CredX Risk Analytics Case Study
BFS Capstone Project

Team members:

1. Phani Raj Parakala
2. Sivakrishna Somalanka
3. Naresh Yeluguri
4. Priya Pullipaka

Problem Statement

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. Need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Business Objective

To help CredX identify the right customers using predictive models by defining the factors affecting credit risk and *f*orming strategies to mitigate them.

Solution Approach:

This solution is driven by binary classification problem. We aim at building models such as Logistic Random forest, regression, and other concepts 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:

DATA UNDERSTANDING

1.Demographic data: This dataset contains the information provided by the applicants at the time of credit card application. It comprises customer-level data on age, gender, income, marital status, etc.

2.Credit bureau data: This data is taken from the credit bureau and comprises variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

Data Analysis:

The credit bureau data consists of 71295 observations with 19 variables.
The demographic data consists of 71295 observations with 12 variables.
Application ID is the common key between the two datasets for merging.
Performance Tag is the target variable that indicates if customer has defaulted the amount or not. Non-Defaults are denoted by 0 and defaulted with 1.

DATA CLEANSING AND PREPARATION

DATA QUALITY Checks:

There are 1425 rows are removed from the data as they indicates no Credit Card was given to the Applicant as there are no performance tag

Three duplicate Application ID data are not considered in data set -765011468, 653287861, 671989187

65 records have been excluded from the data set as the they don’t meet minimum age 18 for issuing Credit card variables

Since 18 is the minimum age to grant credit card, records with age <18 has been excluded from the dataset.

Data Issue

Data	No# Missing Values	Error records
Performance Tag	1425	
Education	119	
Profession	14	
Type of residence	8	
Marital Status	6	
No of dependents	3	
Gender	2	
Application ID	-	3 Duplicate Applications Ids
Age	-	65 Applicants with age less than 18
Income	-	81 Applicants have income less than 0

Data	No# of missing values	Error data
Performance Tags	1425 Applicants	
Avgas CC Utilization - Last 12 months	1058 Applicants	
Presence of open home loan	272 Applicants	
Outstanding Balance	272 Applicants	
No of trades opened in last 6 months	1 Applicant	
Application ID	0 Applicants	

Weight Of Evidence (WOE) AND Information Value Analysis (IV)

WOE and IV values are calculated using woe.binning package in R. Continuous quantitative variables that has WOE values are not monotonically changing across bins. Coarser bins were made by decreasing the number of bins until monotonic behavior is observed across bins. Above mentioned 9 variables with Missing values has the variable values replaced by their corresponding WOE values.

New variable – reverse_perf.Tag with inversed relationship for IV analysis as package treats 1 as 'good'

Information Values Analysis values denotes that demographic data don't play much significant role in forecast.

Top 12 Variables that have IV values of 0.1 to 0.3 has medium projecting influence and are considered significant and no significant variable that has strong predictive authority.

Information Value Analysis

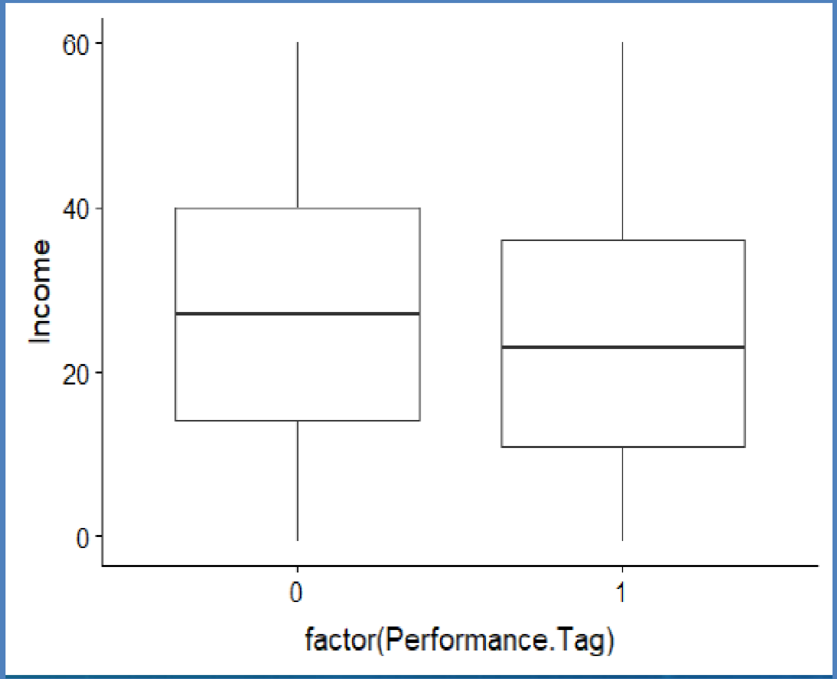
Variable Values	Information Values
Avgas.CC.Utilization in last 12 Months	0.260755415
Number of enquiries in last 6 months	0.092939144
Number of enquiries in last 12 months	0.271544682
Number of times 30 DPD or worse in last Six Months	0.241562739
Number of times 30 DPD or worse in last Twelve Months	0.198254858
Number of times 60 DPD or worse in last Six Months	0.205833876
Number of times 60 DPD or worse in last Twelve Months	0.185498873
Number of times 90 DPD or worse in last Six Months	0.160116924
Number of times 90 DPD or worse in last Twelve Months	0.213874838
Number of trades Opened in last 12 Months	0.194337383
Number of PL trades Opened in last 6 Months	0.124743691
Number of PL trades Opened in last 12 Months	0.176644264
Total Number of Trades	0.182235069
Income	0.0424178
Age	0.003349157
Application ID	0.001504195
Number of months in Current residence	0.078943527
Presence of Open Home Loan	0.017626529
Number of months in Current Company	0.021754413
WOE Professional binned	0.002182094
WOE Gender Binned	0.000324971
WOE of Martial Status Binned	0.0000952
WOE of Type of Residence binned	0.000289274

EXPLORATORY DATA ANALYSIS

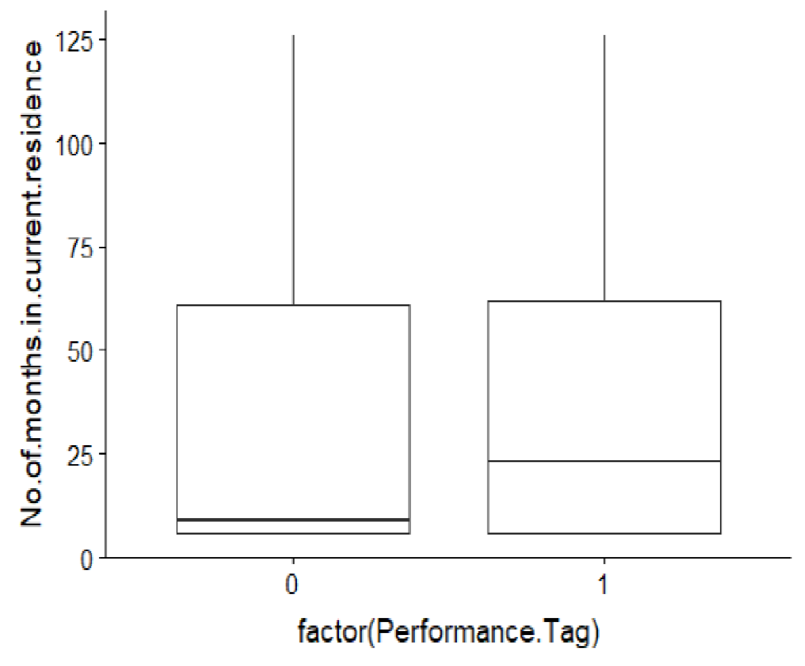
MEDIAN value Defaulters Vs Non-Defaulters

- ✓ The median values for income of defaulters are lower than that of non-defaulters.
- ✓ The median values for Number of months for current residence of non-defaulters are lower than that of defaulters.
- ✓ The median Number of months for current Company of non-defaulters is slightly lower than that of defaulters.

