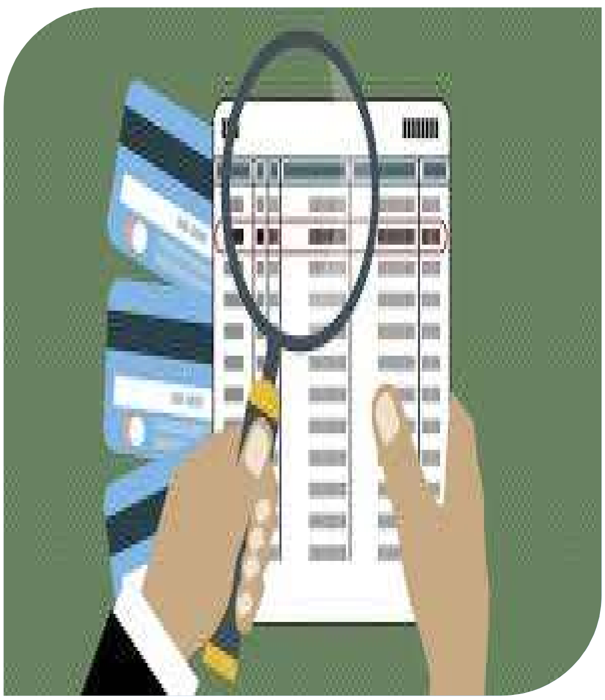
CREDX RISK ANALYTICS CASE STUDY

BFS Capstone Project

Batch Jun-2018



By:

SHARATH ATHREY

SUMIT KAUSHIK

VANDHANA SHRI

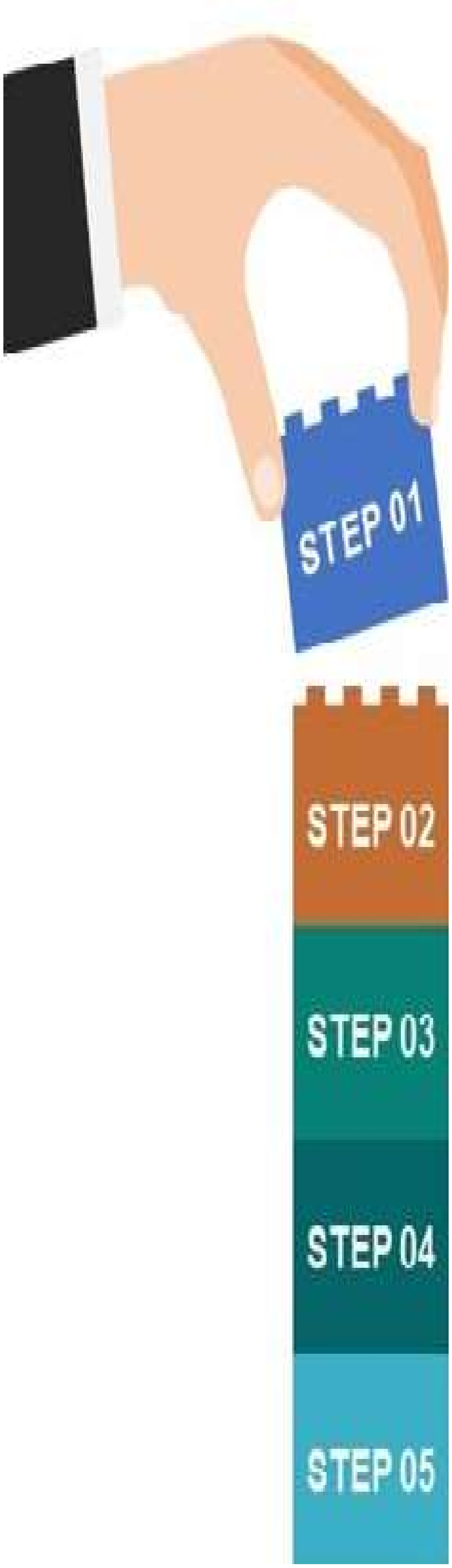
Problem Statement:

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. The best strategy to mitigate credit risk is to 'acquire the right customers'.

Objective:

- Identify the right customers using predictive models.
- Using past data of the bank's applicants, determine the factors affecting credit risk.
- Create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

APPROACH TO SOLUTION



Business Understanding/Data Understanding

Data Preparation/EDA

Data Transformation/Model Building

Model Evaluation

Score Card Application

DATA UNDERSTANDING

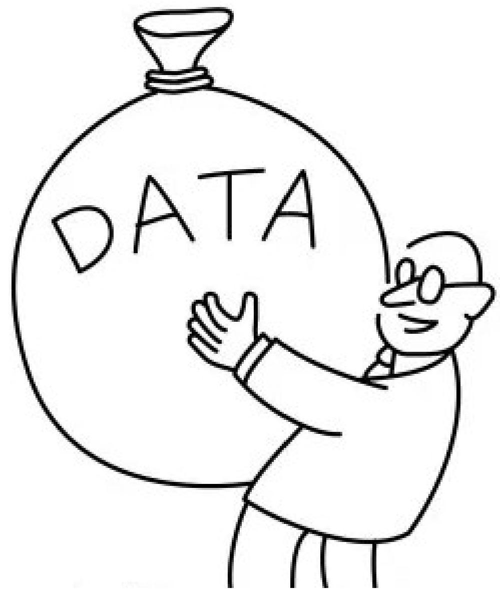
Two Datasets provided namely,

- Demographic Data:

    Information provided by the applicants at the time of Credit card application.

- Credit Bureau Data :

    It provides all details of past transaction taken from the credit bureau.



DEMOGRAPHIC DATA	CREDIT BUEREAU DATA
Application ID	Application ID
Age	No of times 90/60/30 DPD or worse in last 6 months
Gender	No of times 90/60/30 DPD or worse in last 12 months
Income	No of Trades opened in last 6/12 Months
Marital status	No of PL Trades opened in last 6/12 Months
Education	Total Number of trades
Profession	No of Inquiries in last 6/12 Months (Excl Home/Auto Loans)
Number of dependants	Presence of open Home/Auto Loan
Number of Months in Curreny Company	Avg CC Utilzation in last 12 months
Number of Months in Curreny Residence	Outstanding Balance
Performance Tag	Performance Tag

TARGET VARIABLE with values: 0-Non-Default,1-Default

ASSESSING DATA QUALITY

OBSERVATIONS:

- 71295 rows in both datsets
- Common Unique Variable - Application ID
- Target variable -Performance Tag
- 3 Duplicate Application ID's
  Observed -Excluded from
  further analysis

	Variables	NA's Found	Others
	Performance Tag	1425 in both Datasets -excluded from further analysis	
DEMOGRAPHIC	Age		65 with Age <18-Capped to 18
	Income		with Income <0-
	Gender	2	
	Marital Status	6	
	No of dependents	3	
	Education	118	
	Profession	13	
	Type of residence	8	
CREDIT	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	

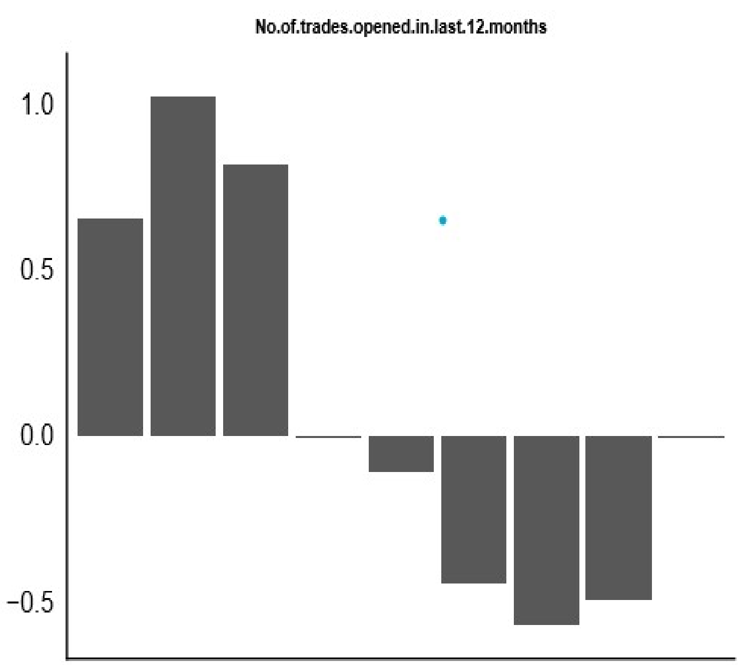
WEIGHT OF EVIDENCE(WOE)

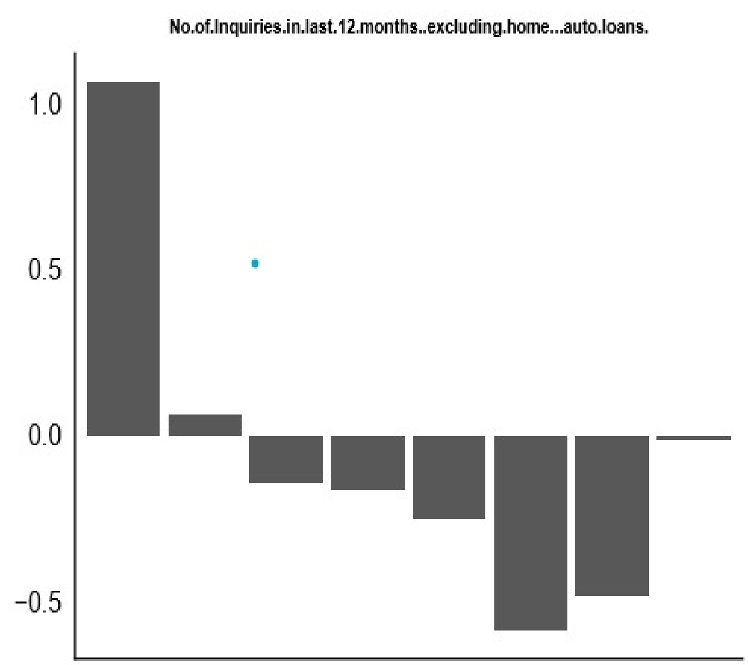
- Compute predictive power of a variable in relation to the dependent variable.
- Impute missing values
- Handle outliers

INFORMATION VALUE (IV)

- Select and rank the most important variables in a predictive model
- No variables from Demographic dataset play a significant role.
- Since the below 6 variables did not monotonically change across bins from the CREDIT DATA, number of bins were reduced such that monotonic behavior is observed across bins. No.of.trades.opened.in.last.12.months,No.of.PL.trades.opened.in.last.6/12.months,No.of.Inquiries.in.last.6/12.months,Total.No.of.Trades

