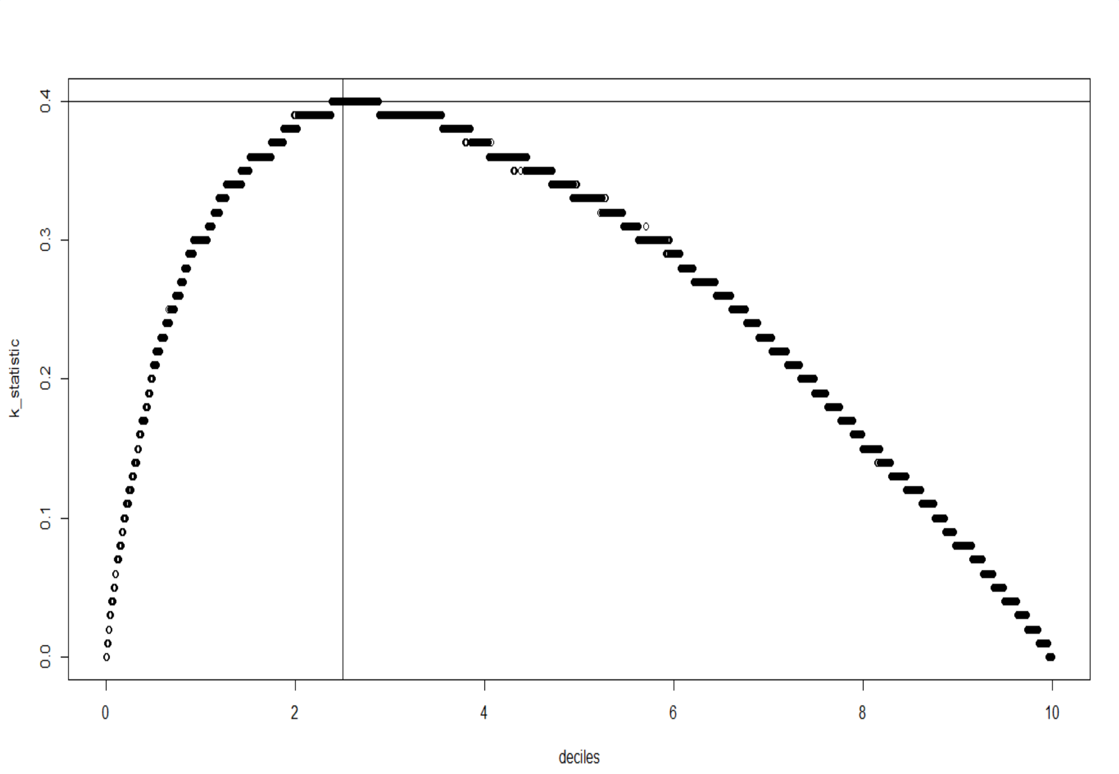
CONFUSION MATRIX

Prediction	0	1
0	14001	404
1	6058	905

Statistics	Values
Accuracy	0.6976
Sensitivity	0.6980
Specificity	0.6914

KS STATISTIC FOR THIS MODEL IS 0.40 AND LIES WITHIN IN FIRST 3 DECILES.

KS Chart



LOGISTIC REGRESSION MODEL ON MERGED DATASET WITH REJECTED 1425 RECORDS

CROSS VALIDATION ON OTHER TEST DATA SETS :