Median values for income of defaulters are lower than that of non-defaulters

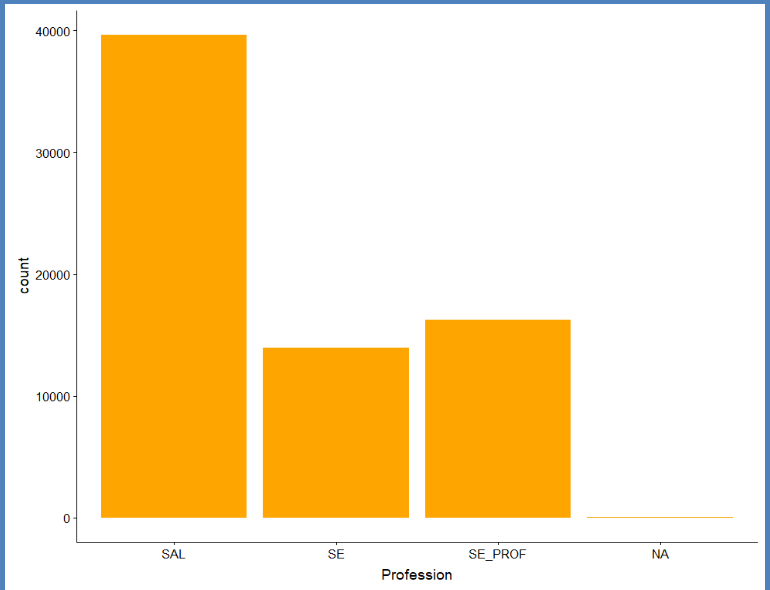


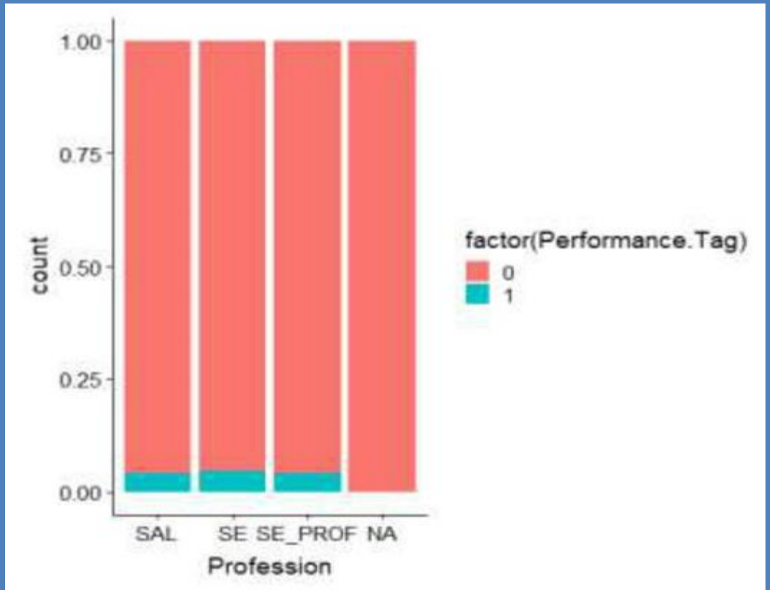
Median current residence duration of non-defaulters are lower than that of defaulters.



EDA ANALYSIS

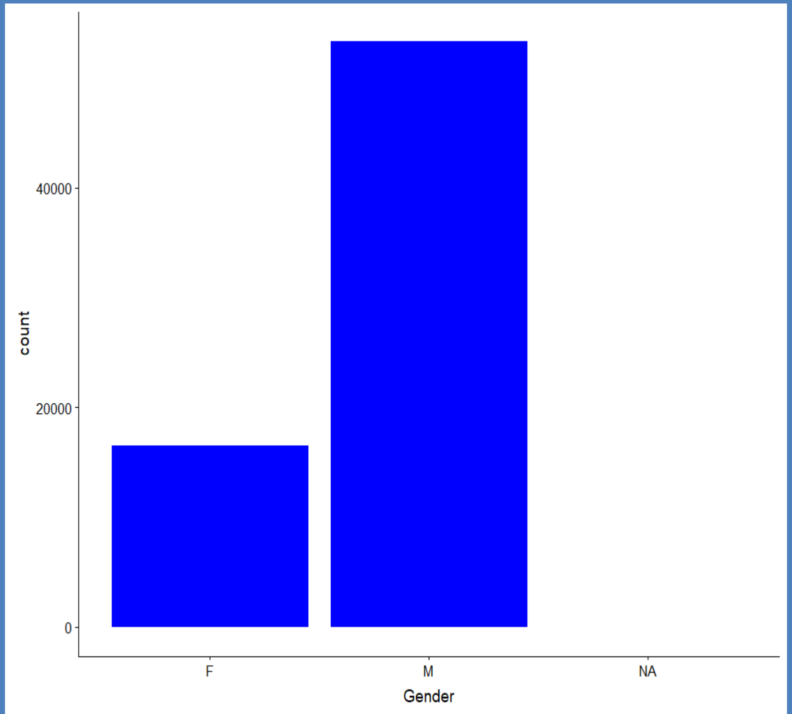
There are more applicants whose profession is SAL but there is no difference in default rates.

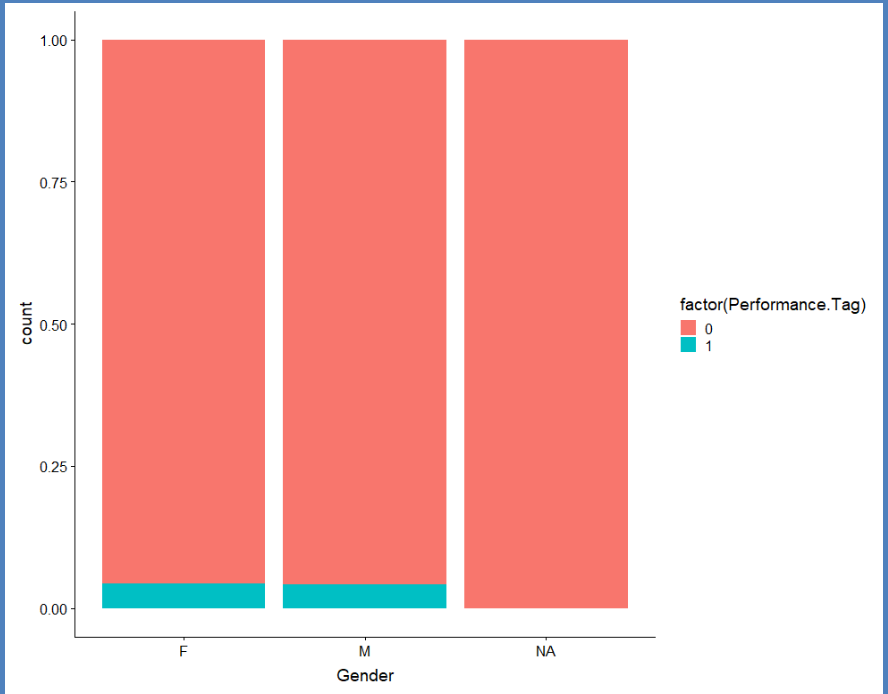




EDA ANALYSIS

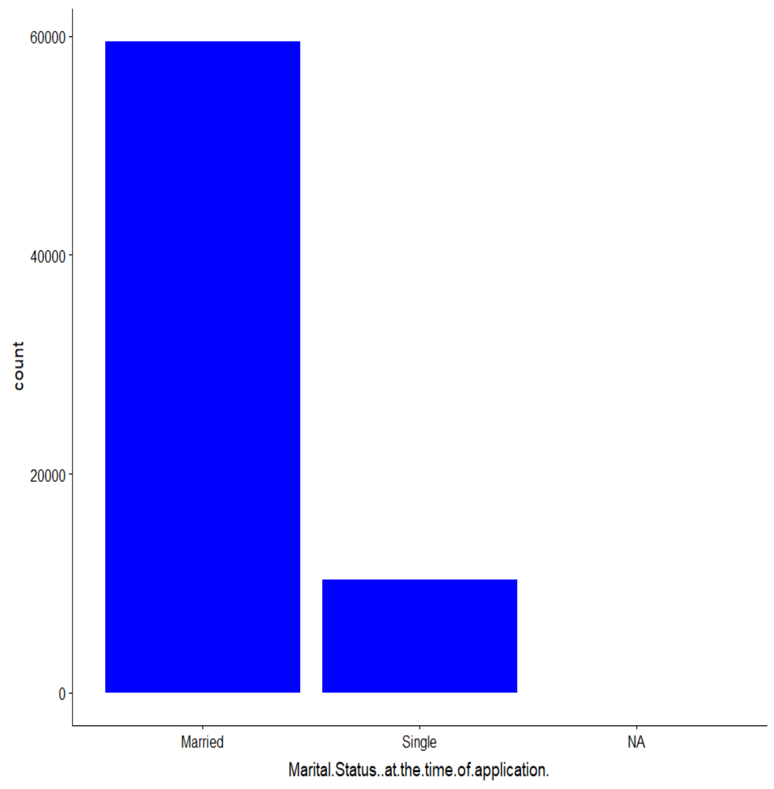
There are more male applicants than female applicants but there is no difference in default rates.