Eg: Non-monotonic



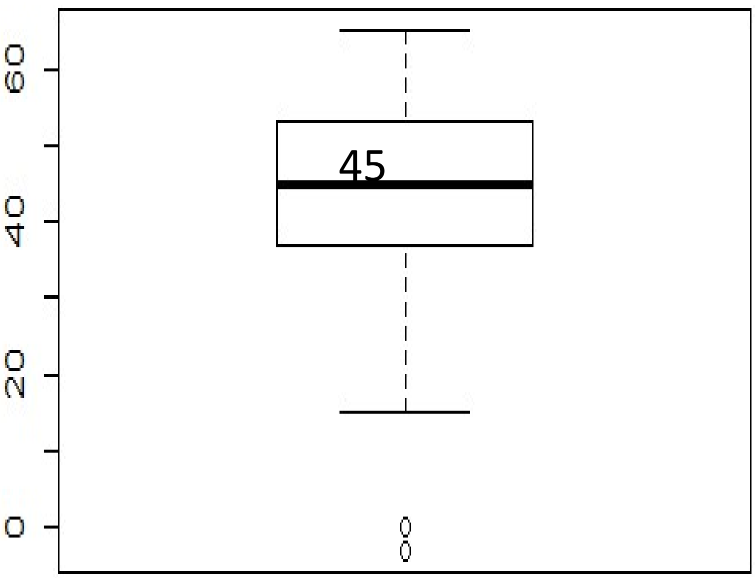


SIGNIFICANT VARIABLES FROM IV VALUES

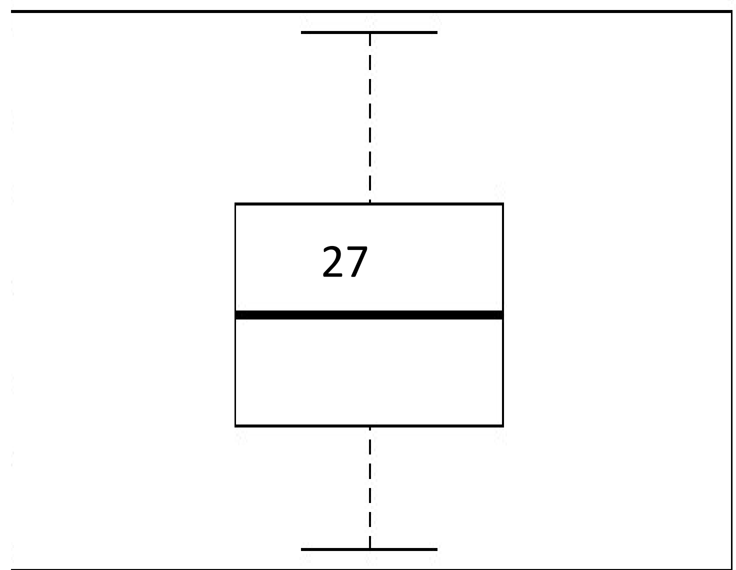
Variable	IV
No of Inquiries.in.last.12.month -excl home,auto loan	0.27154
Avgas.CC.Utilization.in.last.12.month	0.26076
No.of.times.30.DPD.or.worse.in.last.6.month	0.24156
No.of.times.90.DPD.or.worse.in.last.12.month	0.21387
No.of.times.60.DPD.or.worse.in.last.6.month	0.20583
No.of.times.30.DPD.or.worse.in.last.12.month	0.19825
No.of.trades.opened.in.last.12.months	0.19433
No.of.times.60.DPD.or.worse.in.last.12.months	0.18543
Total.No.of.Trades	0.18223
No.of.PL.trades.opened.in.last.12.months	0.17664
No.of.trades.opened.in.last.6.months	0.16977
No.of.times.90.DPD.or.worse.in.last.6.months	0.16011