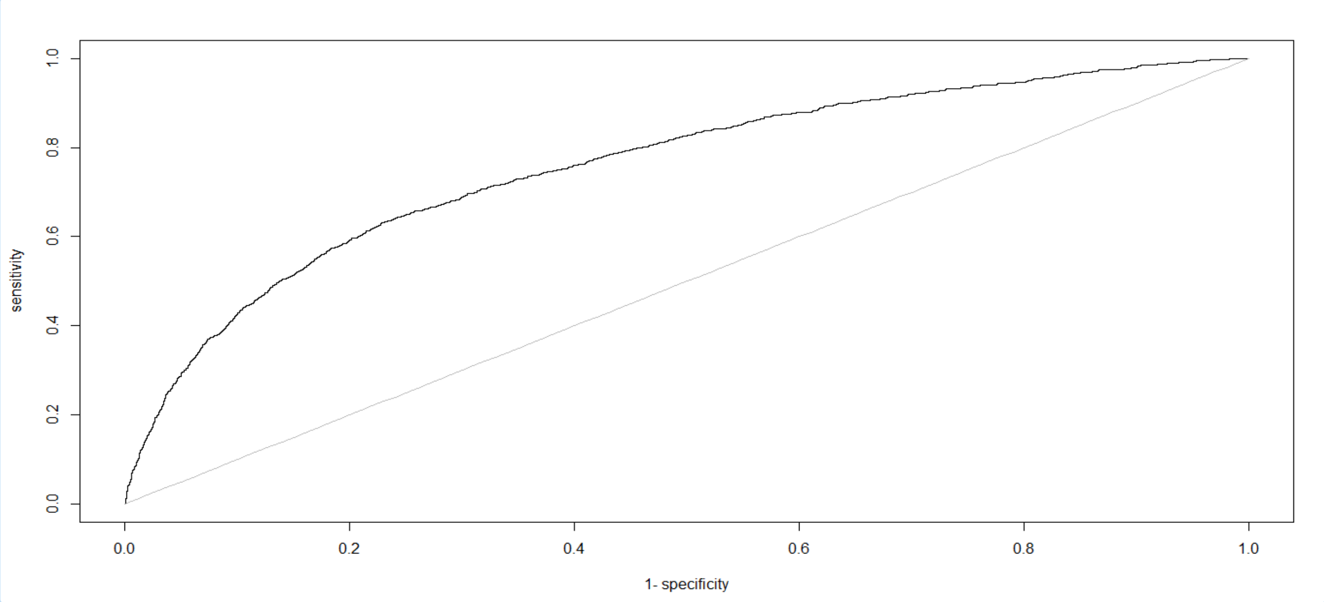
On Test Data set 1 :

- sensitivity=69%
- specificity=69%
- accuracy=69%

On Test data set 2:

- sensitivity=69%
- specificity=68%
- accuracy=69%

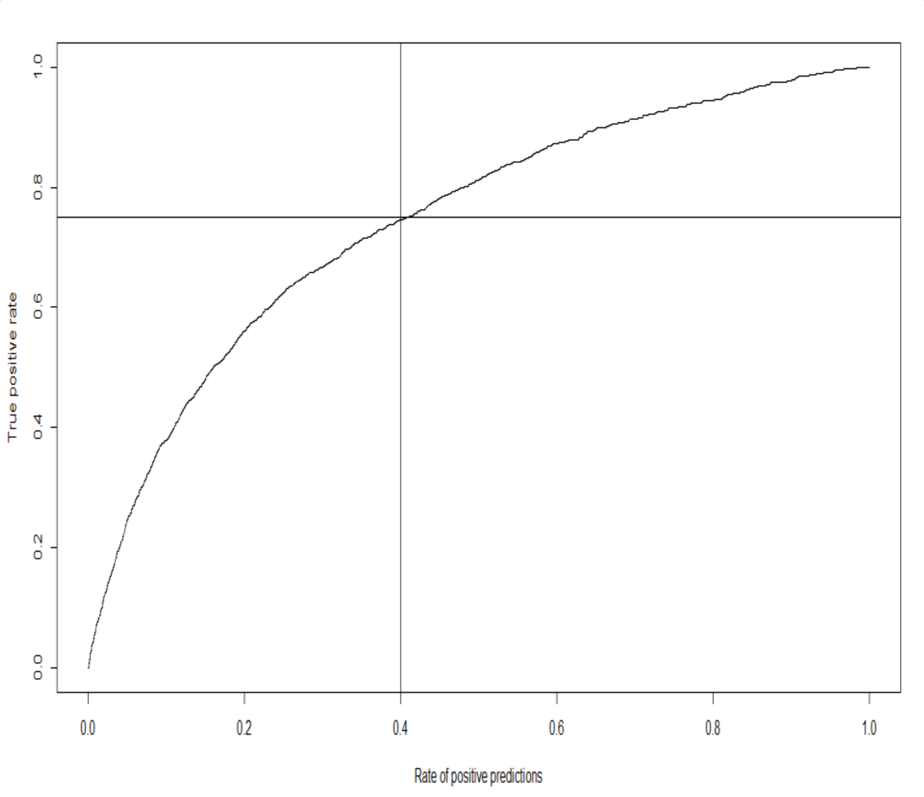
AREA UNDER THE CURVE :