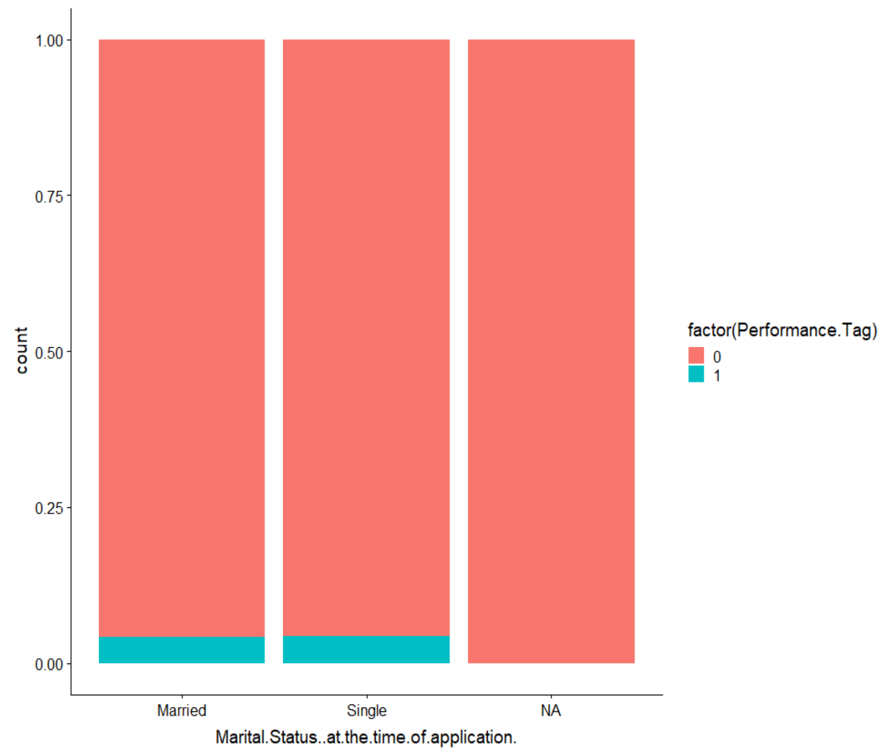




EDA ANALYSIS

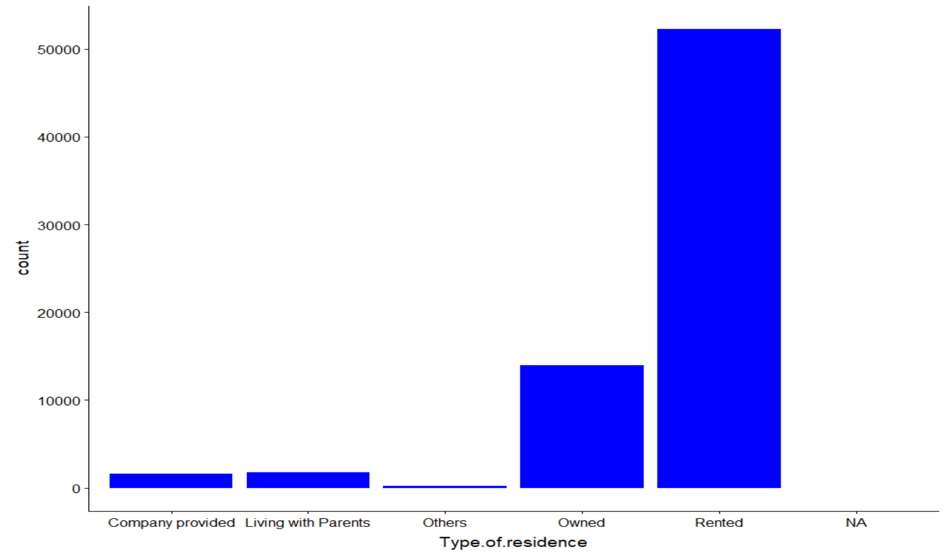
There are more married applicants than single applicants but there is no difference in default rates.

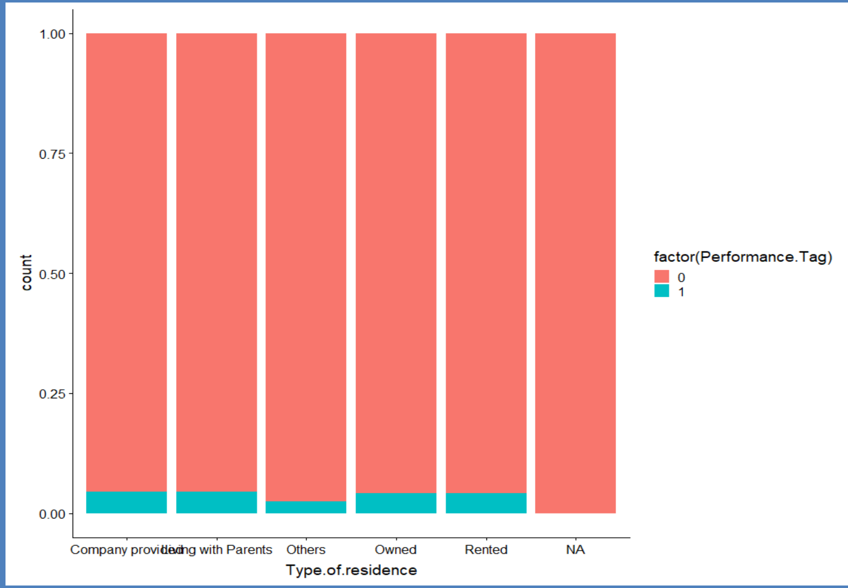




EDA ANALYSIS

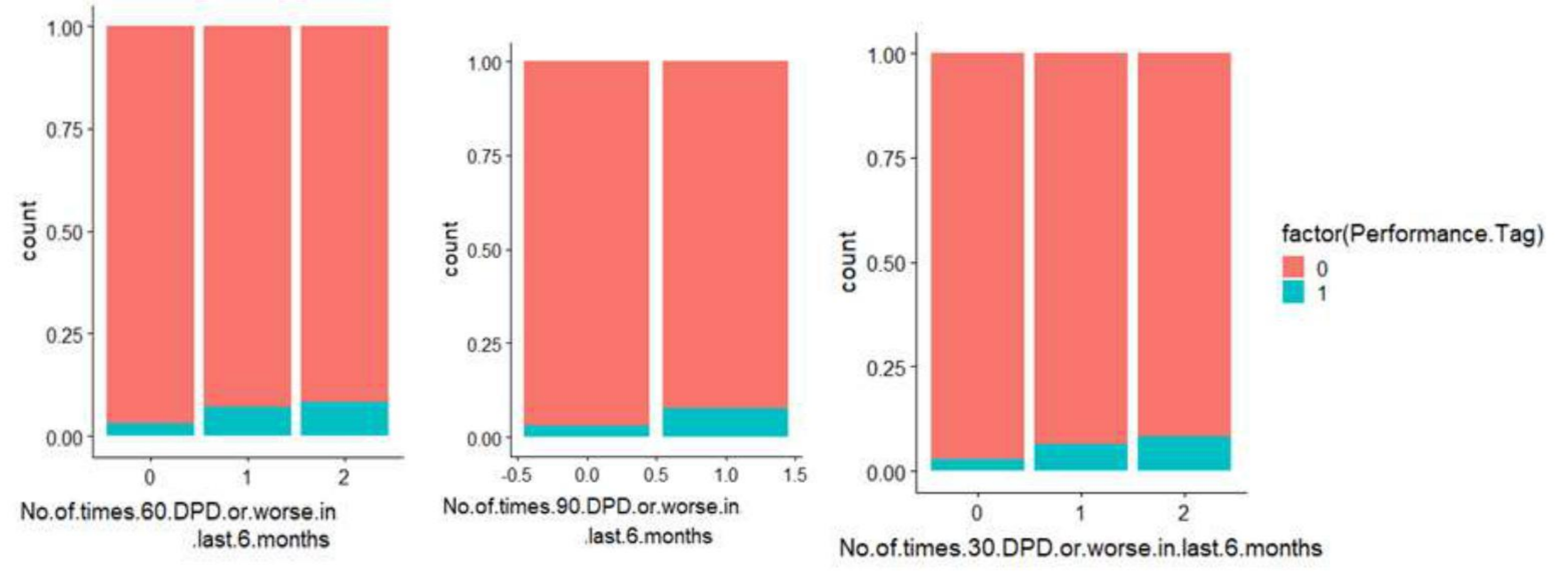
There are more Applicants staying for rent rather than owning or Company Provided