EDA-UNIVARIATE ANALYSIS

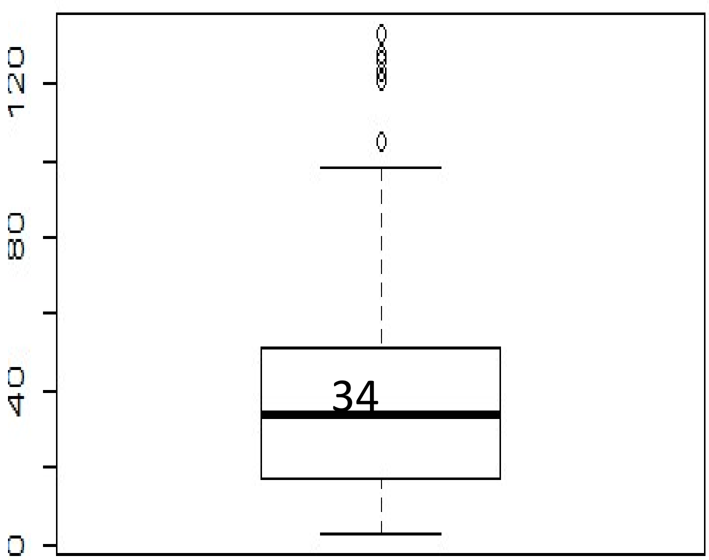
DEMOGRAPHIC DATA



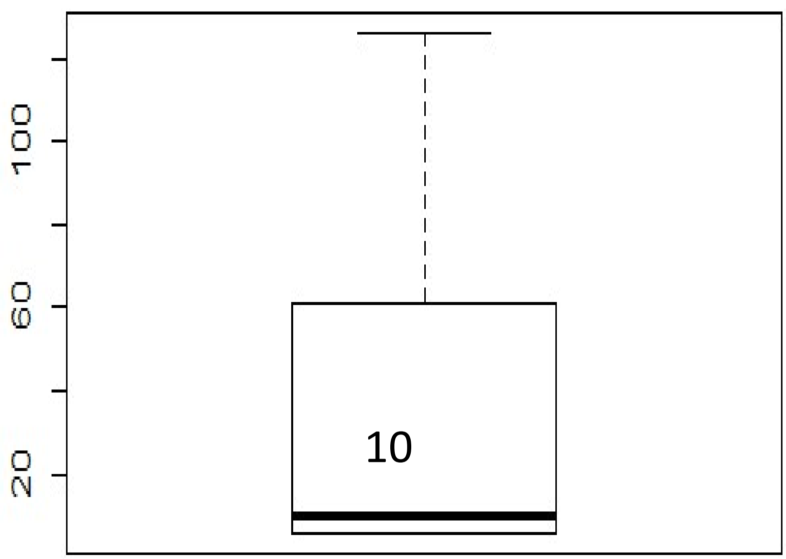
Age



Income



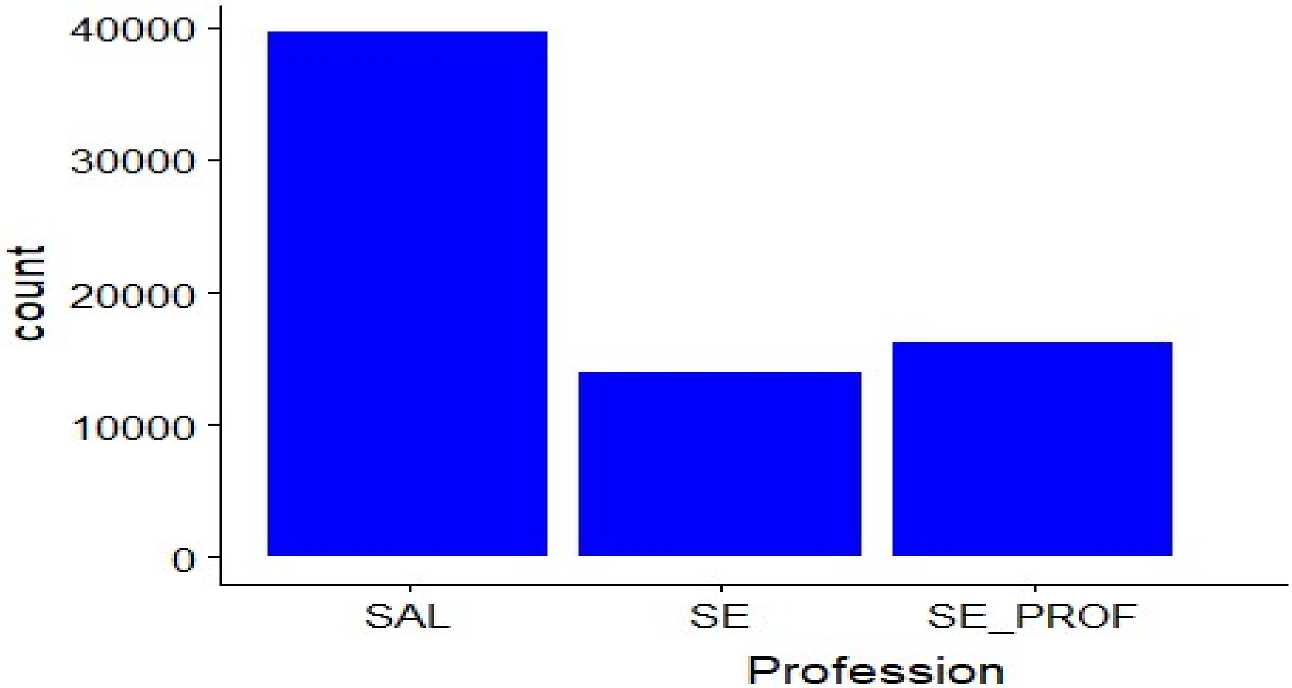
No of Months in Current Company



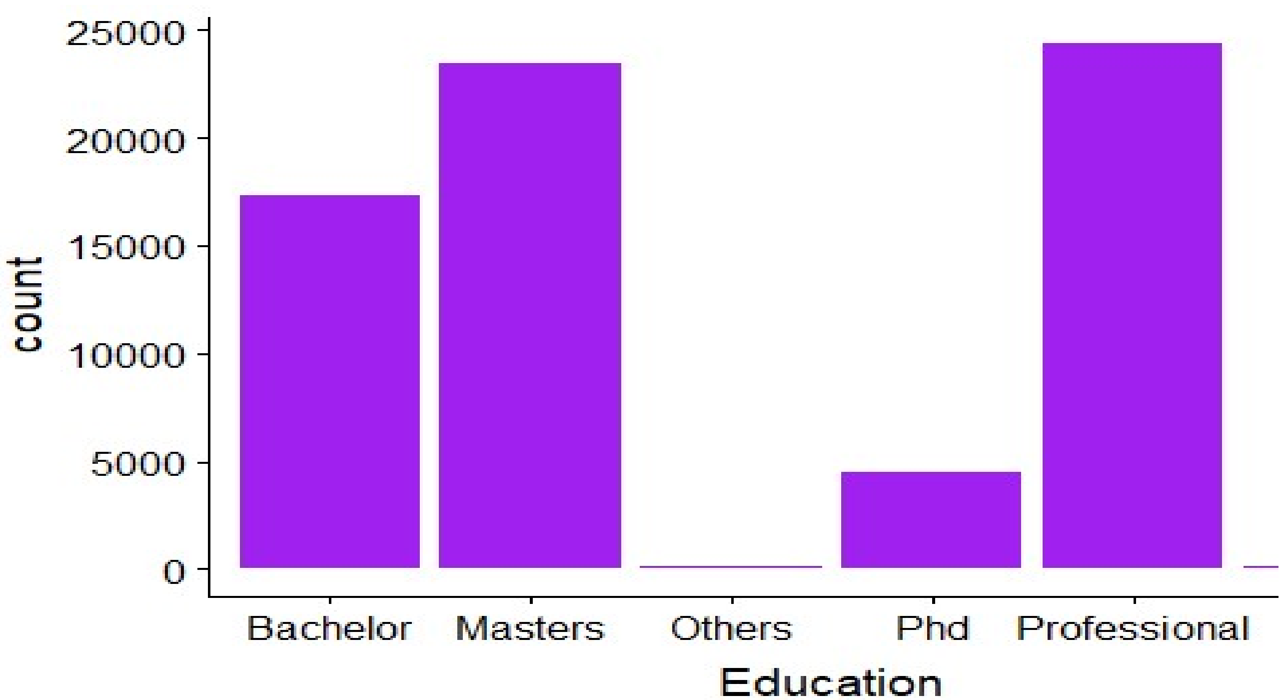
No of months in current Residence

Few Outliers in Age, No of months in current Company

EDA-UNIVARIATE ANALYSIS DEMOGRAPHIC DATA

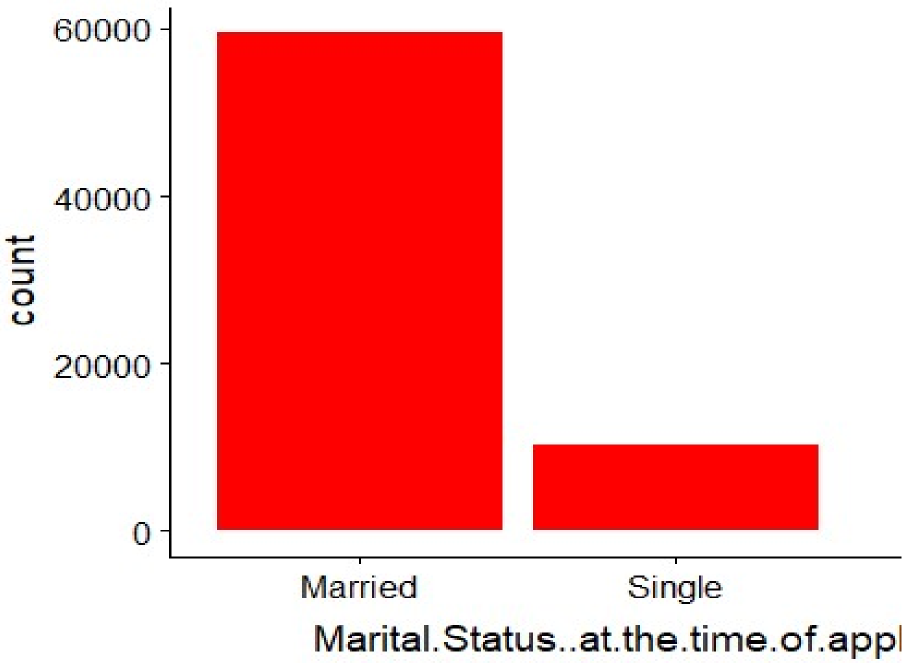


Salaried Professionals apply for credit card more.

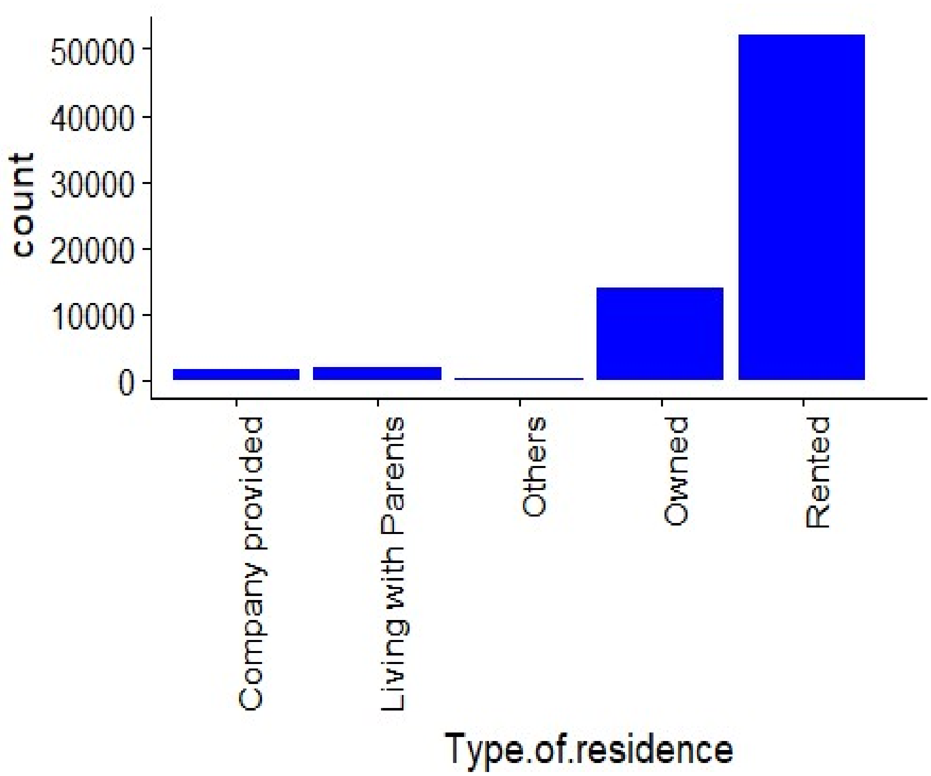


Those with professional degree apply for credit card in large numbers

EDA-UNIVARIATE ANALYSIS
DEMOGRAPHIC DATA

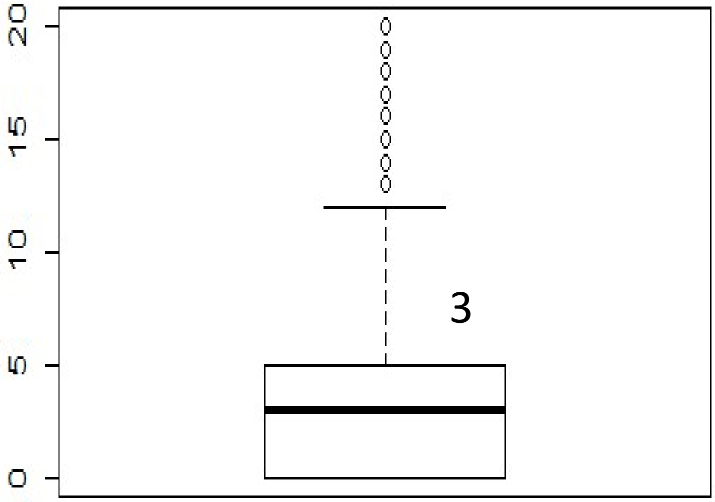


Applicants who are married are higher in number

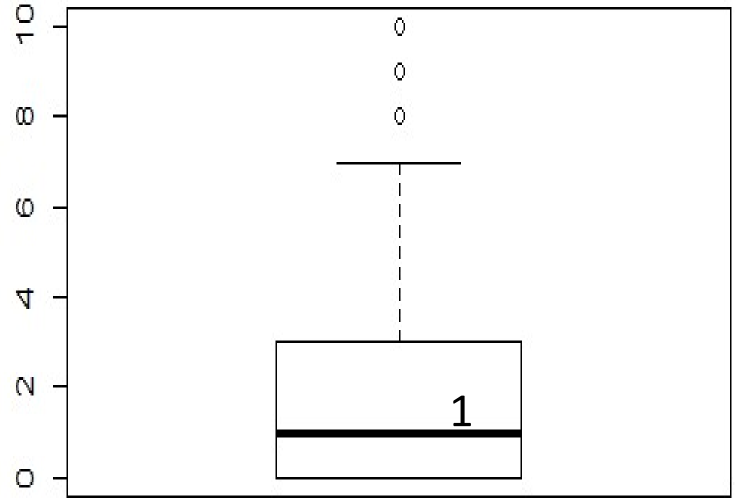


Those who are in rented accommodation tend to apply for credit cards in large numbers

