Case Study: BFS Capstone Project – Mid Submission

Group members:

1. Piyush Gaur
2. Priya Gupta
3. Ria Nag
4. Sahana K

BUSINESS UNDERSTANDING

Objective:

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

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.

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.

1.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.

2. 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 CLEANSING 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.
- 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.

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. Continuous quantitative variables for which WOE values were not monotonically changing across bins, were made by the Information package in R by default, coarser bins were made by decreasing the number of bins until monotonic behavior is observed across bins. For the above 9 variables with Missing values, the variable values were replaced by their corresponding WOE values.
- Since Information package treats 1 as 'good', we created a new variable - Reverse.Performance.Tag with inversed relationship for IV analysis.
- 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.

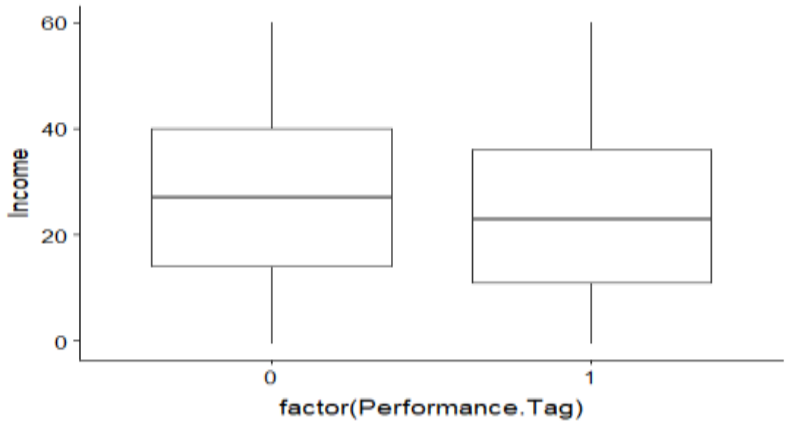
Variable	IV
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.271544682
Avgas.CC.Utilization.in.last.12.months	0.260755415
No.of.times.30.DPD.or.worse.in.last.6.months	0.241562739
No.of.times.90.DPD.or.worse.in.last.12.months	0.213874838
No.of.times.60.DPD.or.worse.in.last.6.months	0.205833876
No.of.times.30.DPD.or.worse.in.last.12.months	0.198254858
No.of.trades.opened.in.last.12.months	0.194337383
No.of.times.60.DPD.or.worse.in.last.12.months	0.185498873
Total.No.of.Trades	0.182235069
No.of.PL.trades.opened.in.last.12.months	0.176644264
No.of.times.90.DPD.or.worse.in.last.6.months	0.160116924
No.of.PL.trades.opened.in.last.6.months	0.124743691
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.092939144
No.of.months.in.current.residence	0.078943527
Income	0.0424178
No.of.months.in.current.company	0.021754413
Presence.of.open.home.loan	0.017626529
Outstanding.Balance	0.014239503
Age	0.003349157
woe.Profession.binned	0.002182094
Presence.of.open.auto.loan	0.00165482
Application.ID	0.001504195
woe.Gender.binned	0.000324971
woe.Type.of.residence.binned	0.000289274
woe.Education.binned	0.000269428
woe.Marital.Status..at.the.time.of.application..binned	9.52E-05