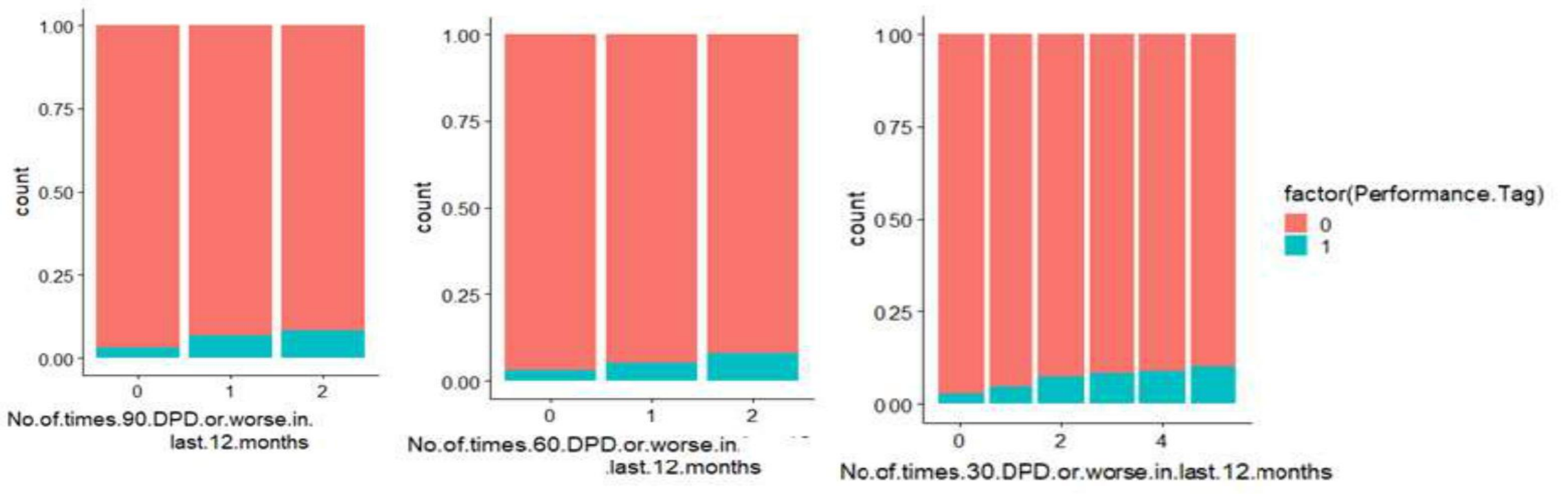
EDA ANALYSIS

For last 6months Variable values , the defaulters numbers are increasing with increase in Number of 30/60/90 DPD or worse. This variable is another significant predictor.



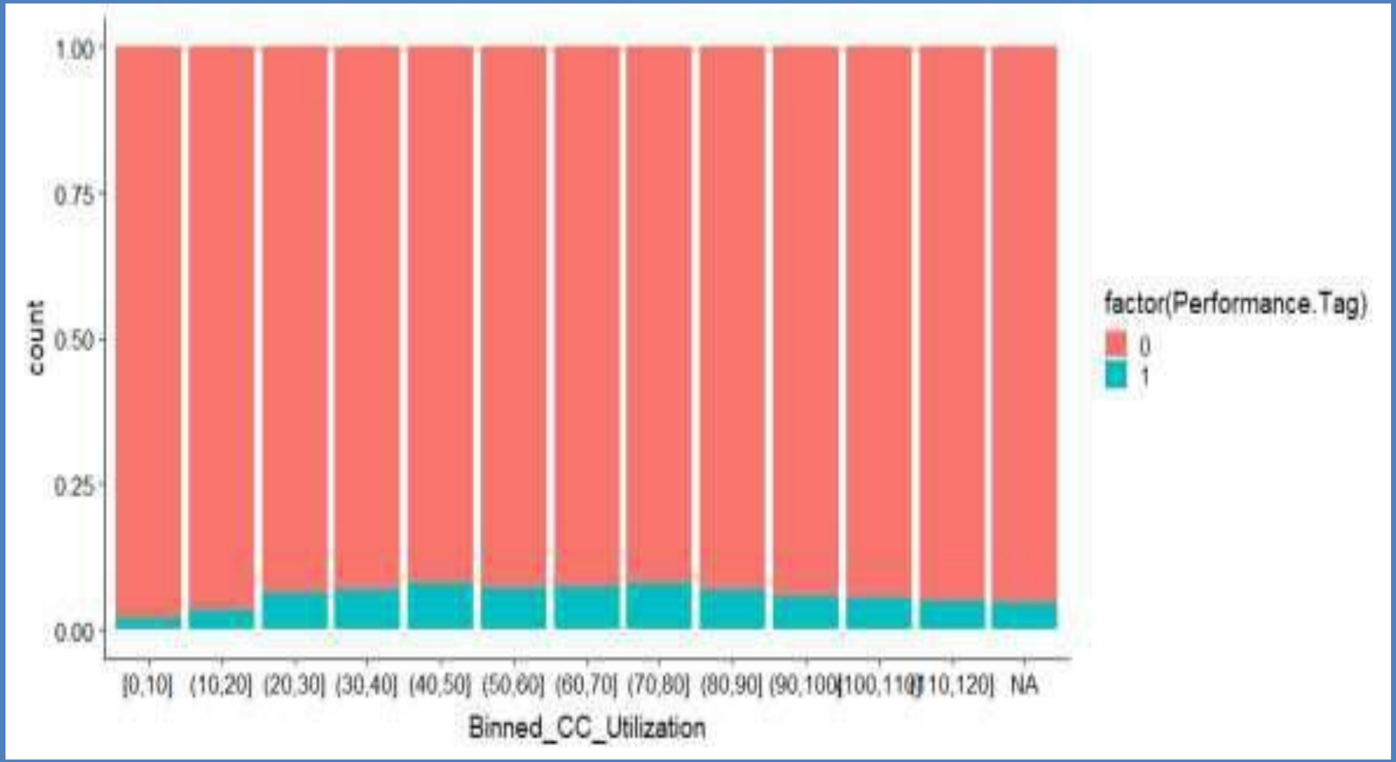
EDA ANALYSIS

Defaulters Numbers are increasing with increase in 30/60/90 DPD or worse in last 12 months' variable values. This reflects it is an important predictor.



EDA ANALYSIS

Average Credit Card Utilization does not show any significant Pattern details.

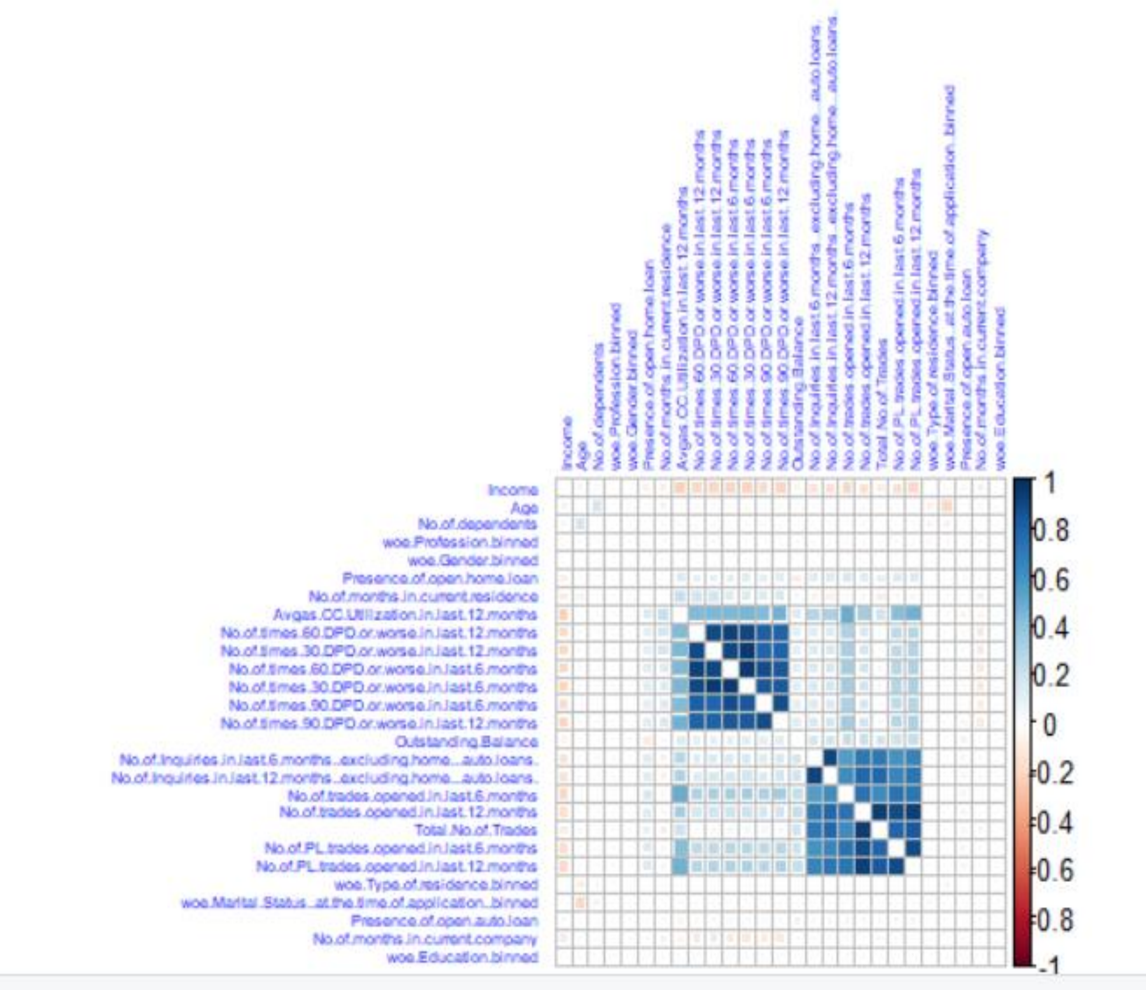


CORRELATION MATRIX

Correlated with variables within the group

A-GROUP

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



B-GROUP

- ✓ 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.ofTrades
- ✓ No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
- ✓ No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

DATA TRANSFORMATION

Sample of data:

As the data is not balanced ROSE package was used for balancing data sets. It is a Smoothed Bootstrap and helps in analysis of data using sampling methods

Train and Test DATA ratio:

After all exclusions, the dataset contains 69,799 records for analysis and the dataset is split into Train and Test in 70:30 ratio.