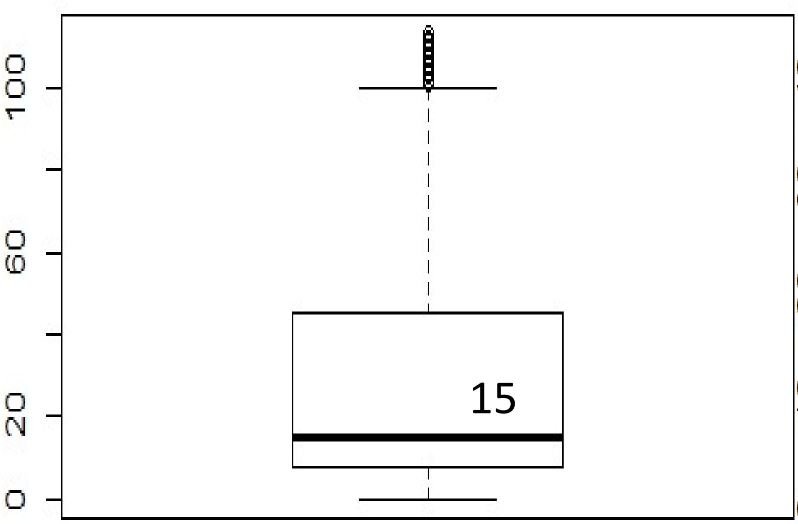
EDA-UNIVARIATE ANALYSIS CREDIT DATA



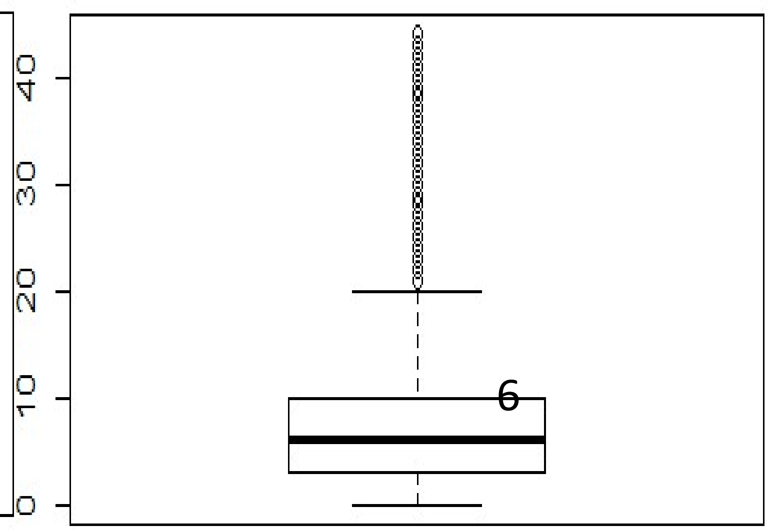
Number of Inquiries in last 12 Months



No of Inquiries in the last 6 months

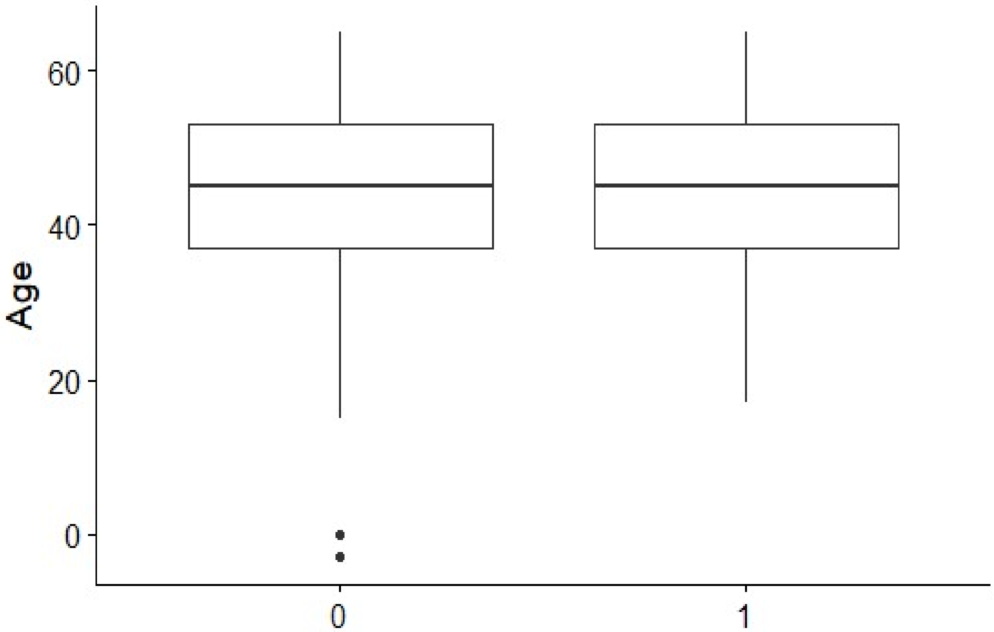


Avgas CC utilization

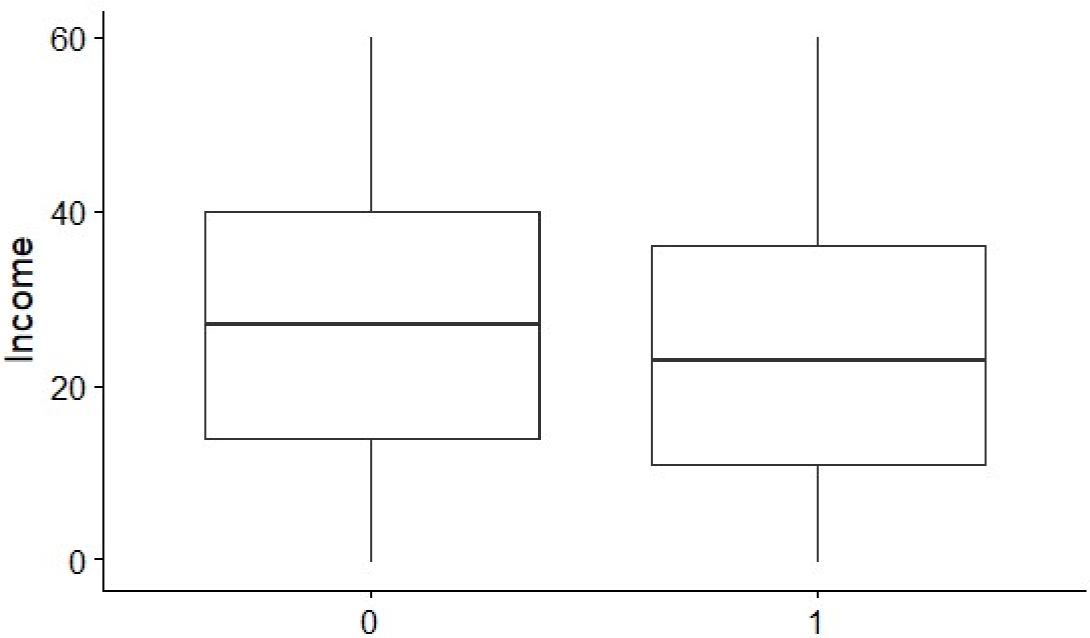


Total number of Trades

EDA-BIVARIATE ANALYSIS DEMOGRAPHIC DATA

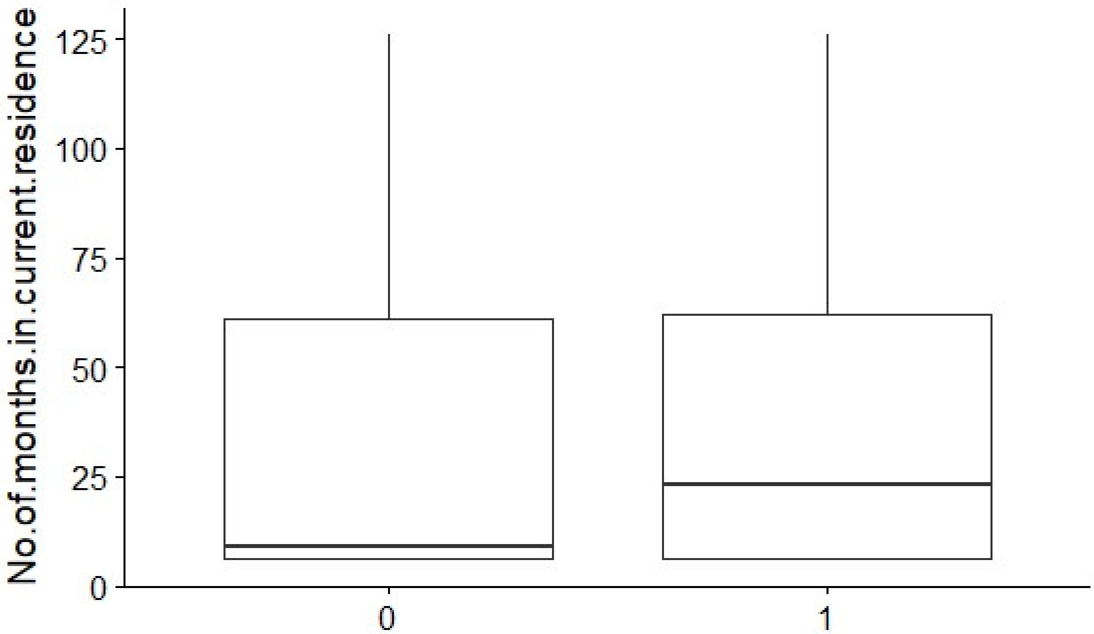


Age does not make a difference in default parameter

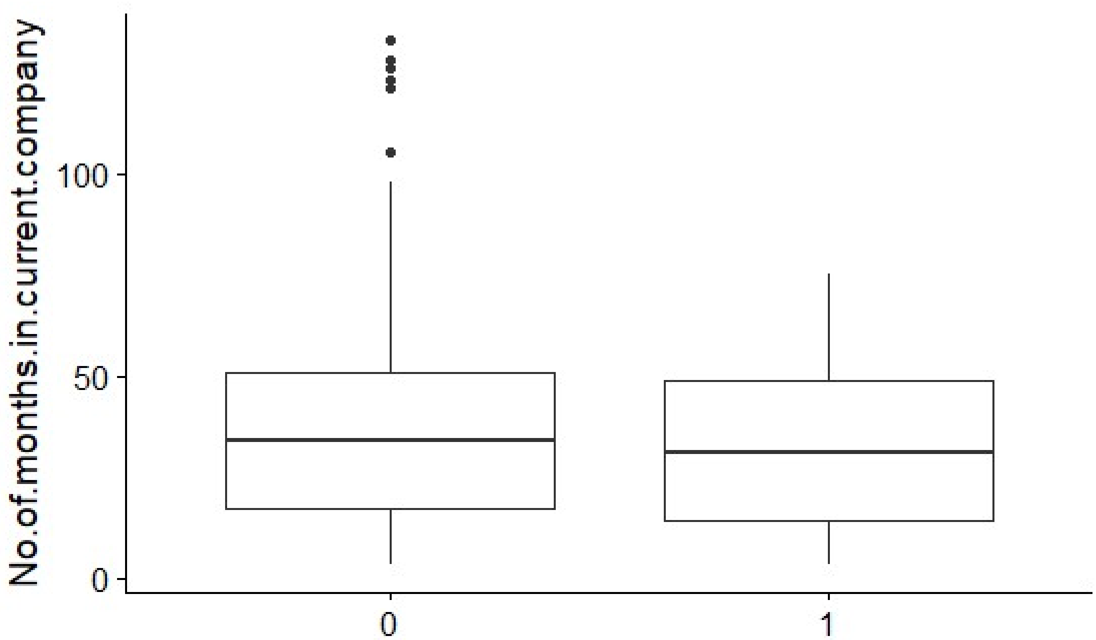


Applicants with a slightly lower income tend to default

EDA-BIVARIATE ANALYSIS DEMOGRAPHIC DATA

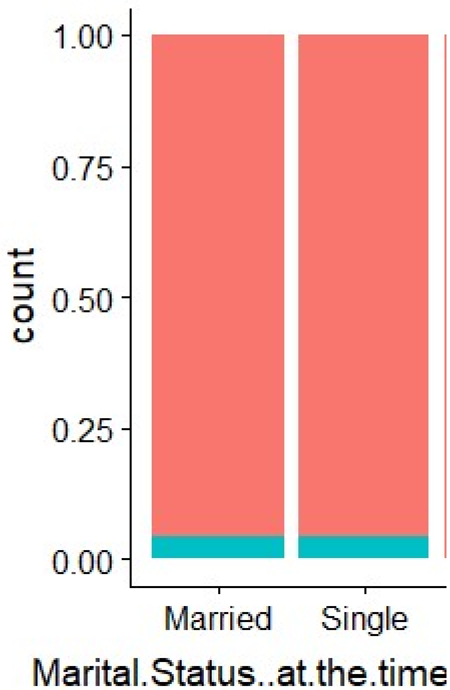


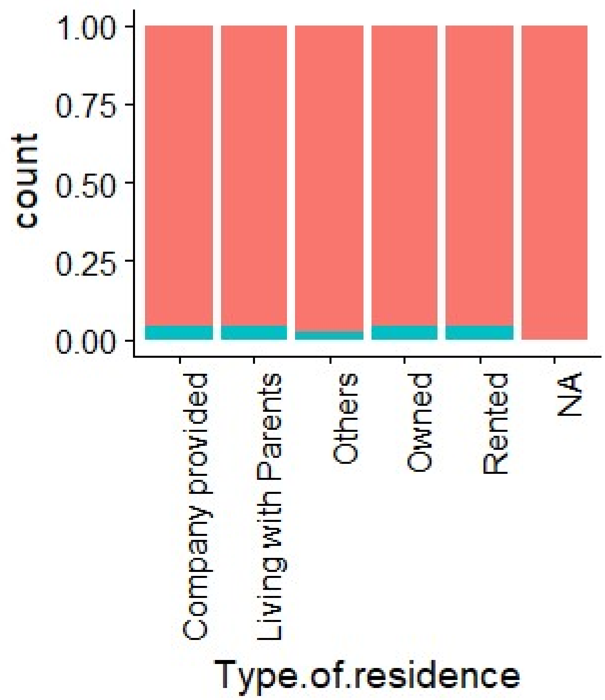
People with higher residence tenure default more