AREA UNDER ROC CURVE = 0.759

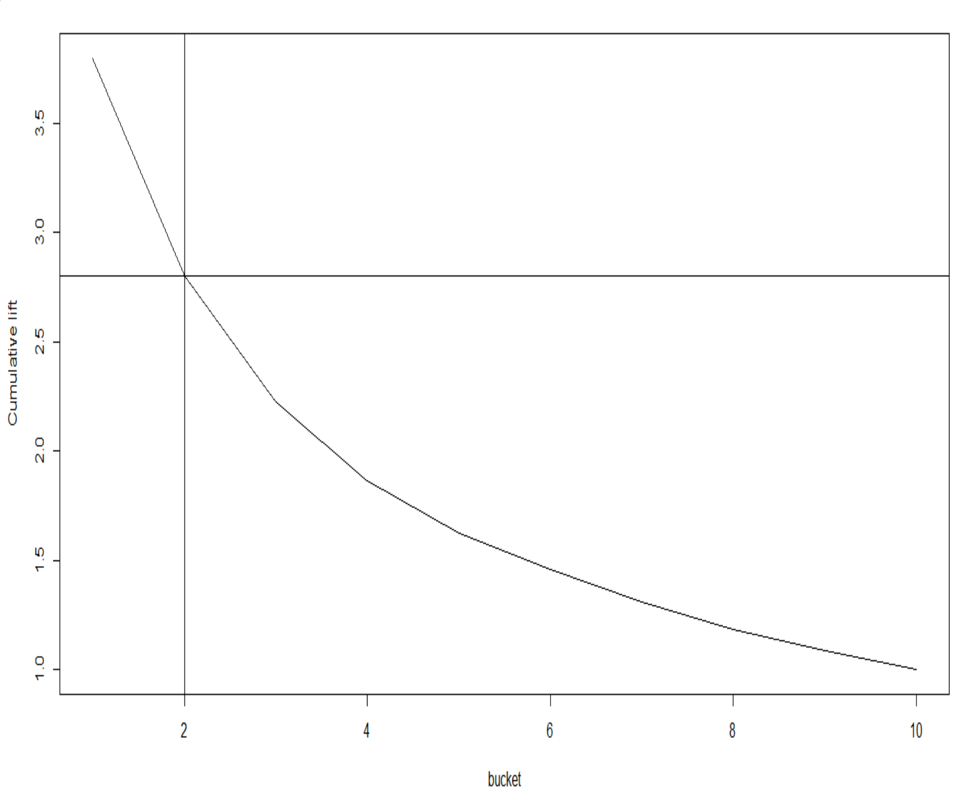
LOGISTIC REGRESSION MODEL ON MERGED DATASET WITH REJECTED 1425 RECORDS

Gain Chart



Within first 4 deciles as per the model we are able to predict 75% of defaulters correctly.

Lift Chart



A lift of 2.8 times is achieved with the model within first 2 deciles compared to random mode;

DIFFERENT MODEL’S ACCURACY, SENSITIVITY & SPECIFICITY ON MERGED DATASET WITH REJECTED 1425 RECORDS

DIFFERENT MODEL’S ACCURACY, SENSITIVITY & SPECIFICITY ON MERGED DATASET WITH REJECTED 1425 RECORDS WHICH HAVE PERFORMANCE TAGS

Models	Accuracy	Sensitivity	Specificity
Logistic Regression on the Final Merged Dataset (with rejected 1425 records)	69.76%	69.80%	69.14%
Random Forest on the Final Merged Dataset (with rejected 1425 records)	69.62%	69.63%	69.44%

CONCLUSION: FOR MERGED DATA WITH PERFORMANCE TAG MISSING RECORDS, LOGISTIC REGRESSION MODEL IS PERFORMING BETTER COMPARED TO RANDOM FOREST. HENCE CONSIDERING LOGISTIC REGRESSION MODEL AS FINAL MODEL FOR APPLICATION SCORECARD.

Application Scorecard

Final application scorecard was made using the **Logistic regression** model on the entire dataset which also contained predictions for missing values in "Performance Tag" in 1425 records.

The logistic regression model was chosen since its evaluation metrics were comparable to other models as well it's an easily interpretable simple model.