OUTLIERS:

Box plots used to detect Outliers on Continuous variables and
The variables with outliers have been mapped to nearest non-outlier values.

SCALING:

Application ID and performance tag has been excluded from Scaling to standardize the data into common scale.

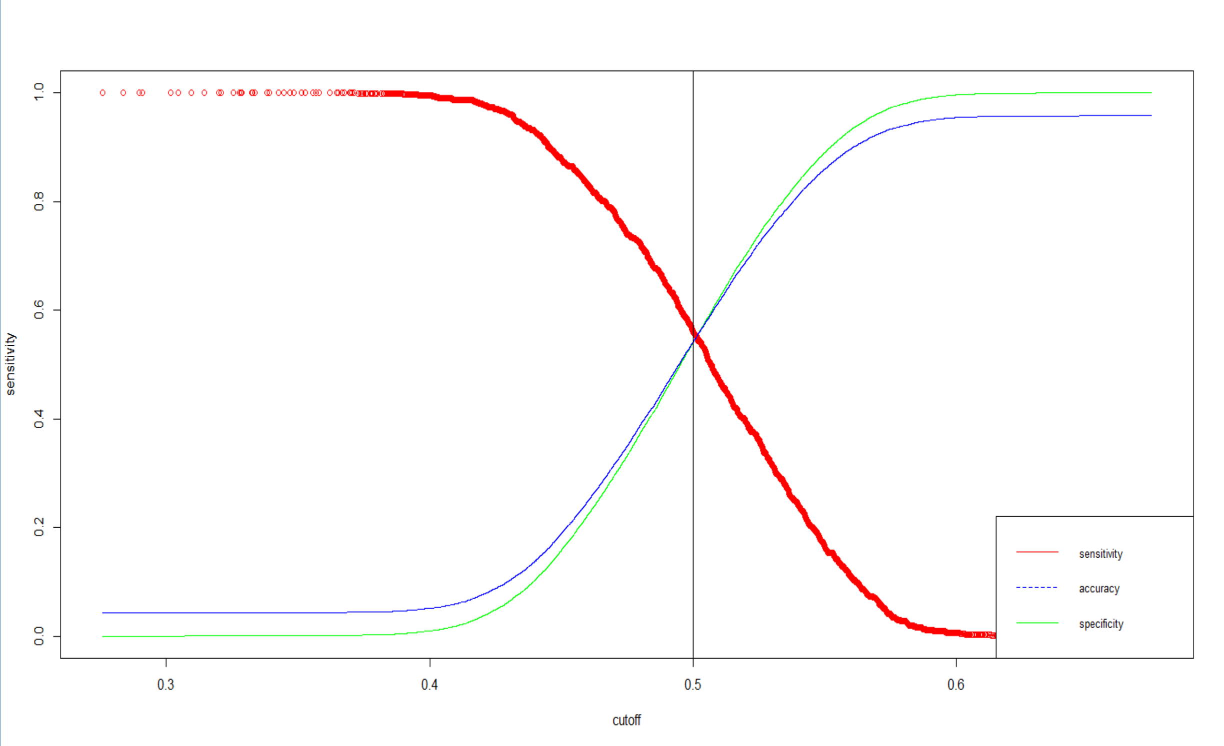
LOGISTICS REGRESSION

Logistic Regression with below two algorithms
1. Based on AIC Stepwise variable selection
2. Based on VIF and p value Backward variable selection

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

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

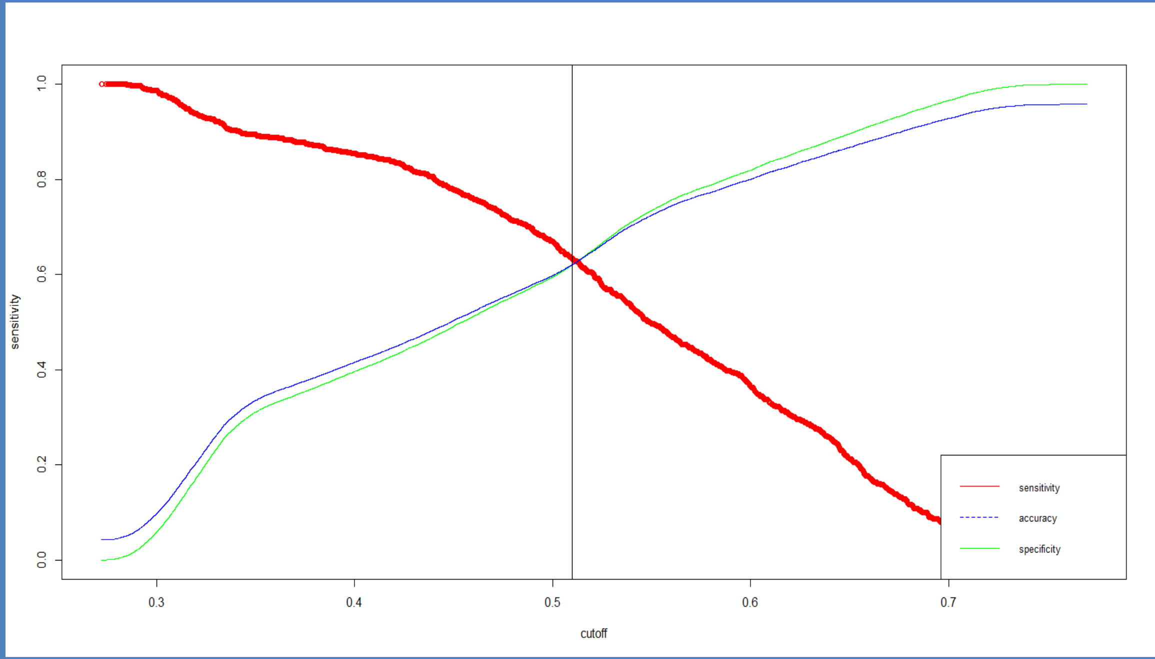
- ✓ INCOME
- ✓ AGE
- ✓ NO.OF.MONTHS.IN.CURRENT.RESIDENCE
- ✓ NO.OF.MONTHS.IN.CURRENT.COMPANY
- ✓ WOE.PROFESSION.BINNED
- ✓ WOE.TYPE.OF.RESIDENCE.BINNED
- ✓ WOE.EDUCATION.BINNED



LOGISTICS REGRESSION

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 :