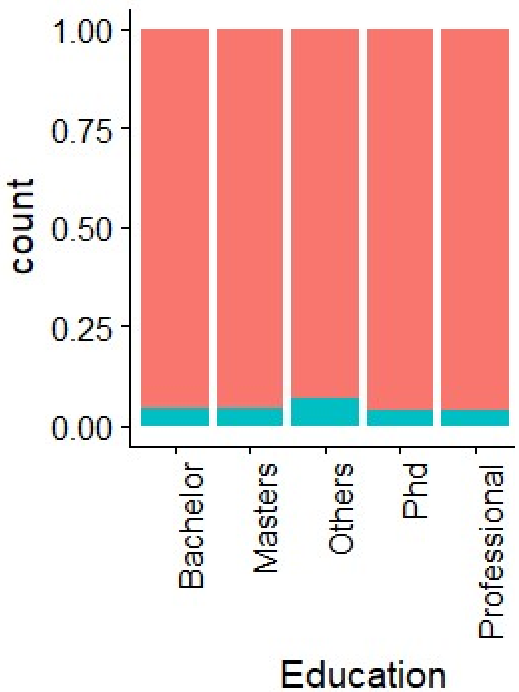


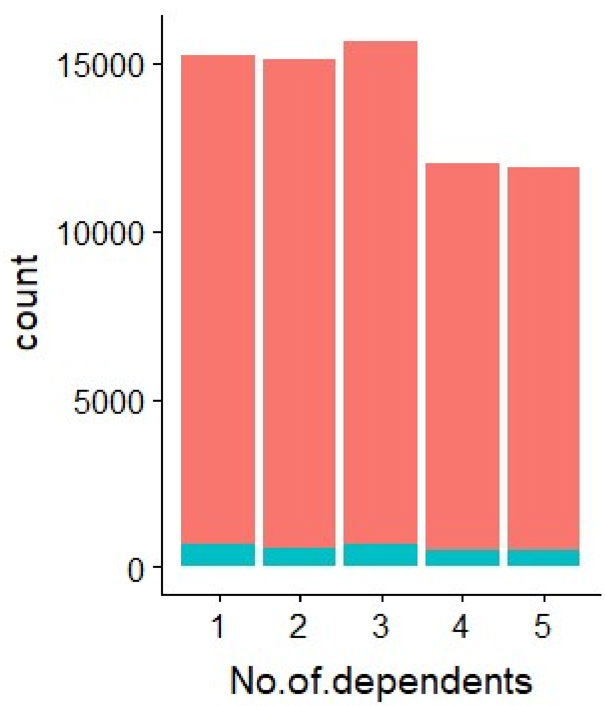
Tenure in a company doesn't play a role in default rate

