BFS Capstone Project

CredX Case study

Group Name:

1. Anusha Kulkarni
2. Husn Ara

BUSINESS UNDERSTANDING

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.

Objective:

To identify the right customers using predictive models by determining the factors affecting credit risk and creating strategies to mitigate them.

Solution Approach:

This is a binary supervised classification problem. We aim at building models such as Logistic regression, Random forest etc. 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

- Two datasets are provided, demographic data and credit bureau data.
- **Demographic/application data**: This dataset contains the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- **Credit Bureau data:** 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 CLEANING AND PREPARATION

- **DATA QUALITY ISSUES:**
  - The 1425 rows with no performance tag indicates that the applicant is not given credit card, hence they are removed and stored in another dataset for future use.
  - 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, the 65 records with age <18 has been excluded from the dataset.
  - The above rejected records have been saved separately and would be used for scorecard verification and not for EDA/ modelling.

DATA CLEANING AND PREPARATION

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 (INFORMATION VALUE):

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers. **"Bad Customers"** refers to the customers who defaulted on a loan. and **"Good Customers"** refers to the customers who paid back loan.

Steps of Calculating WOE:

- For a continuous variable, split data into 10 parts (or lesser depending on the distribution).
- Calculate the number of events and non-events in each group (bin)
- Calculate the % of events and % of non-events in each group.
- Calculate WOE by taking natural log of division of % of non-events and % of events

Benefits of WOE:

- It can treat outliers. These values would be grouped to a class Later, instead of using the raw values, we would be using WOE scores of each classes.
- It can handle missing values as missing values can be binned separately.
- Since WOE Transformation handles categorical variable so there is no need for dummy variables.
- WoE transformation helps you to build strict linear relationship with log odds. Otherwise it is not easy to accomplish linear relationship using other transformation methods such as log, square-root etc

WOE AND IV (INFORMATION VALUE):

What is Information Value (IV)?

Information value is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance

Rules related to Information Value

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

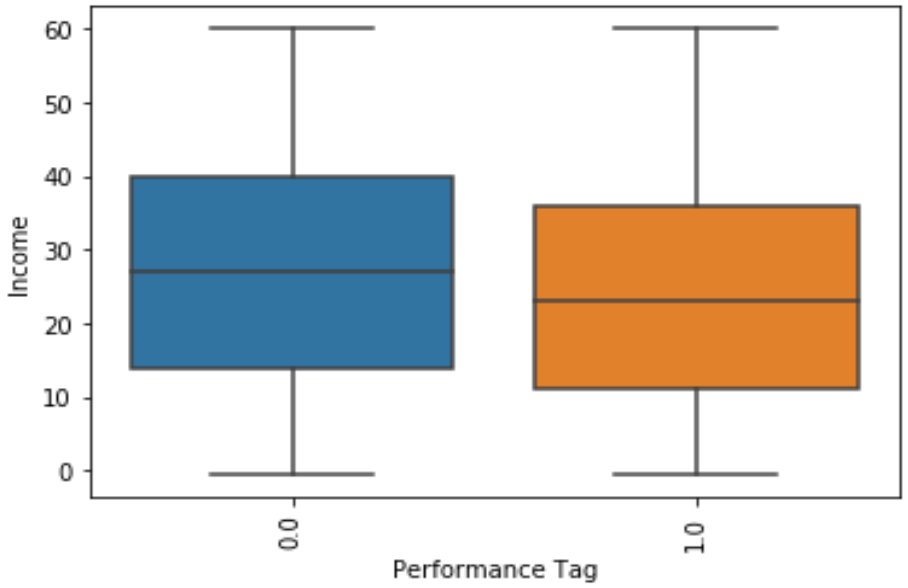
WOE AND IV ANALYSIS

- WOE and IV values are calculated for each of the attributes. For the above 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.
- Following are the Top 9 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.

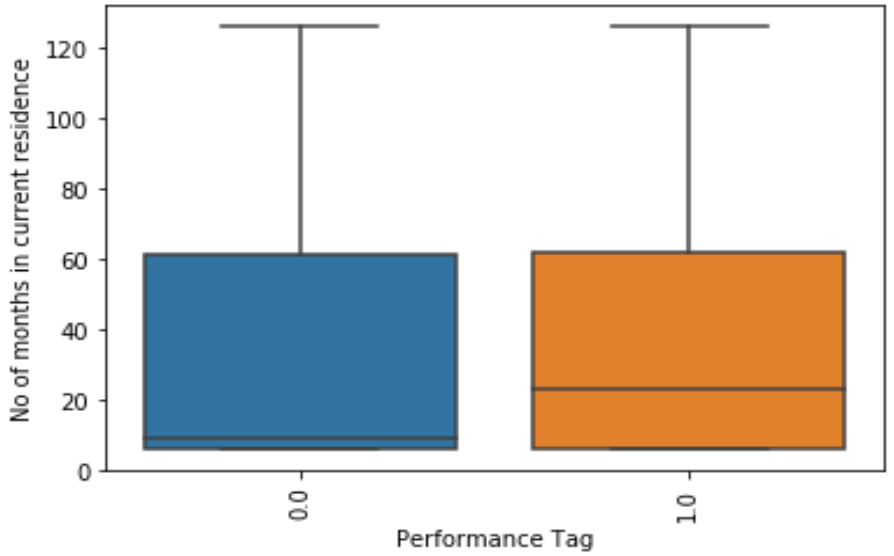
Variable	IV
Avgas CC Utilization in last 12 months	0.293831
No of trades opened in last 12 months	0.257429
No of Inquiries in last 12 months (excluding h..	0.229218
Total No of Trades	0.189907
No of times 30 DPD or worse in last 12 months	0.188045
No of PL trades opened in last 12 months	0.176644
No of times 30 DPD or worse in last 6 months	0.145708
No of times 60 DPD or worse in last 12 months	0.137676
No of PL trades opened in last 6 months	0.124744

EXPLORATORY DATA ANALYSIS

Both Univarate and Bivariate analysis is performed on all the variables of the dataset. Variables of credit bureau dataset showed better insights than demographic variables.