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



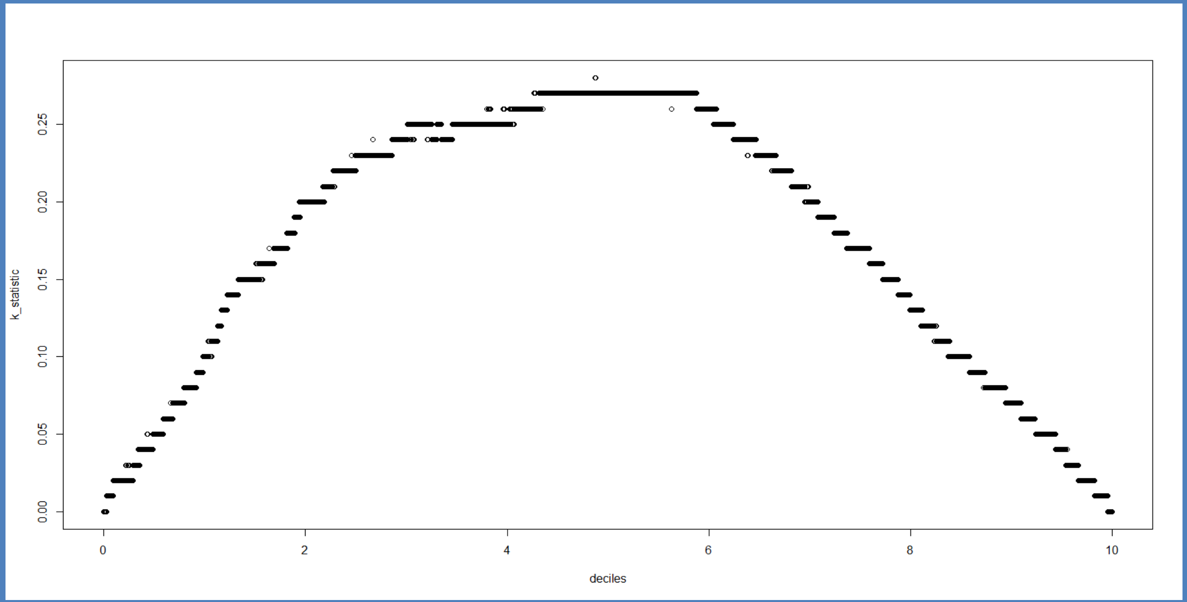
LOGISTICS REGRESSION

CONFUSION MATRIX

Prediction	0	1
0	12436	324
1	7620	560

Statistics	Statistics
Specificity	62%
Sensitivity	63%
Accuracy	62%

KS CHART:
KSSTATISTIC FOR THIS MODEL IS 0.27
AND LIES WITHIN IN FIRST 5 DECILES

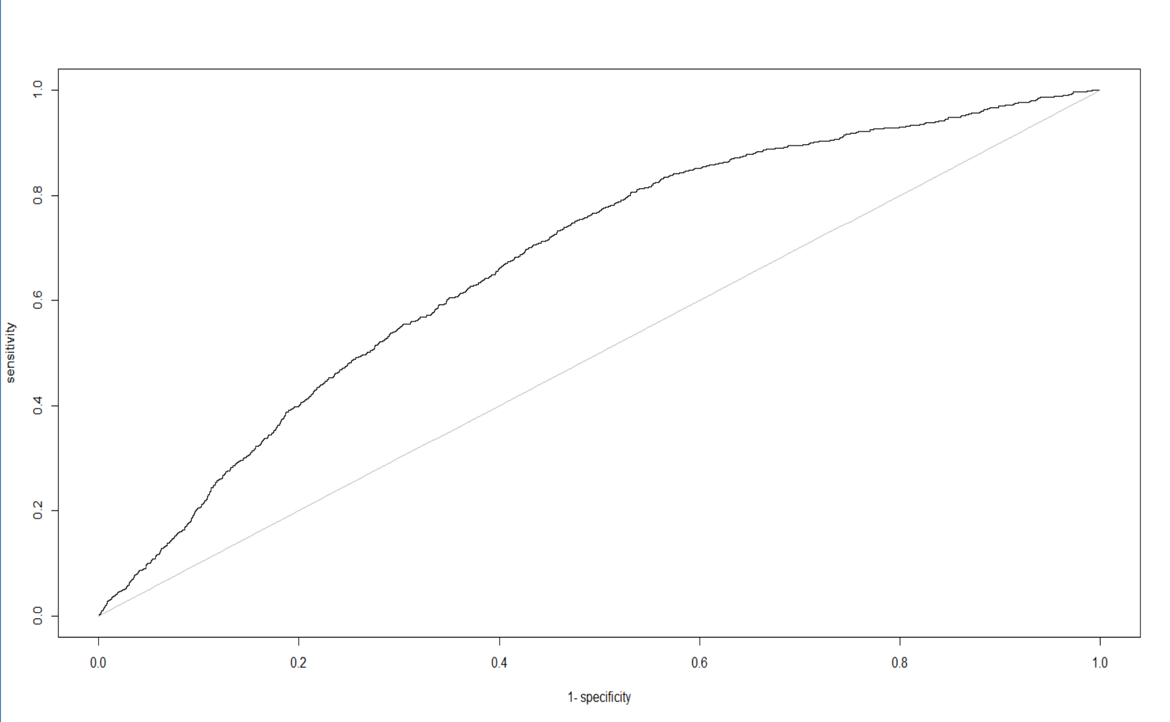


LOGISTICS REGRESSION
AREA UNDER CURVE

TEST DATA- Cross Validation

Test Data- 1 :	Test data -2:
Sensitivity=62%	sensitivity=62%
Specificity=60%	specificity=63%
Accuracy=62%	accuracy=62%

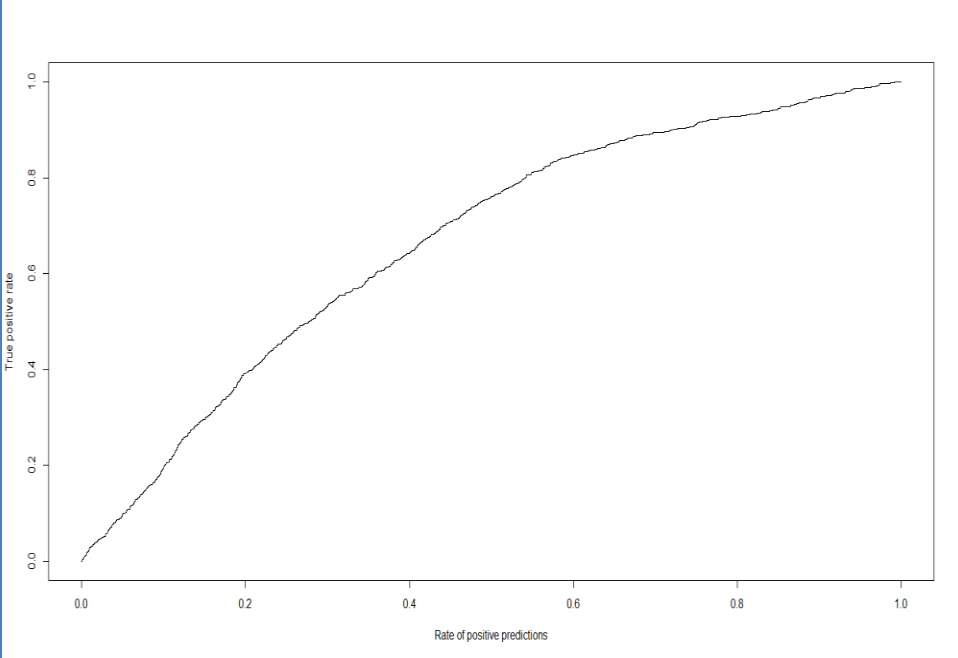
AREA UNDER ROC CURVE = 0.67

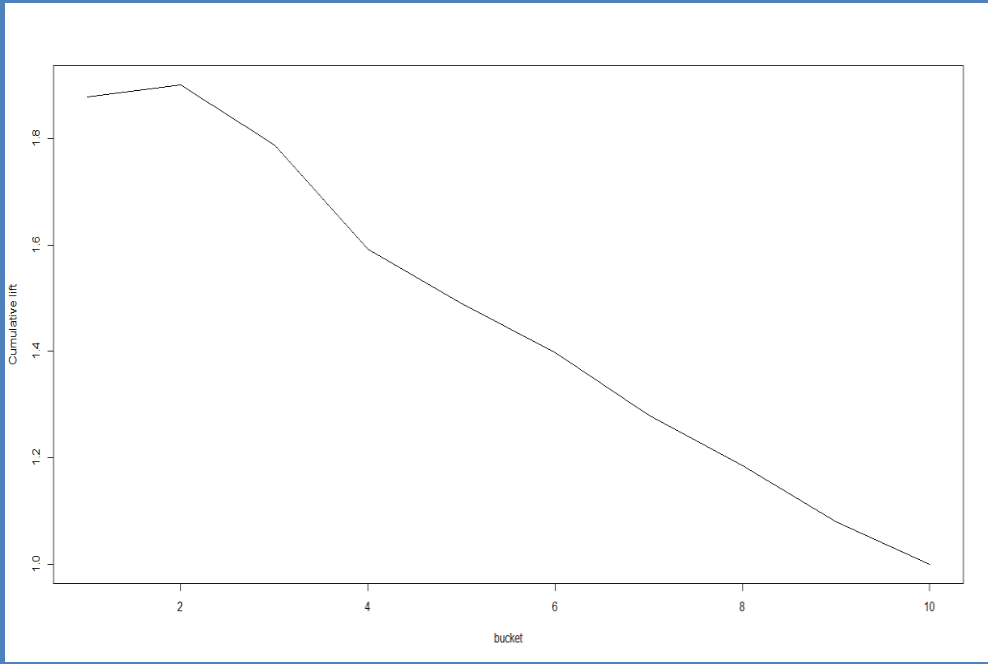


LOGISTIC REGRESSION GAIN & LIFT CHARTS

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





DIFFERENT MODEL'S ACCURACY, SENSITIVITY & SPECIFICITY Without REJECTED Dataset 1425

Models	Accuracy	Sensitivity	Specificity
Logistic Regression on Demographic dataset	55.00%	55.00%	55.00%
Random Forest on Demographic dataset	54.49%	54.56%	52.94%
Logistic Regression on Merged dataset	63.00%	63.00%	63.00%
SVM model with linear Kernel on Merged dataset	53.44%	52.59%	72.51%
SVM model with RBF Kernel on Merged dataset	67.42%	67.86%	56.35%
Random Forest on Merged dataset	63.40%	63.40%	63.60%