EDA-BIVARIATE ANALYSIS DEMOGRAPHIC DATA



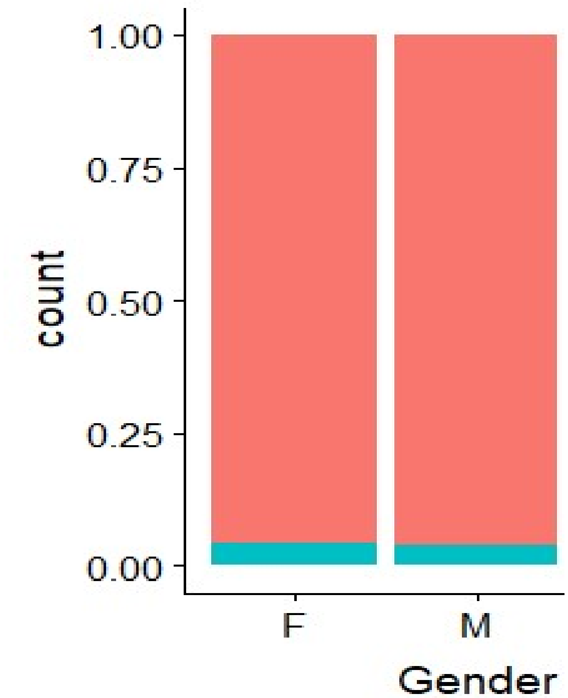


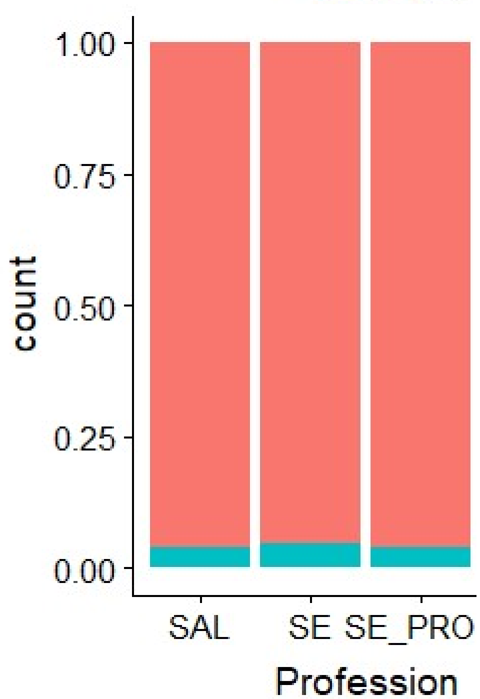




None of the above influence the parameter default rate

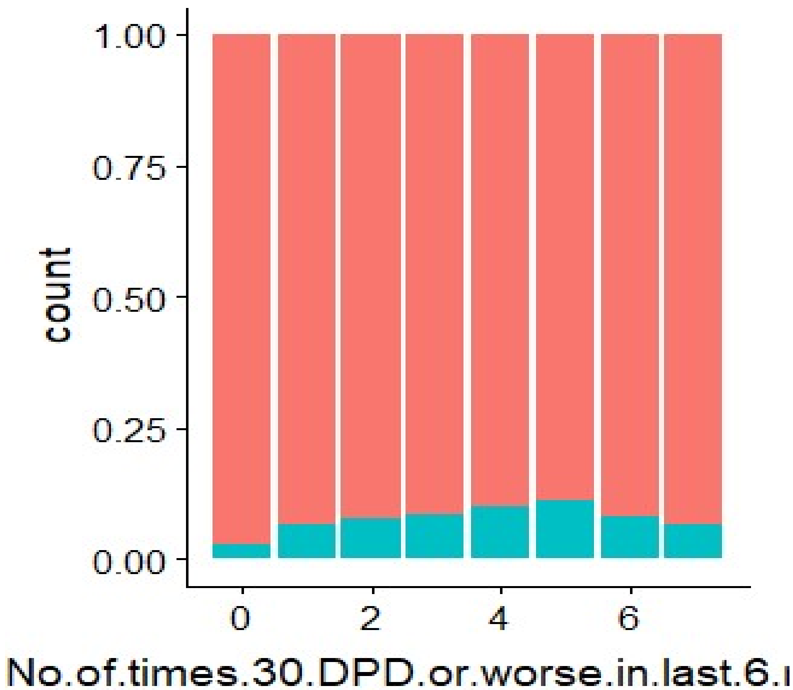
EDA-BIVARIATE ANALYSIS DEMOGRAPHIC DATA

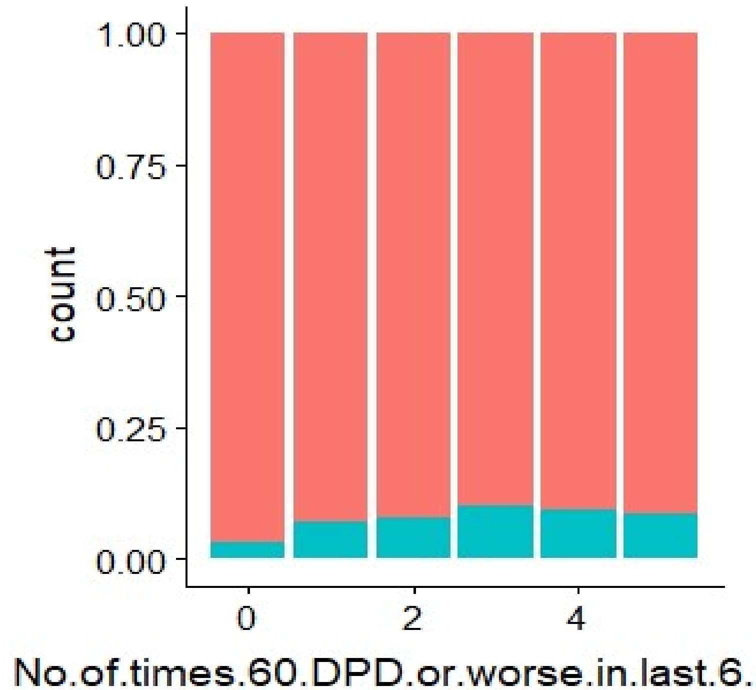


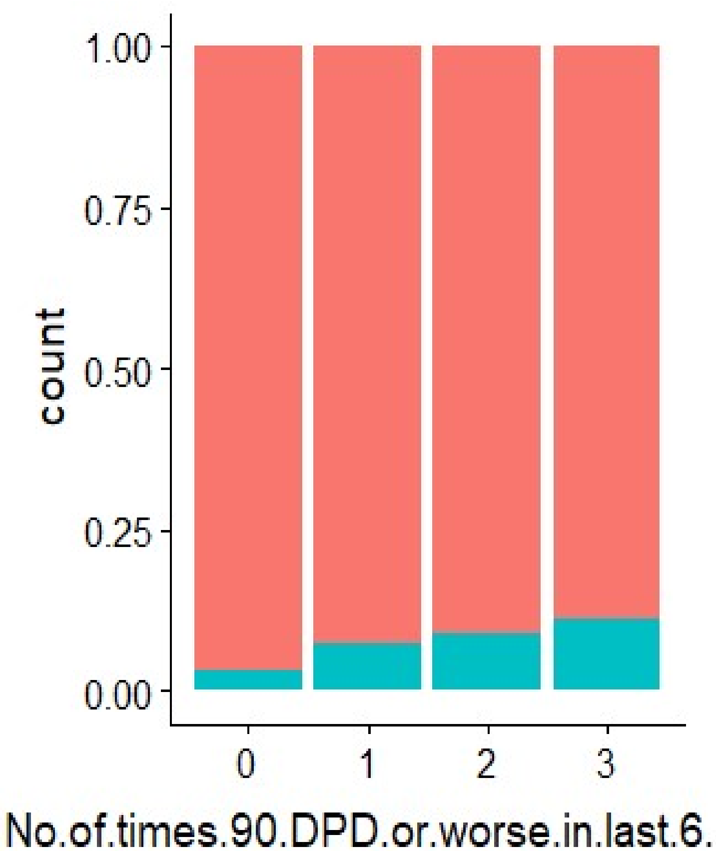


Both Gender and Profession does not affect default rate

EDA-BIVARIATE ANALYSIS CREDIT DATA

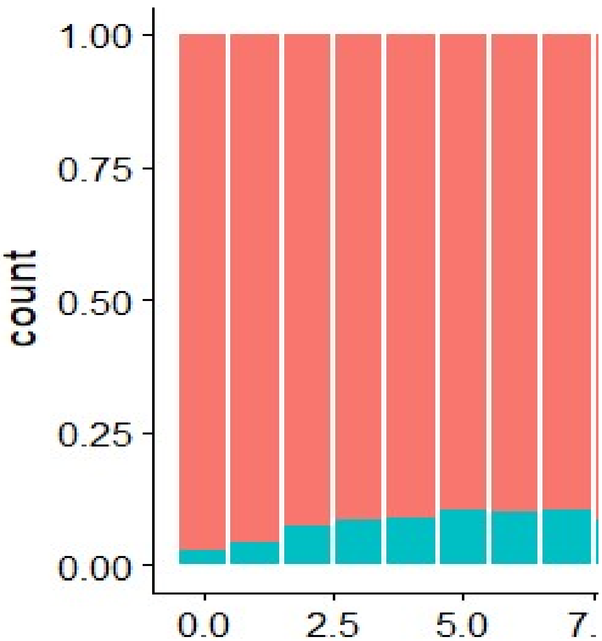




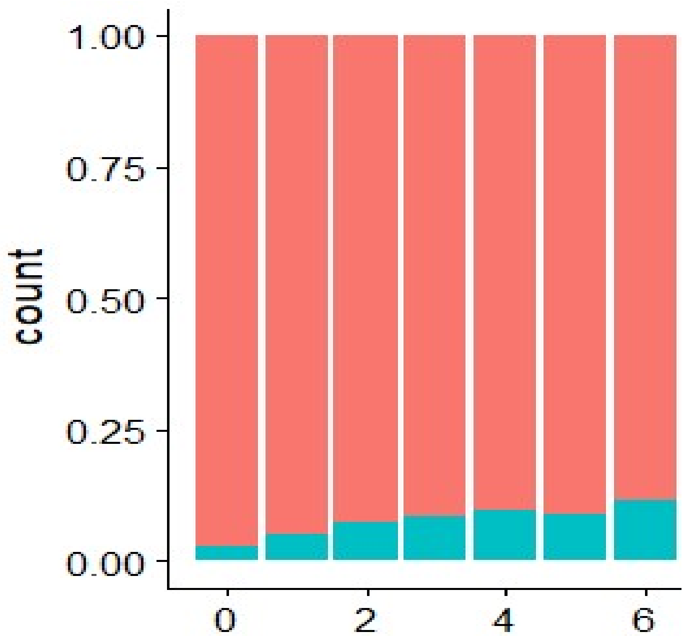


There is a similar trend observed in No of Times 30/60/90 DPD or worse in last six months and hence can be an important predictor.

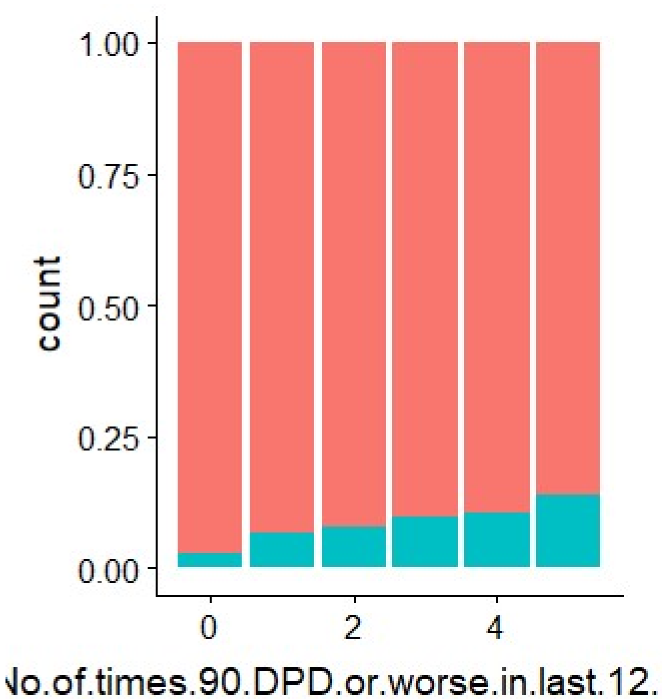
EDA-BIVARIATE ANALYSIS
CREDIT DATA



No of times 30 DPD in last 12 mon

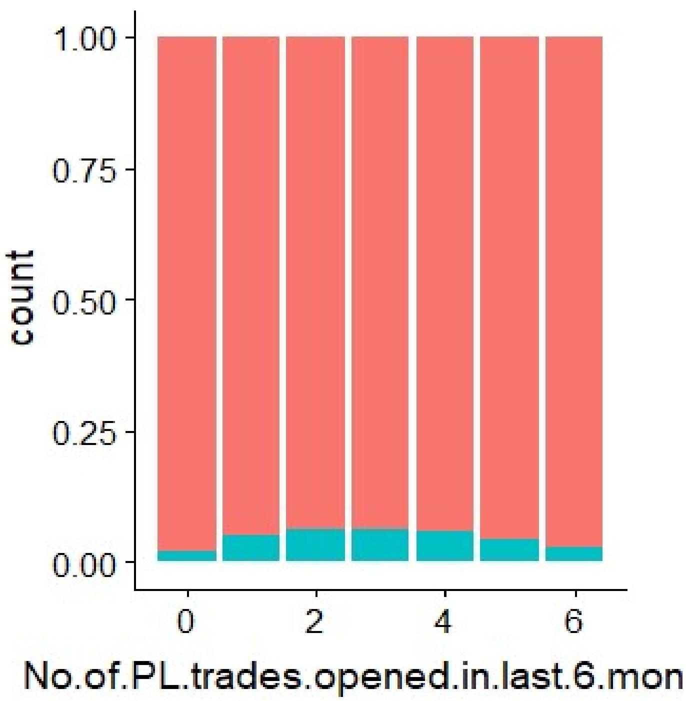


No of times 60 DPD in last 12 mon



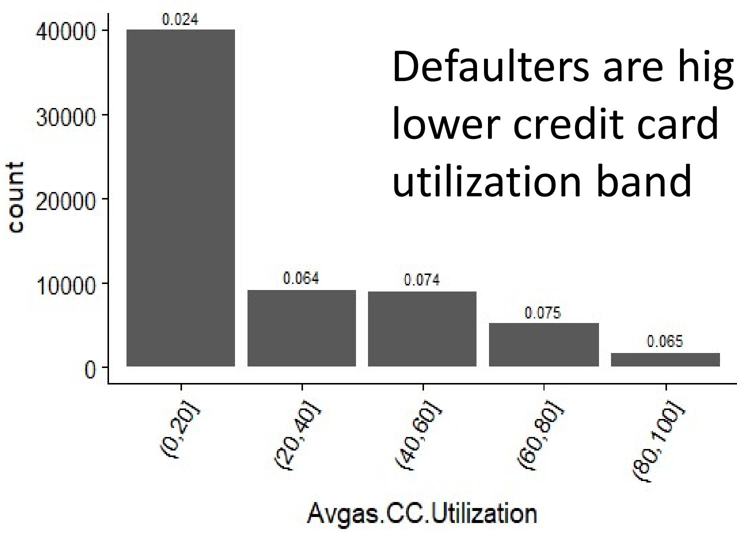
With increase in number of times times 30/60/90 DPD or worse in last 12 mon there is increase in default rate

EDA-BIVARIATE ANALYSIS
CREDIT DATA

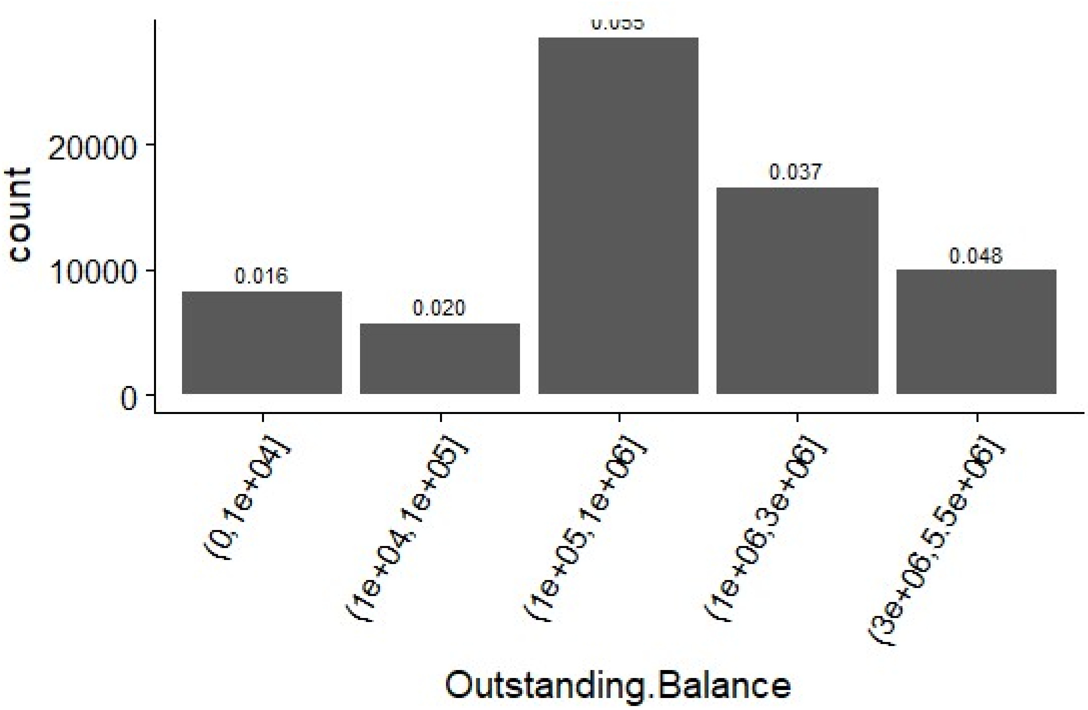


- Percentage of defaulters increases with increase in number of PL trades opened in last 6 months till the 4th month and then decreases.
- Percentage of defaulters is highest amongst applicants who opened 12 PL Trades in last 12 months.