The scorecard was made using the following steps:

1. Application score card was made with odds of 10 to 1 being a score of 400. Score increases by 20 points for doubling odds.
2. Probability of default for all applicants were calculated
3. Odds for good was calculated. Since the probability computed is for rejection (bad customers), Odd(good) = (1-P(bad))/P(bad)
4. ln(odd(good)) was calculated
5. Used the following formula for computing application score card:
   400 + slope * (ln(odd(good)) - ln(10)) where slope is 20/(ln(20)-ln(10))
   Where, slope=20/(log(20)-log(10))

Summary of application_score_card values:

- Scores range from 272.7 to 393.4 for applicants with median score being 349.5.
- Higher scores indicate less risk for defaulting

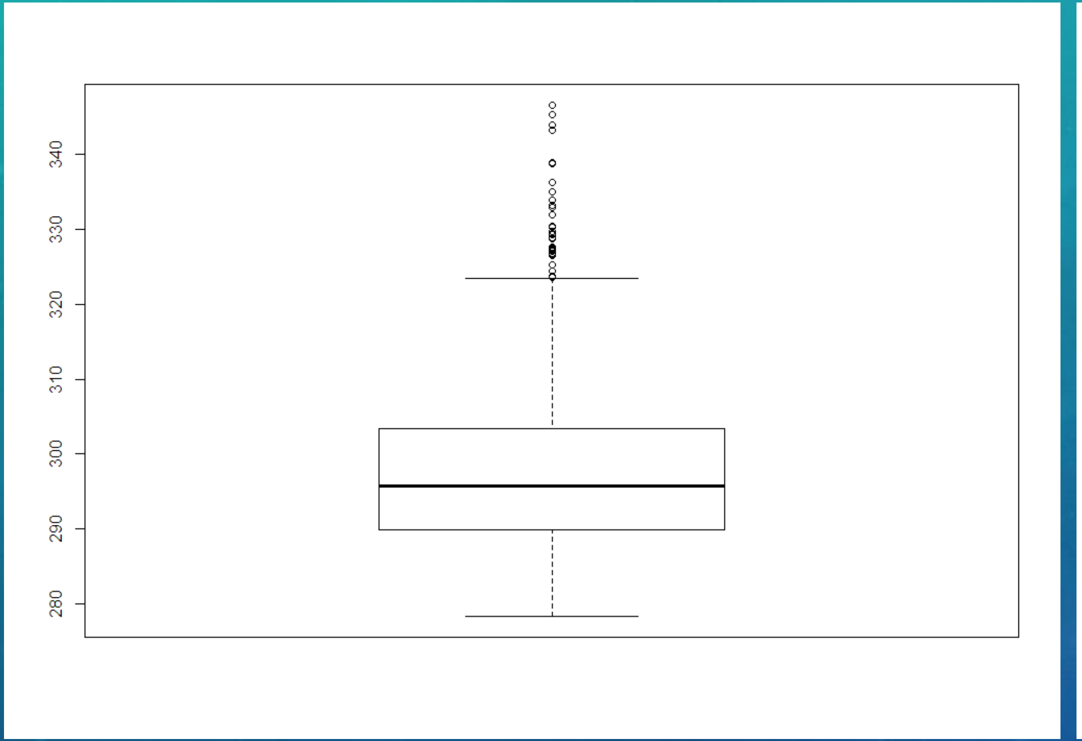
Application Scorecard

CUTOFF SCORE FOR ACCEPTING OR REJECTING AN APPLICATION

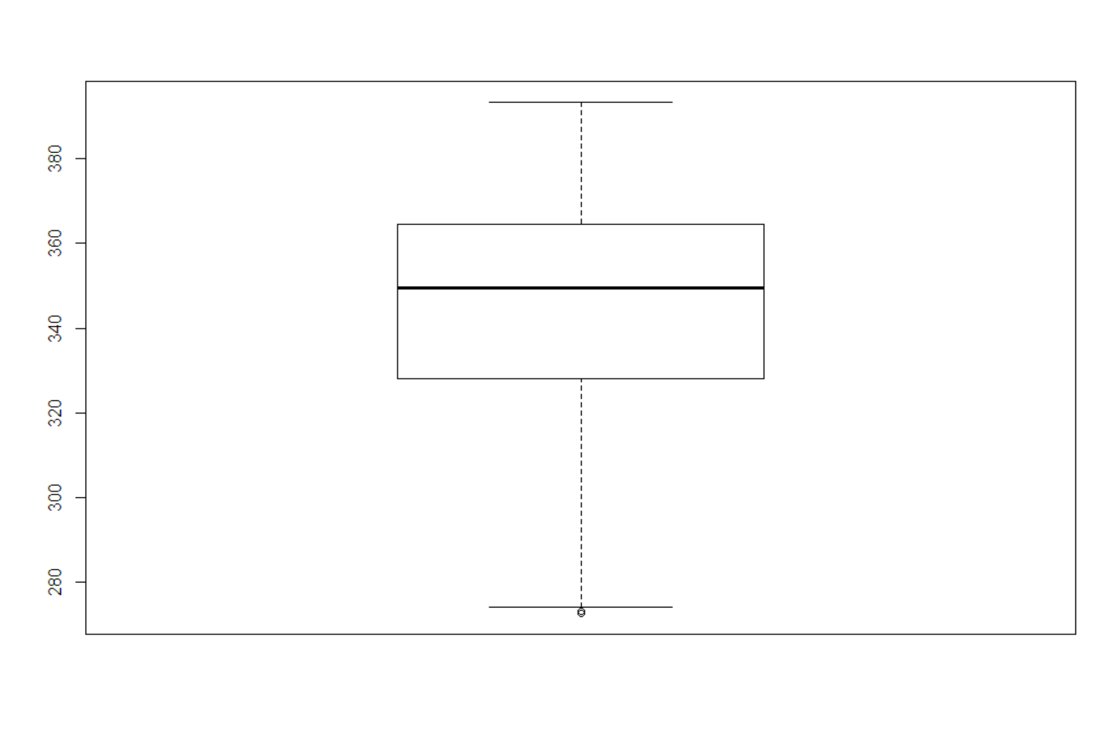
- Cutoff selected for probability of default for logistic regression model was 0.46
- CUTOFF_SCORE= 400 + (slope * (log((1-0.46)/0.46) - log(10)))
- CUTOFF SCORE is equal to **338.18**
- No.of applicants above score 338.18 and thus their credit card application will be accepted as per our model is 47790
- No.of applicants below score 338.18 and thus their credit card application will not be accepted as per our model is 23434

COMPARISON OF SCORES OF APPROVED AND REJECTED CANDIDATES

Boxplot for scores of rejected candidates



Boxplot for scores of approved candidates



- Mean and median score for rejected applicants with missing performance tag is 297.7 and 295.7 respectively.
- Mean and median score for approved customers is 344.1 and 349.5 respectively.
- **Thus mean and median score of approved customers is much higher than those of rejected customers.**

Financial Benefits of the Model