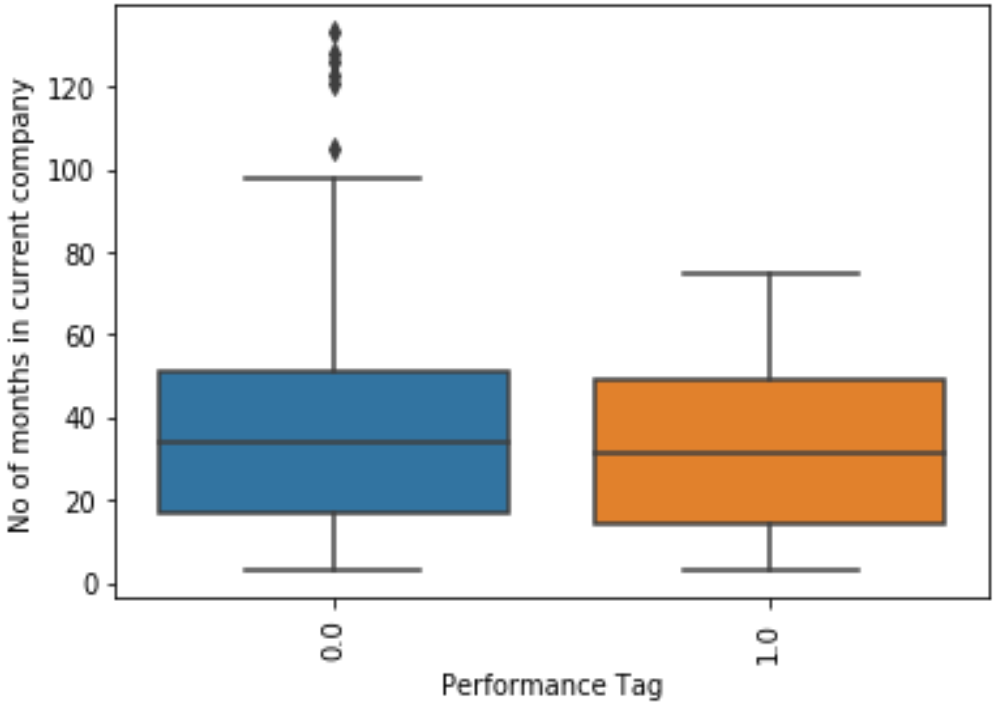
The median values for income of defaulters are lower than that of non-defaulters.



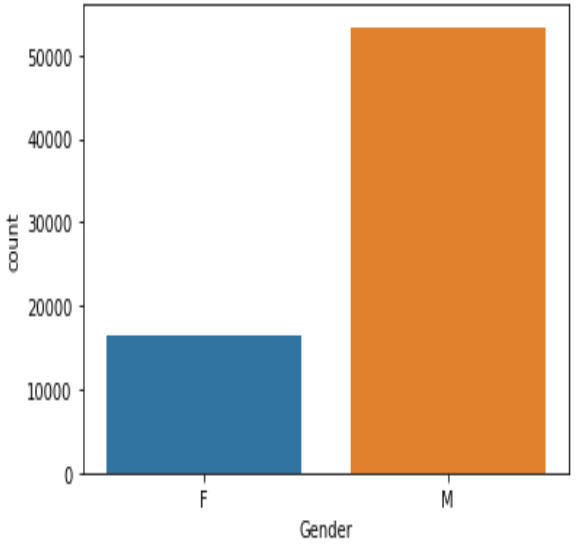
The media value of No.of.months.in.current.residence of non-defaulters are lower than that of defaulters.