EXPLORATORY DATA ANALYSIS

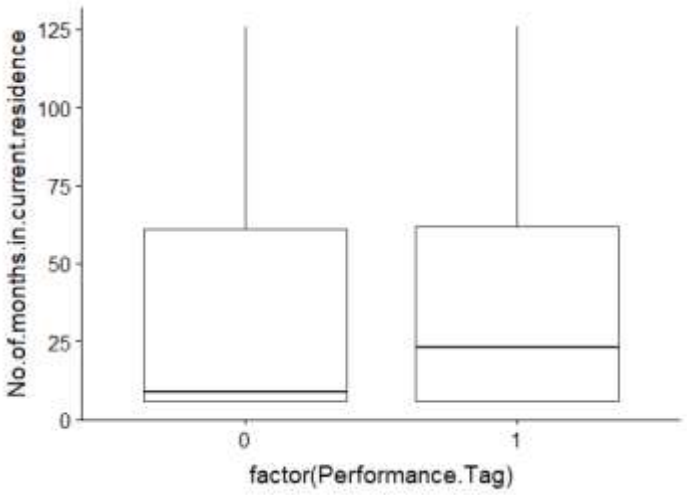
Both Univariate and Bivariate analysis is performed on all the variables of the dataset.

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

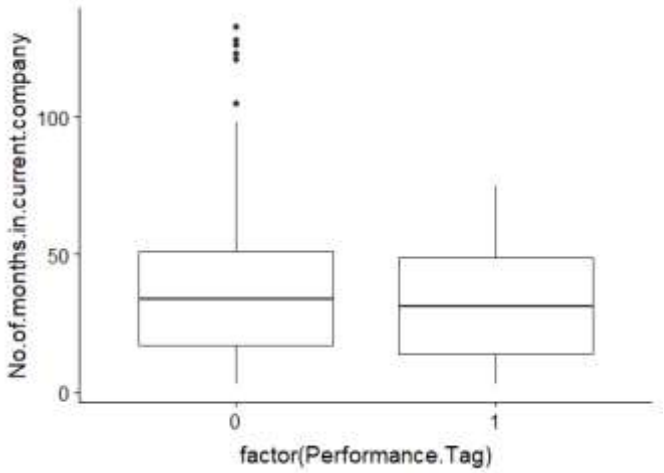
- The median values for income of defaulters are lower than that of non-defaulters.



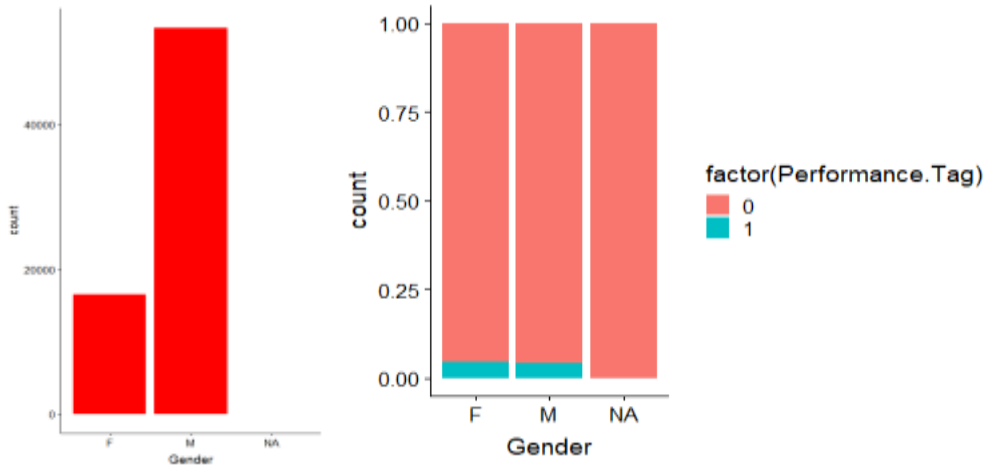
- The median No.of.months.in.current.residence of non-defaulters are lower than that of defaulters.