EDA-BIVARIATE ANALYSIS CREDIT DATA

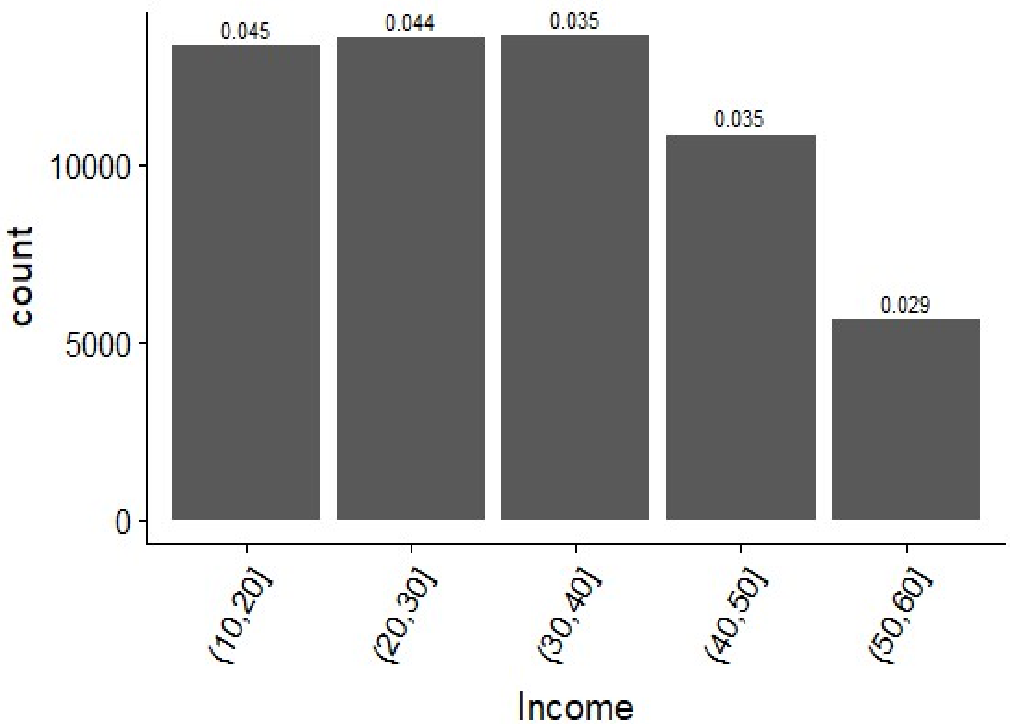


Defaulters are higher in lower credit card utilization band



Defaulters are higher in mid outstanding balance band

EDA-BIVARIATE ANALYSIS
CREDIT DATA



Defaulters reduce at a higher income band of(40-50) and (50-60) but not much difference in trend is observed in lower income bands

Better insights could be obtained from CREDIT DATA in comparison with DEMOGRAPHIC DATA

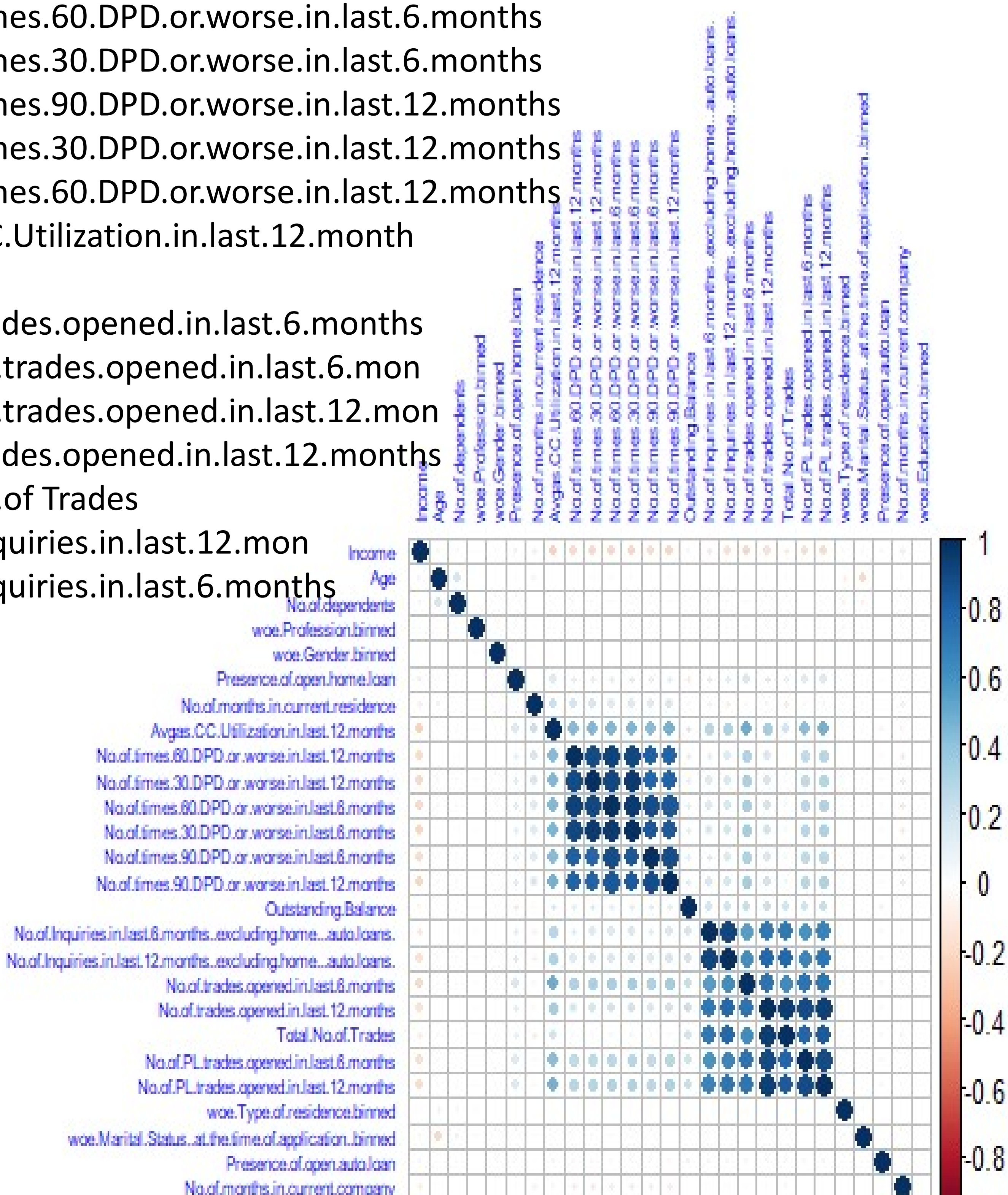
CORRELATION PLOT ON MERGED DATA

Data merged using “Application ID” as common field for join

2 Groups of data having positive correlation with each other..

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.month

No.of.trades.opened.in.last.6.months
No.of.PL.trades.opened.in.last.6.mon
No.of.PL.trades.opened.in.last.12.mon
No.of.trades.opened.in.last.12.months
Total.No.of Trades
No.of.Inquiries.in.last.12.mon
No.of.Inquiries.in.last.6.months



MODEL BUILDING DEMOGRAPHIC DATA

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.

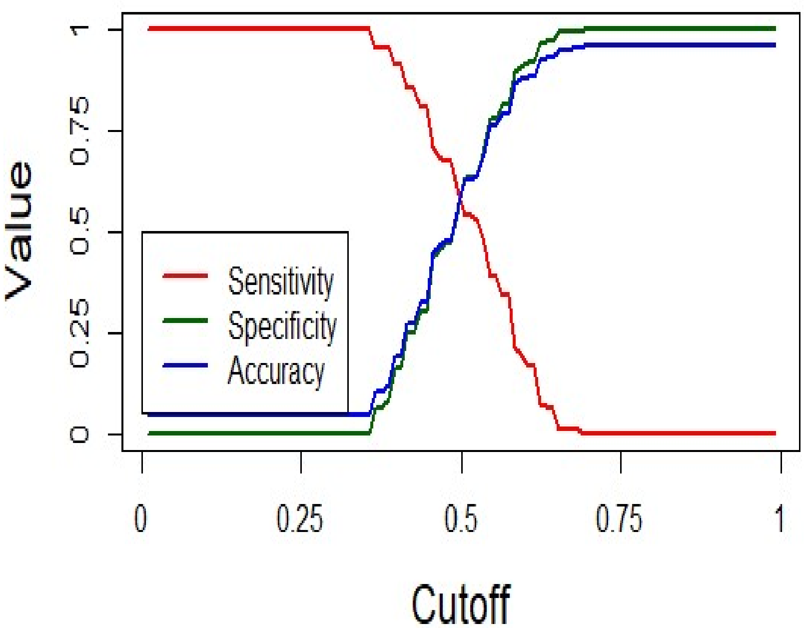
Logistic Regression was built by iteratively removing using these two algorithms

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

LOGISTIC REGRESSION-DEMOGRAPHIC DATA

Important Predictors:

1. WOE.No of months in current residence Binned
2. WOE No of months in current company binned
3. WOE Income Binned



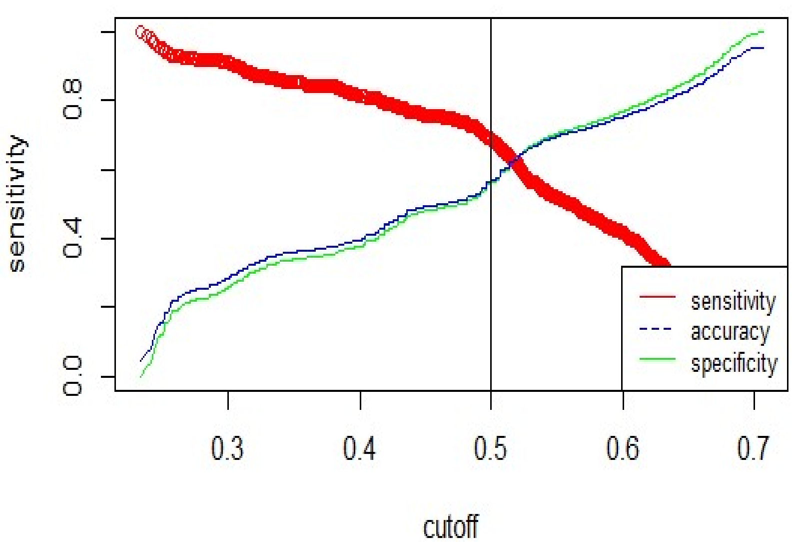
Statistics	Values
Cut-off	0.495
Accuracy	57%
Sensitivity	59%
Specificity	57%

Thus a logistic regression model based only on demographic data seems to have low performance. Hence lets build a model with both demographic and credit data merged.