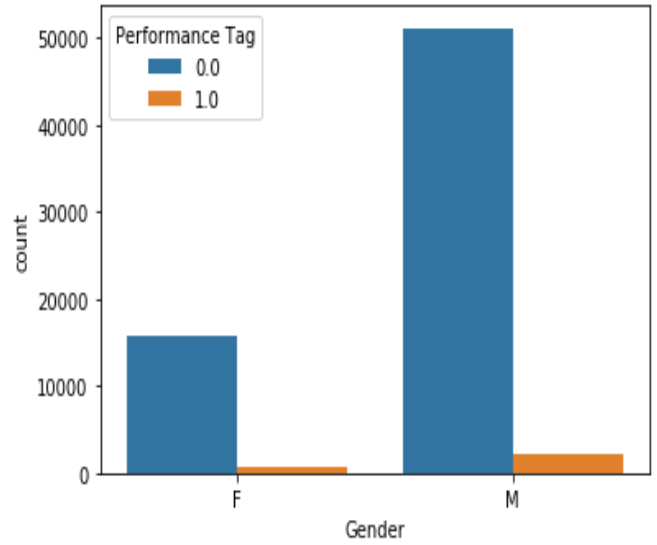
EXPLORATORY DATA ANALYSIS

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



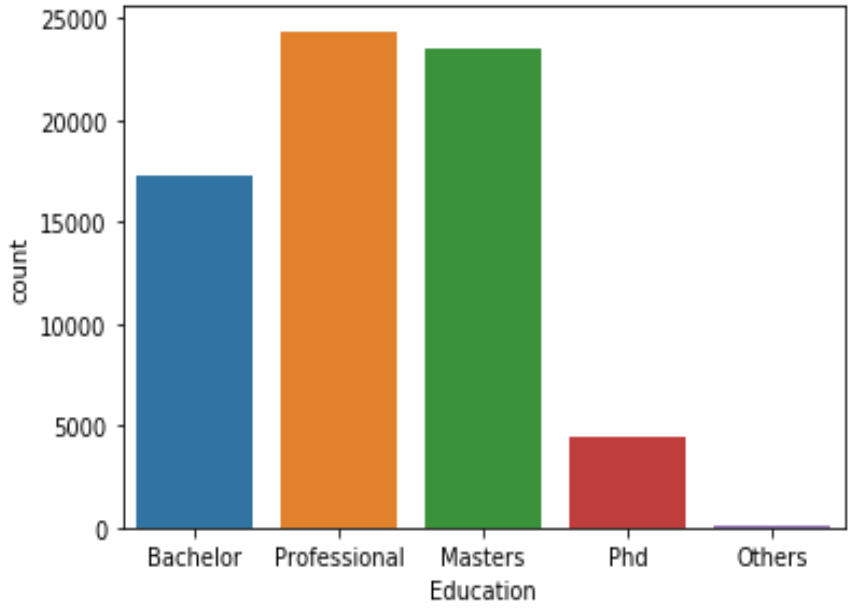
There are more male applicants than female applicants and slightly more default rate of male.



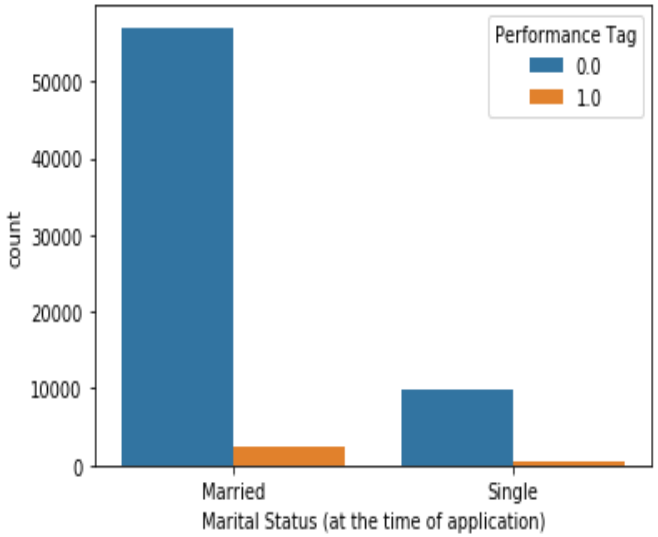


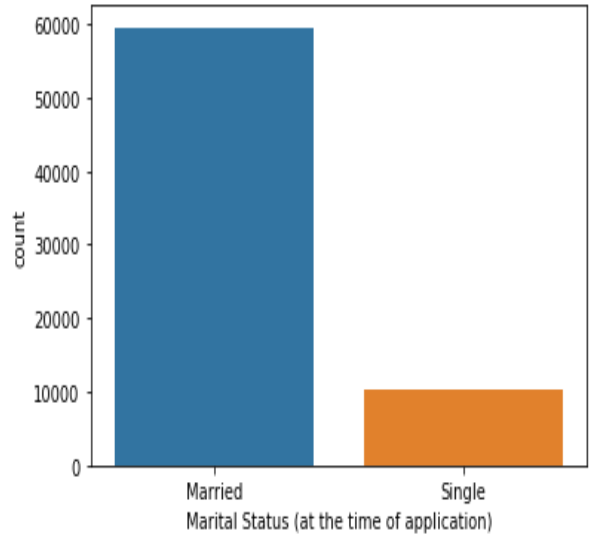
EXPLORATORY DATA ANALYSIS

There are more professional applicants.



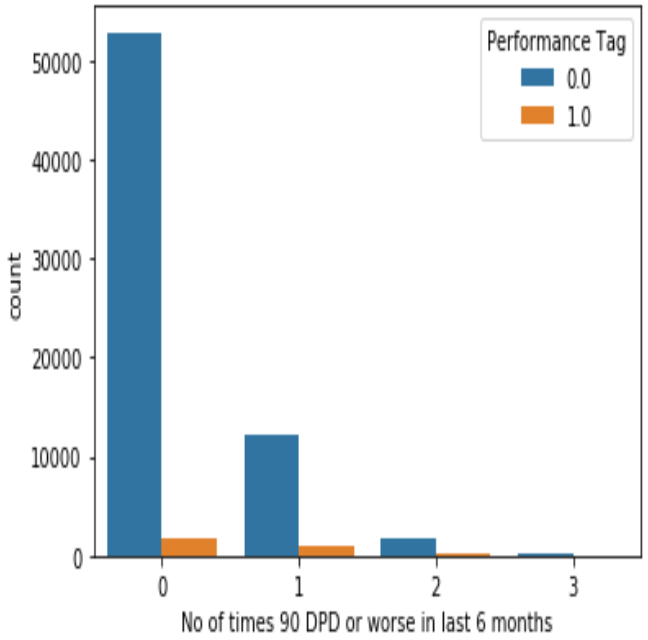
There are more married applicants than single applicants and slightly more default rate of married.