COLCLUSION BASED on MODELS(EXCLUDING REJECTED DATASET)

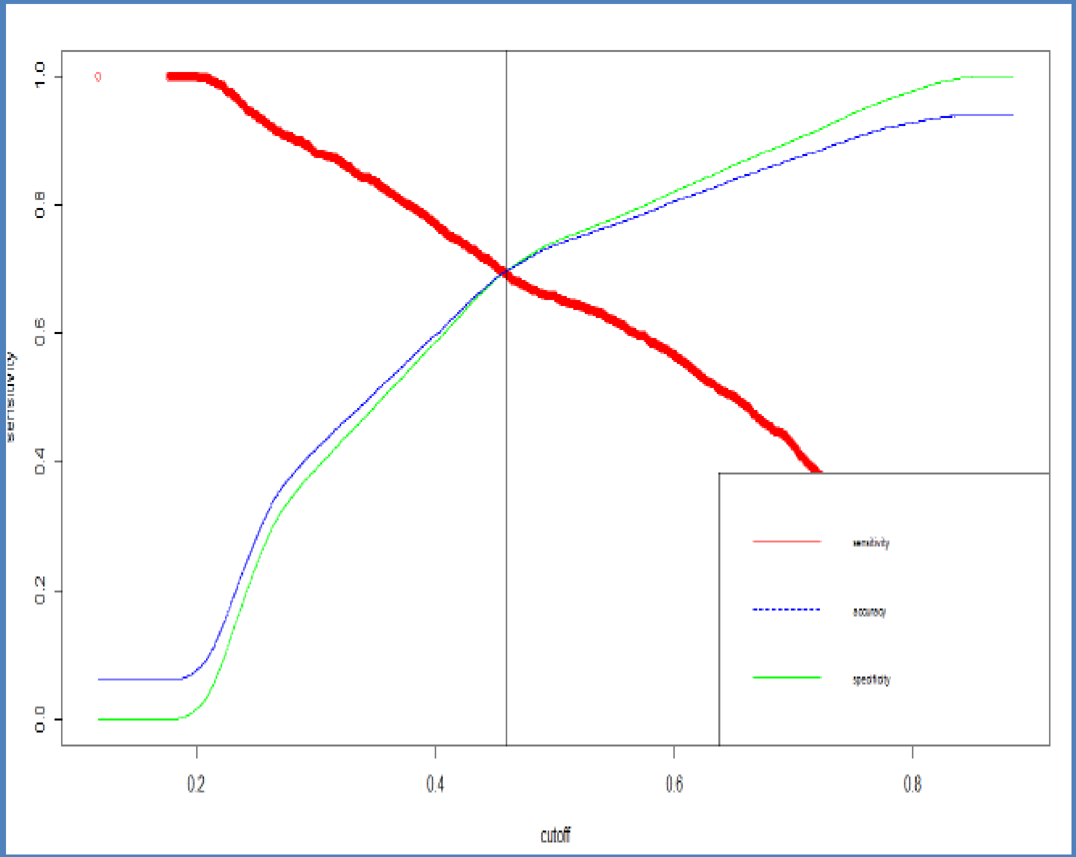
Logistic regression model used on the combined (credit bureau + demographic dataset) to predict the performance tag values missing for 1425 applicants .

All the Models showed low performance evaluation metrics with maximum accuracy being nearly 64%.

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

LOGISTICS MODEL
-INCLUDING REJECTED 1425 RECORDS

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

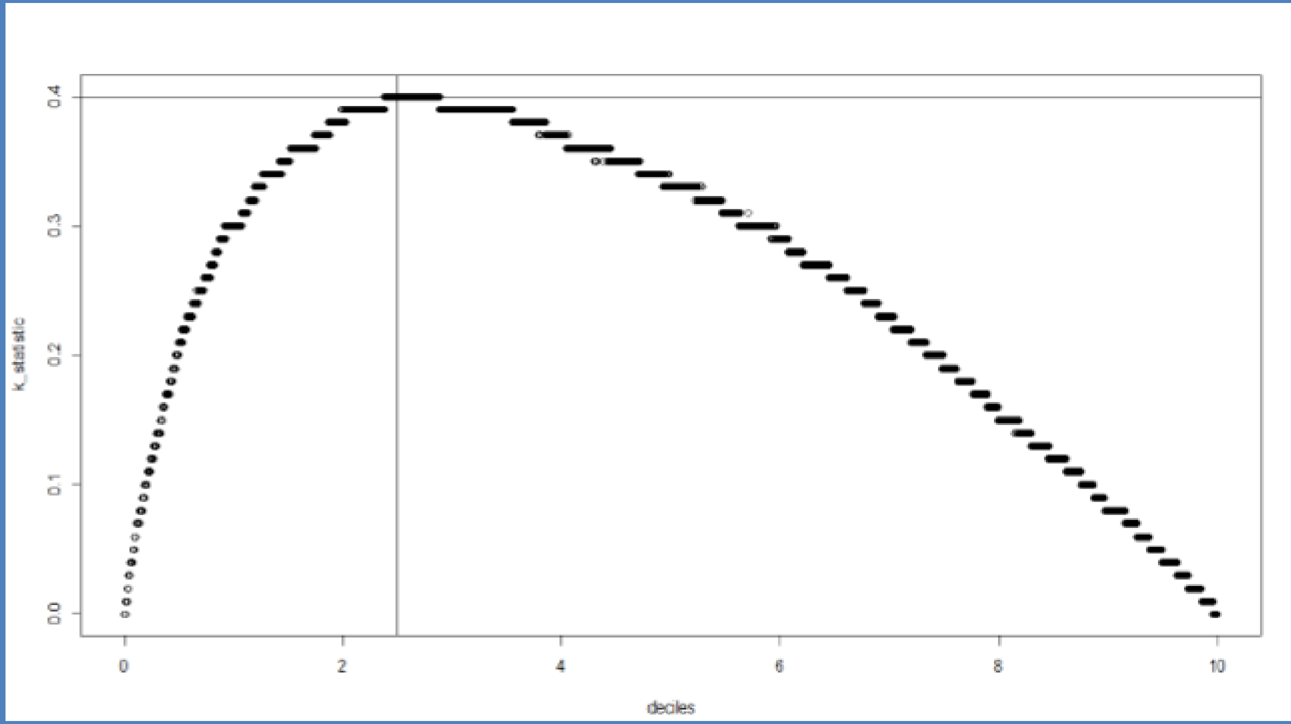


LOGISTIC REGRESSION MODEL ON MERGED DATASET (WITH REJECTED 1425 RECORDS)

Prediction	0	1
0	14001	404
1	6058	905

Statistics	Values
Specificity	69.14%
Sensitivity	69.80%
Accuracy	69.76%

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



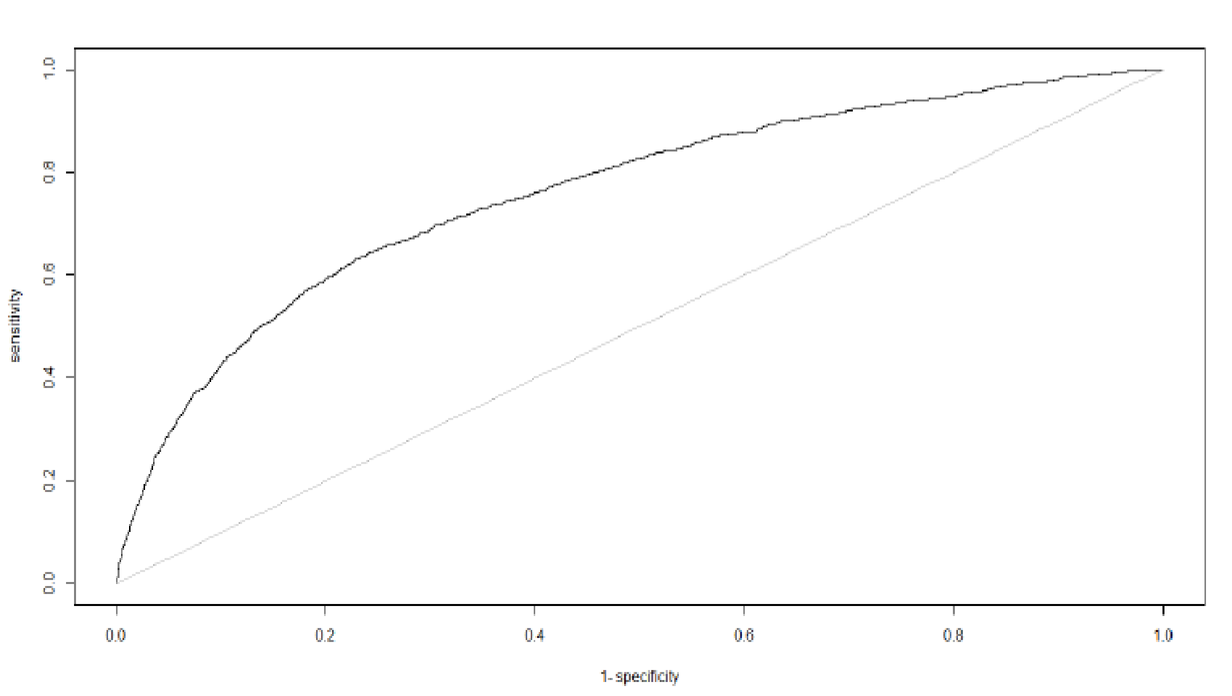
LOGISTIC REGRESSION MODEL ON MERGED DATASET Including Rejected Records
AREA UNDER THE CURVE