The Confusion Matrix for calculating the Financial gain using our model was made on the dataset without missing Performance tag records, since we need to evaluate how much gain was achieved using our model for applicants who were provided with credit card compared to when no model was used.

	Reference	
Prediction	**0**	**1**
0	45873	1311
1	20980	1635

Profit calculations – with model Vs without model

- We have considered an average profit of Rs.5000 from each non defaulters and
- an average loss of Rs.1,00,000 when each accepted applicant defaults

- Net Profit without model = Rs 3.9665 crores

- Profit using model will be total profit due to each true positive and each true negative minus loss from each false positive and each false negative prediction
- Profit with model = Rs15.6865 crores

- Net financial gain with using our model = **Rs. 11.72 crores**
- Percentage financial gain = **295.47%**

Revenue loss and Potential Credit loss saved

Revenue Loss : Occurs when good customers are identified as bad and credit card application is rejected.

- No of candidates rejected by the model who didn’t default – 20980.
- Total No of candidates who didn’t default – 66853
- % of good candidates rejected by our model – 31.38%

- About 31.38% of the non defaulting customers are rejected which resulted in revenue loss.

Credit Loss Saved : The candidates who have been selected by the bank and have defaulted are responsible for the credit loss to the bank.

- % of candidates approved and then defaulted when model was not used = 4.2%
- % of candidates approved and then defaulted when model was used = 1311/69799 = 1.8%

- Credit loss saved => 4.2 – 1.8 = 2.4%

Conclusion

- ❑ Logistic regression model is chosen as the final Model with 70% of Accuracy.
- ❑ Optimal score cut-off value of 338.18 is derived to approve and reject the applications.
- ❑ By this we found out that credit loss % was decreased when we used this model. Hence it is accurate in rejecting the candidate who may default in future.
- ❑ There is Net Financial gain of 295.47% after using the model.