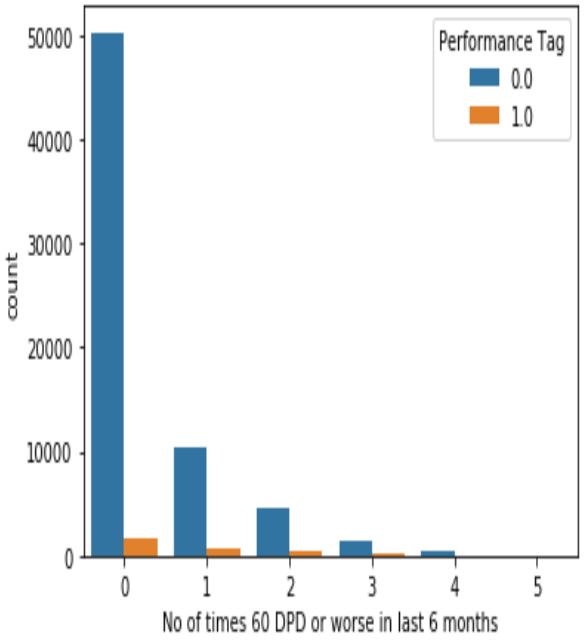


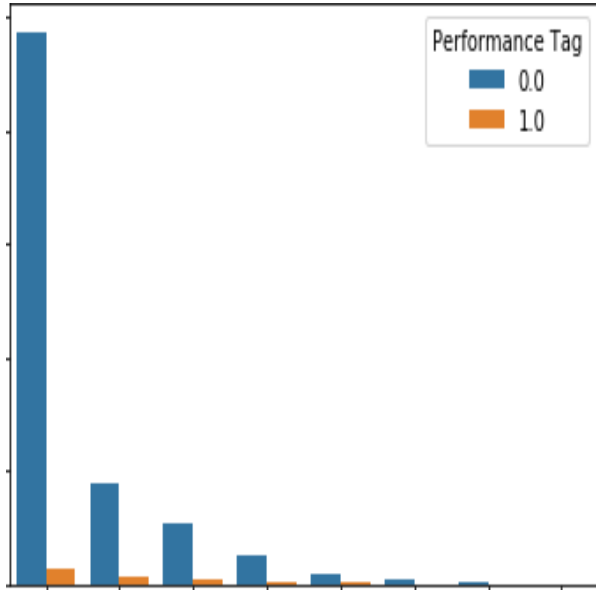


EXPLORATORY DATA ANALYSIS

Number 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

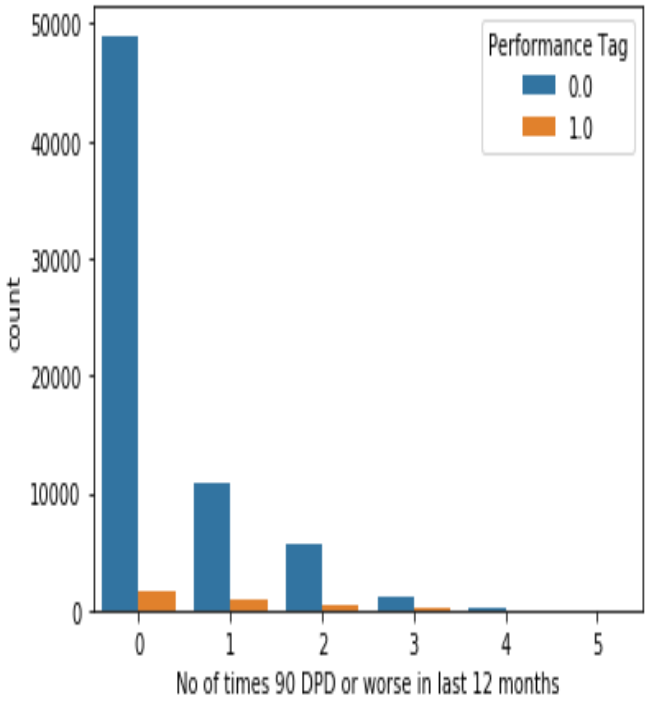


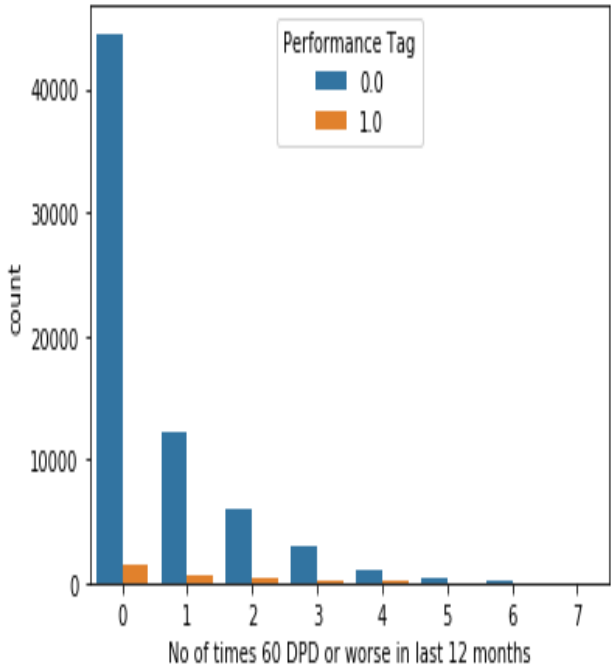


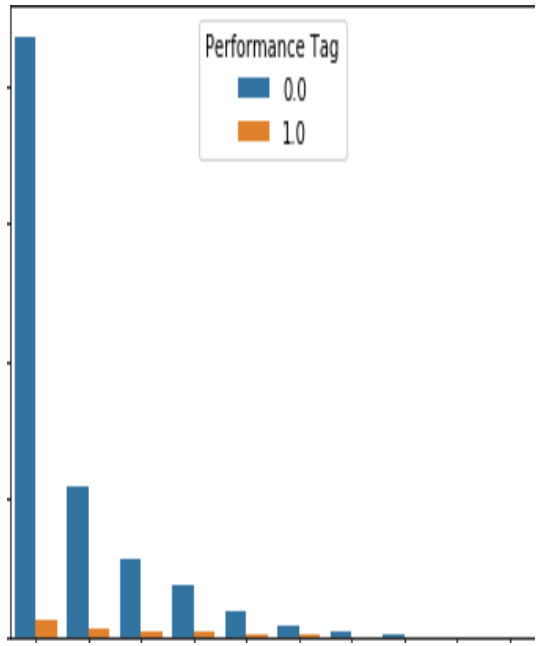


EXPLORATORY DATA ANALYSIS

Number of defaulters are 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.