TEST DATA- Cross Validation

Test Data- 1 :	Test data -2:
Sensitivity=69%	sensitivity=69%
Specificity=69%	specificity=68%
Accuracy=69%	accuracy=69%

AREA UNDER ROC CURVE = 0.759

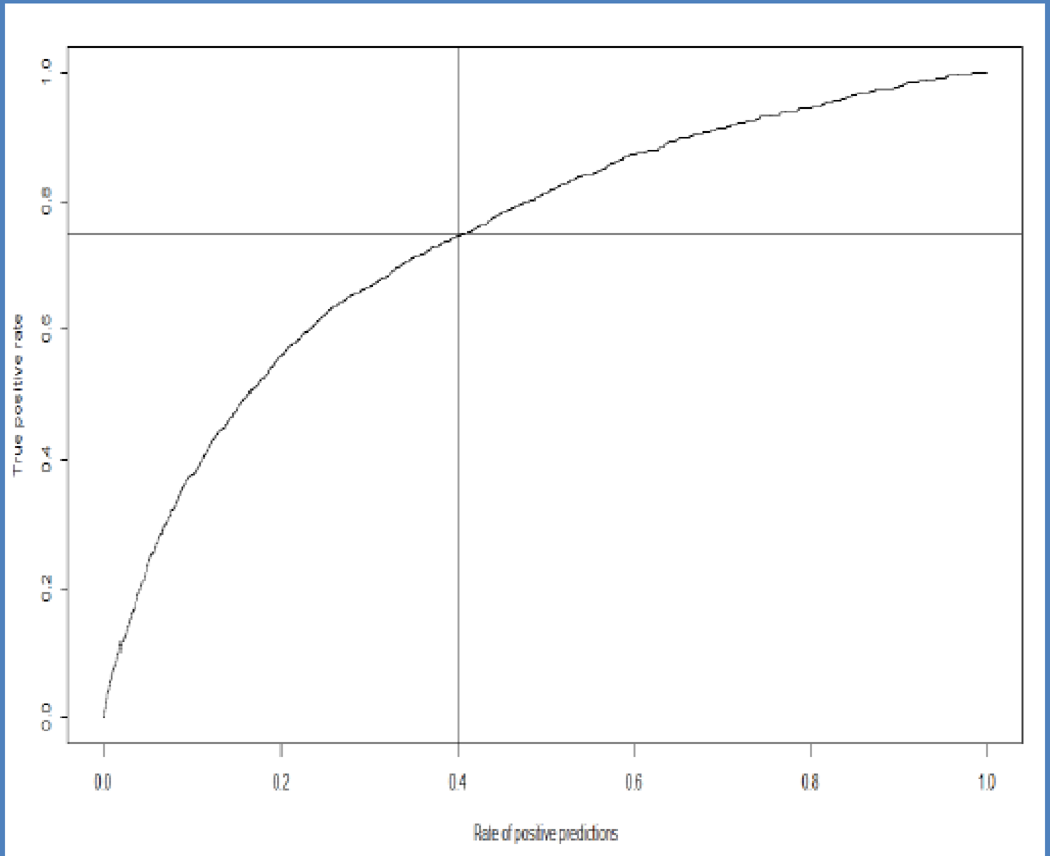
AREA UNDER THE CURVE :



LOGISTIC REGRESSION MODEL ON MERGED DATASET Including Rejected Records
GAIN & LIFT CHARTS

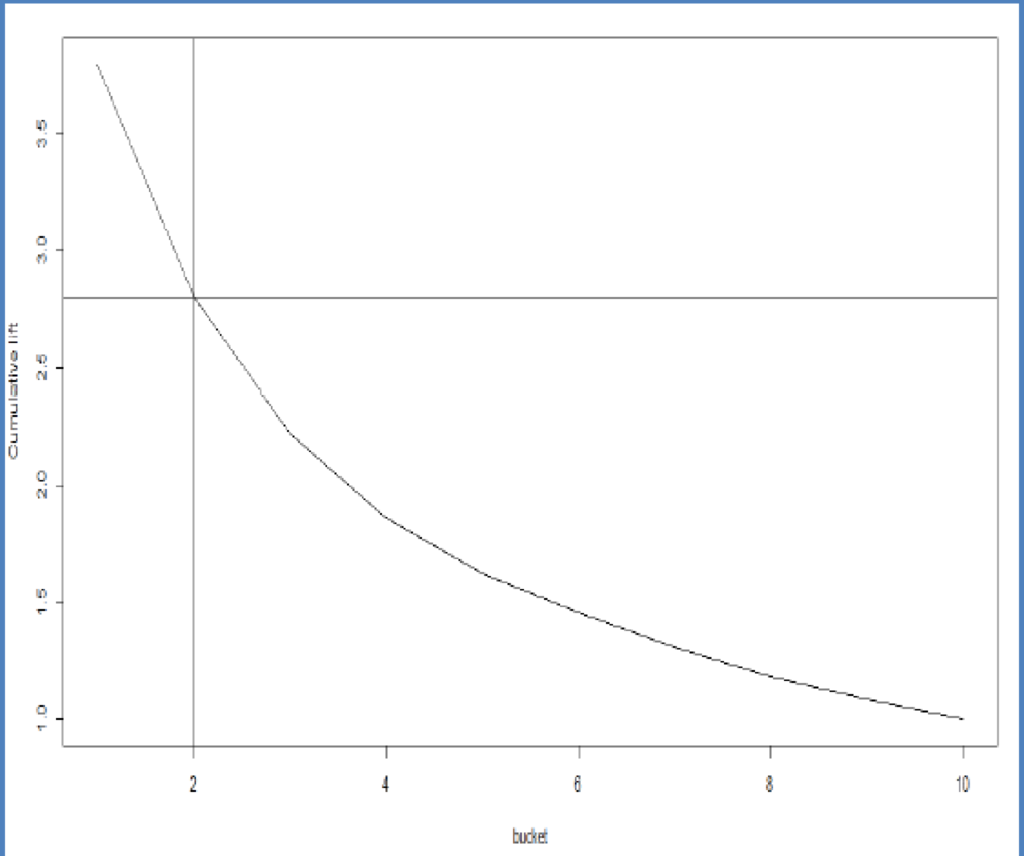
GAIN CHART

Within first 4 deciles able to predict 75% of defaulters correctly using this model.



LIFT CHART

A lift of 2.8 times within first 2 deciles compared to random mode using this model



ACCURACY, Sensitivity and specificity Logistic regression VS Random forest

Statistics	Logistic Regression Including Regression	Random Forest Including Regression
Specificity	69.14%	69.44%
Sensitivity	69.80%	69.63%
Accuracy	69.76%	69.62%

INFERENCE:
Merged Data with Missing Tags records, Logistic regression model is performing better than Random Forest

APPLICATION SCORE CARD

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 scorecard was derived from following:

1.Odds for good was calculated.

Since the probability computed is for rejection (bad customers), Odd(good) = (1-P(bad))/P(bad)

2. Probability is Calculated for all default for applicants

3.Odds of 10 to 1 being a score of 400 where Score increases by 20 points for doubling odds.

3.Probability is Calculated for all default for applicants

4.ln(odd(good)) was calculated

5) 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 Cardvalues:

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

APPLICATION SCORE CARD

Cutoff 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

- Number of applicants above score 338.18 and these credit card application will be accepted as per this chosen model is 47790

- Number of applicants below score 338.18 and these credit card application will be not be accepted as per our model is 23434

Potential Credit loss and Revenue loss saved

Potential 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%

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.

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.

Profit calculations –with model Vs without model

- ✓ Considered an average profit of Rs.5000 from each non defaulters
- ✓ An average loss of Rs.1,00,000 when each accepted applicant defaults
- ✓ Net Profit without model = Rs3.9665 crores
- ✓ Profit using model will be total profit due to each true positive
- ✓ 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%

FINAL MODEL SELECTION

- ✓ Logistic regression model is Selected 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 percentage was decreased when we used this model and it was appropriate refusing the candidate who may default in future.
- ✓ Net Financial gain of 295.47% after using the model.