LOGISTIC REGRESSION-MERGED DATA

Important Predictors:

1. **Income**
2. **No of Months in current company**
3. **No of times 90 DPD or worse in last 12 months**
4. **No of times 60 DPD or worse in last 12 months**
5. **No of times 30 DPD or worse in last 12 months**
6. **Avg CC Untilization in last 12 months**
7. **No of trades opened in last 6 months**
8. **No of PL trades opened in last 12 months**
9. **No of Inquiries in last 12 months exl home and auto loans**



Confusion Matrix:

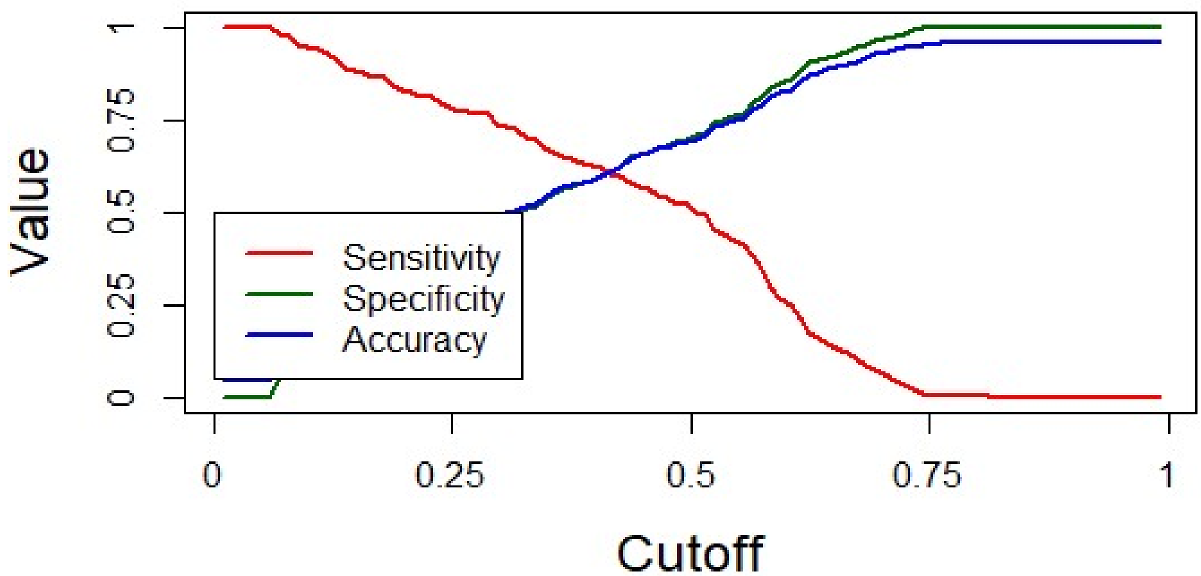
Prediction	No	Yes
No	12671	336
Yes	7404	548

KS-Statistic-25%

Statistics	Values
Cut-off	0.519
Accuracy	62%
Sensitivity	62%
Specificity	63%

RANDOMN FOREST MODEL FOR MERGED DATA

Random Forest model is also applied to check if it performs better in comparison with logistic regression model



Statistics	Values
Cut-off	0.
Accuracy	62%
Sensitivity	62%
Specificity	61%

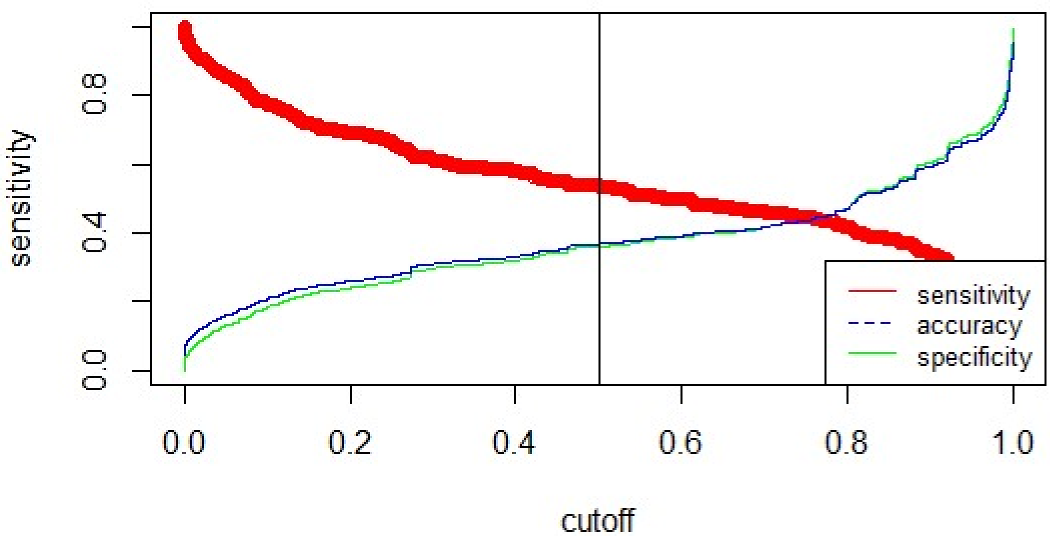
Confusion Matrix

Prediction	No	Yes
No	12522	353
Yes	7553	531

Since performance of the model is not as expected, let us try neural network.

NEURAL NETWORK

Neural network is applied to check if it performs better in comparison with logistic regression model



Accuracy 43%

CONCLUSION:
Logistic Regression model on merged data is the final model for Application Scorecard.

MODEL EVALUATION USING REJECTED DATA

- Merge rejected data(those without Performance Tag) from DEMOGRAPHIC &CREDIT Data using Application ID

- 35 NA's Observed inAvgas.CC.Utilization.in.last.12.months-Excluded

- With a cutoff of 0.52 for the logistic regression model, below is the predictions:
  NO: 3
  YES: 1387

- CONCLUSION: Model accuracy is over 99% on the rejected data

APPLICATION SCORECARD

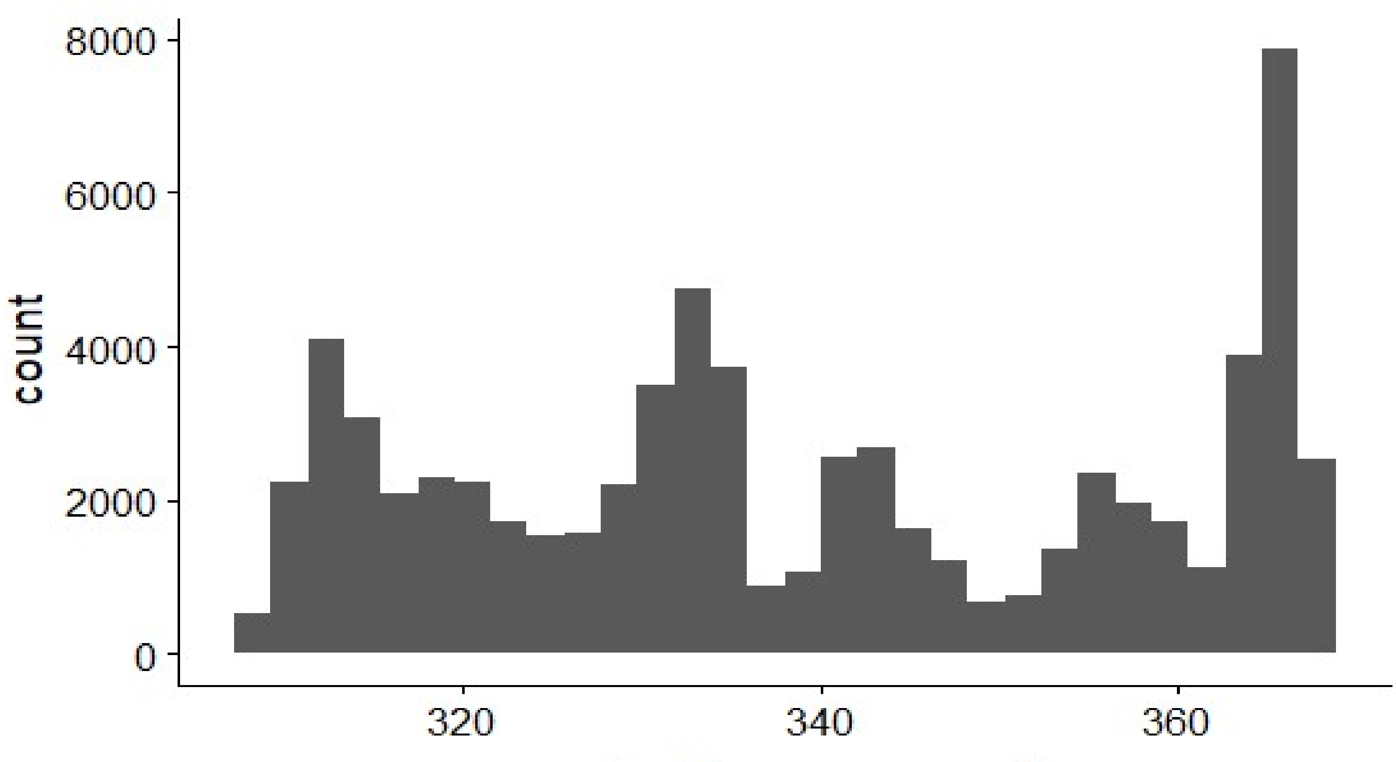
- Final application scorecard was made using the Logistic regression model on the entire merged dataset.
- The logistic regression model was chosen since its evaluation metrics were better in comparison with the other models.
- Probability of default for all applicants were calculated
- Odds for good was calculated. Since the probability computed is for rejection (bad customers),
  Odd(good) = (1-P(bad))/P(bad)
  ln(odd(good)) was calculated
- 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))

APPLICATION SCORE CARD VALUES

- Scores range from 308.3 to 367.9
- Mean score of approved customers is 339.3
- High score means increased risk of defaulting

CUTOFF SCORE FOR ACCEPTING OR REJECTING AN APPLICATION

- Cutoff for final Logistic Regression model 0.52
- CUTOFF_SCORE 331.25
- Number of Applicants above 331.25(ACCEPTED)—43295
- Number of Applicants below 331.25(REJECTED)—26569



CALCULATING BANK PROFIT

Confusion matrix can be used to predict the bank's gain on using the model vs when no model is used

Prediction	No	Yes
No	42216	1079
Yes	24701	1868

- ✓ Applicants with scorecard greater than 331.25 are filtered
- ✓ Profit of #2338483691 units to the bank is calculated as profit based on the outstanding balance of non-defaulters using the model