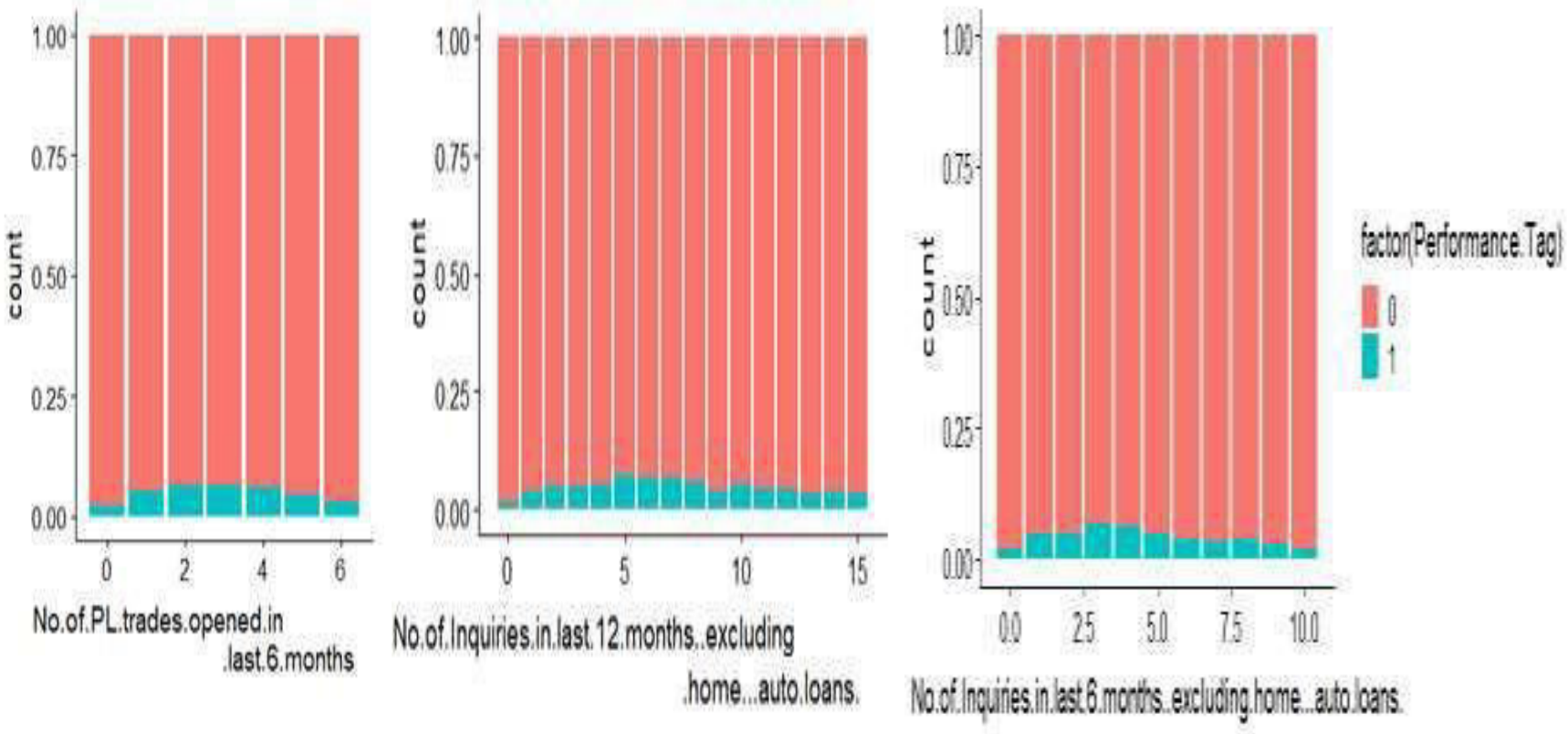






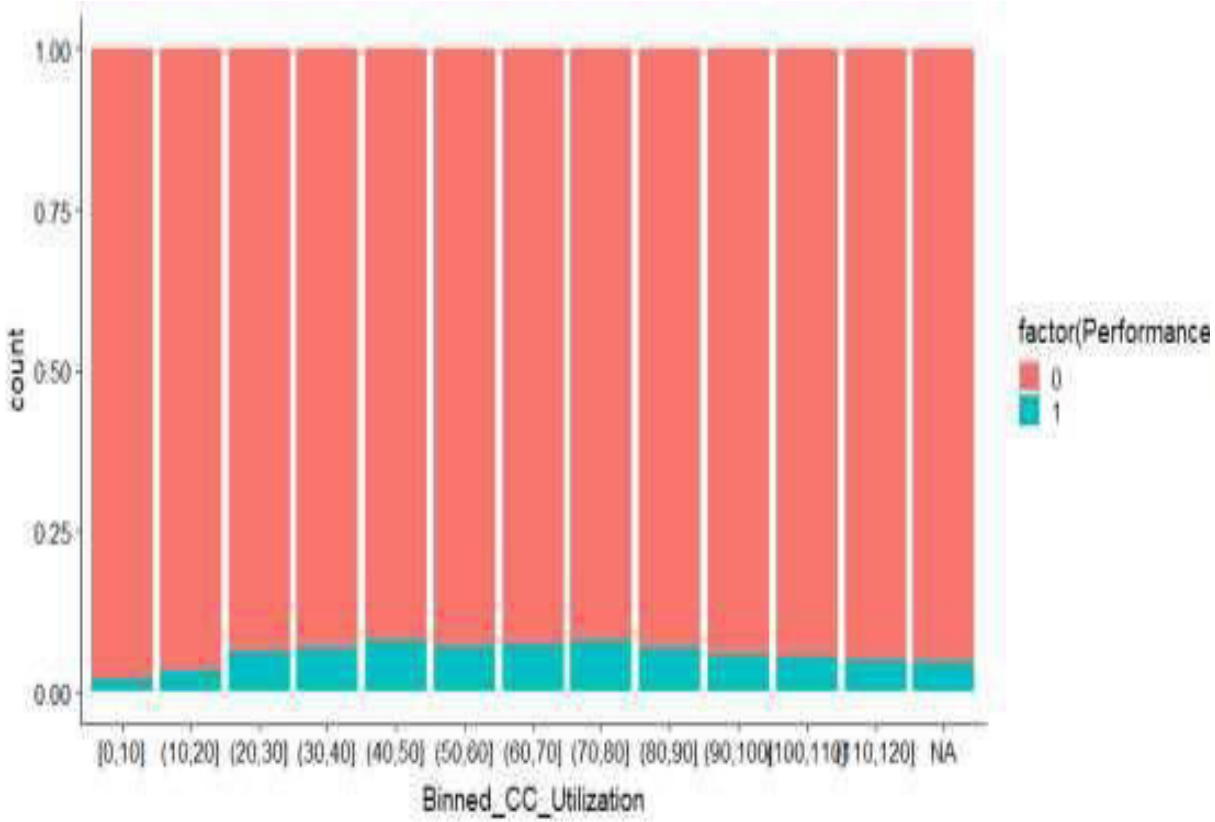
EXPLORATORY DATA ANALYSIS

Number of enquiries and numbers of enquiries fields don't show any pattern

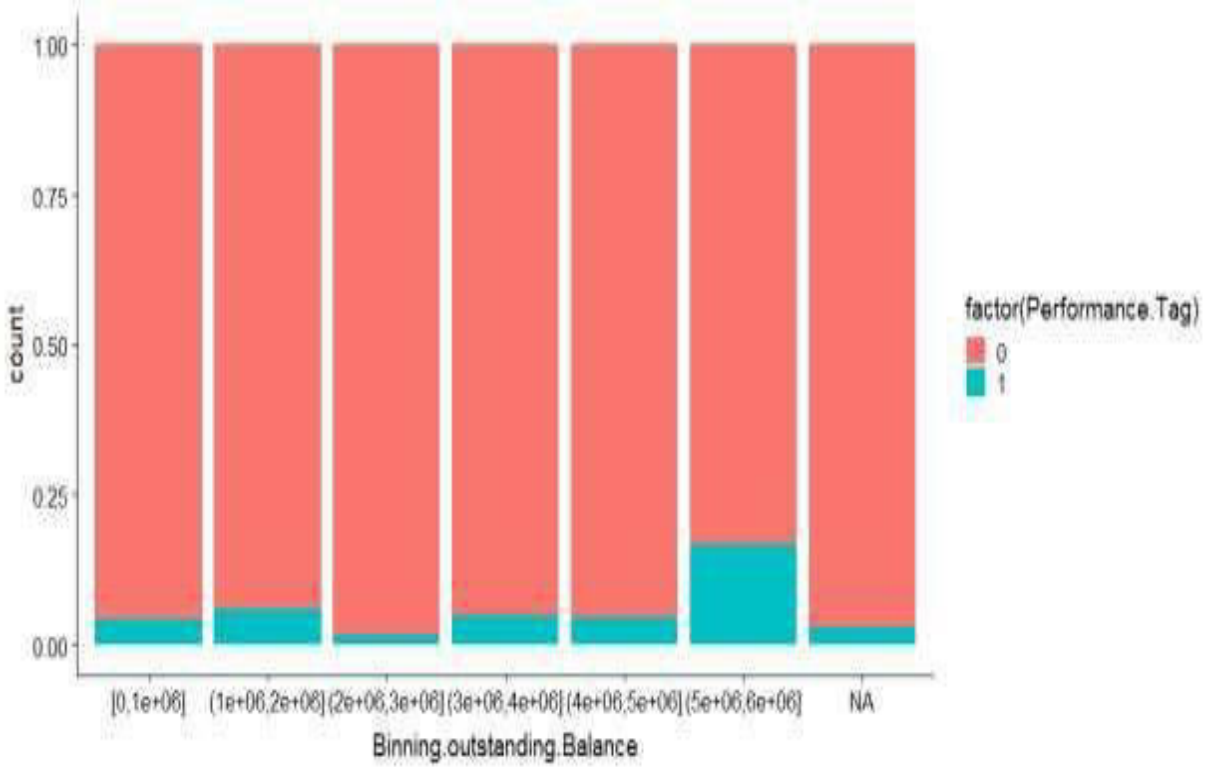


EXPLORATORY DATA ANALYSIS

There is no appropriate pattern found by Avg. Credit cased utilization details.



Outstanding balance field shows increase in defaulter in 50L-60L range. This can be an important predictor.



DATA TRANSFORMATION

- 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 contains 69,799 records and the dataset is split into Train and Test in 70:30 ratio for model building.

- DATA SAMPLING:

The data is highly imbalanced. Only 4.2% of total data is about the defaulters. So we use SMOT sampling method for data sampling

MODEL BUILDING APPROACH

The following models are built on the given datasets to check the performance of the different models.

1.Demographics Dataset

- Logistic Regression Model
- Random Forest model

2.Master Dataset(Demographics + Credit Bureau Dataset)

- Logistic Regression Model
- Random Forest model

There are 1425 records without performance tag. We stored that records in separate data frame(Master_data_rejected) for predicting the performance.
After selecting the final model we will perform prediction on that data.

Assumption:
The 1425 records without performance tag are assumed to be rejected by the bank.

Model Evaluation Results

	Accuracy	Sensitivity	Specificity	Precision	Recall
Demographics Logistic Regression	50%	53%	50%	5%	62%
Demographics Random Forest	92%	9%	93%	5%	4%
Master Data (Demographics + Credit Bureau) Logistic Regression	57%	69%	56%	6%	70%
Master Data (Demographics + Credit Bureau) Random Forest	95%	99%	99%	6%	0%

Model Evaluation Results

OBSERVATIONS:

1. Logistic regression model based only on demographic data seems to have low performance.
2. The overall performance is very low for models made from demographic data alone.
3. Around 60% accuracy, sensitivity and specificity was achieved with a logistic regression model on combined demographic and Credit bureau data i.e. Master data.
4. Random Forest Performed low on combined demographic and Credit bureau data i.e. Master data.

MODEL SELECTION:

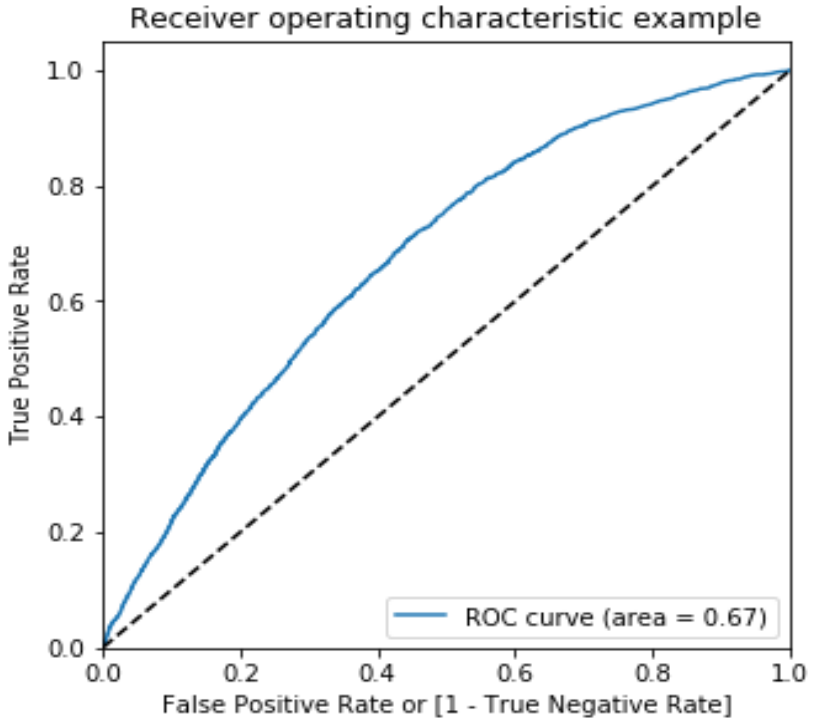
Logistic Regression model with the combined demographic and Credit bureau data i.e. Master data is selected as the best model for the given business problem.

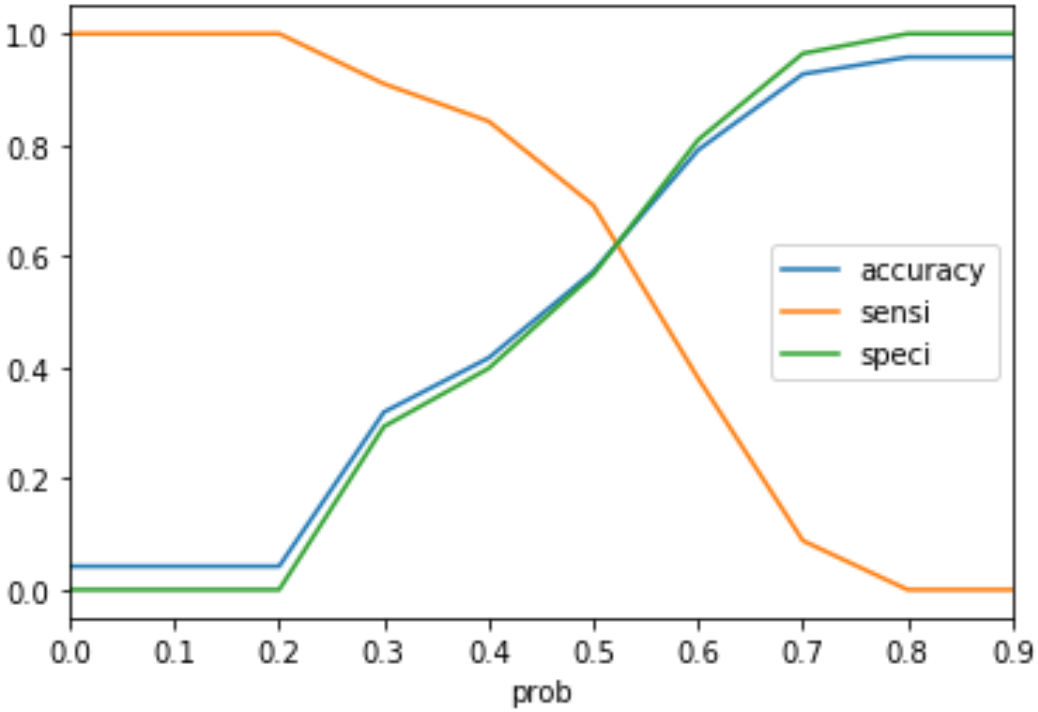
Results of Logistic Regression on Master Data

Accuracy: 57%
Sensitivity: 69%
Specificity: 56%
AUC: 0.67
Cutoff Point: 0.50

Confusion Matrix:

Prediction	0	1
0	26569	20271
1	637	1427





Application ScoreCard

Final application scorecard was made using the selected **Logistic regression** model on Mater Datadataset.

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. Used the following formula for computing application score card:

   Score=offset+factor*log(Odd_good)

   where factor=pts_double_odds/log(2)

   offset=target_score-factor*log(target_odds)

Summary of application_score_card values:

- Scores range from 308.38 to 393.4 for applicants.
- 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 0.50
- CUTOFF_SCORE= 400 + (slope * (log((1-0.50)/0.50) - log(10)))
- CUTOFF SCORE is equal to **333.56**
- No.of applicants above score 333.56 and thus their credit card application will be accepted as per our model is 27202
- No.of applicants below score 333.56 and thus their credit card application will not be accepted as per our model is 21702

Results of Logistic Regression on Records with no performance Tag i.e. Master_Data_rejected

As Assumed Total number of rejected customers by bank : 1425
After Applying Logistic Regression model on Master_data_Rejected we found that 6 customers out of 1425 should have been given the credit card.
No : 1419
Yes : 6
Credit Score Cut off : 333.56

Financial Benefit Of The Model

Assumptions:

- Let us assume bank makes Rs.5000 per year from 1 credit card customer.
- Considering an average loss of Rs.5000 when each non defaulters application is rejected
- and an average loss of Rs.1,00,000 when each accepted applicant defaults

Revenue Loss For Bank:

Revenue loss occurs when good customers are identified as bad and their credit card application is rejected.

Bank refused 6 potential credit card customer, amounting to Rs.30,000 annual loss to the bank.

No of candidates rejected by the model who didn't default - 637
Total No of candidates who didn't default - 26569+637 = 27206
Revenue_loss = (637/(27206))*100

2.34% of the non defaulting customers are rejected which resulted in revenue loss.

Financial Benefit Of The Model

Net financial gain due to using our model

Net Profit when no model is considered:
Total number of applicants who are non defaulters:66917
Total no of applicants who are defaulters:2947
Profit_without_model=(66917*5000)-(2947*100000)
So net profit without model is 39885000

Net Profit when model is considered:
Using the confusion matrix given from logistic regression model:
Profit_with_model=(26569*5000)+(1427*100000)-(20271*5000)-(637*100000)
So net profit with model is 110490000

Prediction	0	1
0	26569	20271
1	637	1427

Net financial gain due to using our model: 70605000
% financial gain due to using our model:177.02

Financial Benefit Of The Model

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

Credit loss without model:
Credit Loss without model is: 0.91

Credit loss with model:
Credit loss with model is :0.41

Credit loss saved with model: 0.91-0.41=0.50

Conclusion

- Logistic regression model is chosen as the final Model with 57% of Accuracy.
- Optimal score cut-off value of 333.56 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 to reject the candidate who may default in future.
- There is Net Financial gain of 177.02% after using the model.