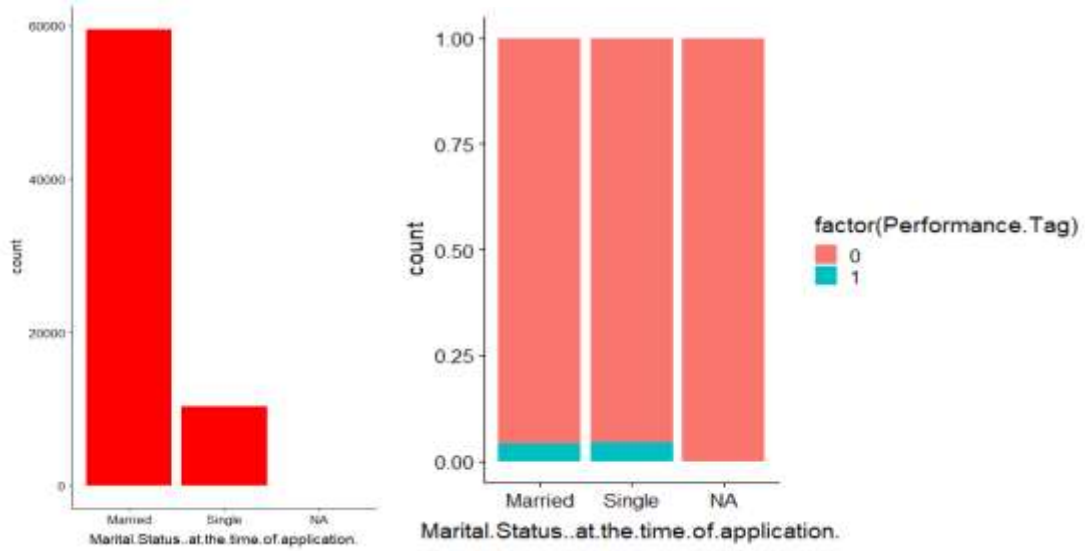
- The median No.of.months.in.current.Company of non-defaulters is slightly lower than that of defaulters.



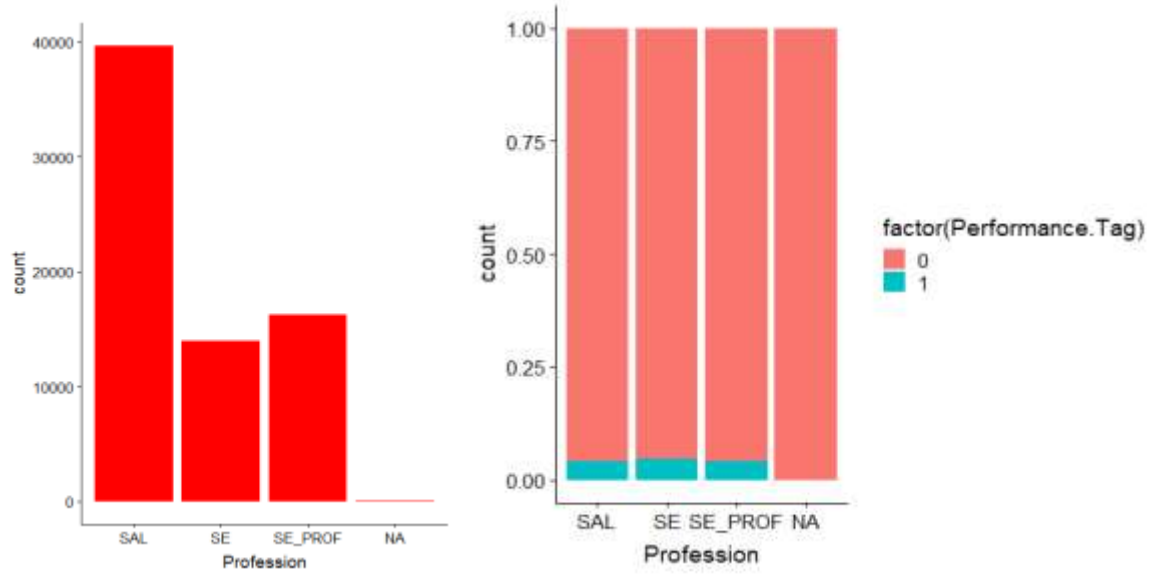
- There are more male applicants than female applicants but there is no difference in default rates.



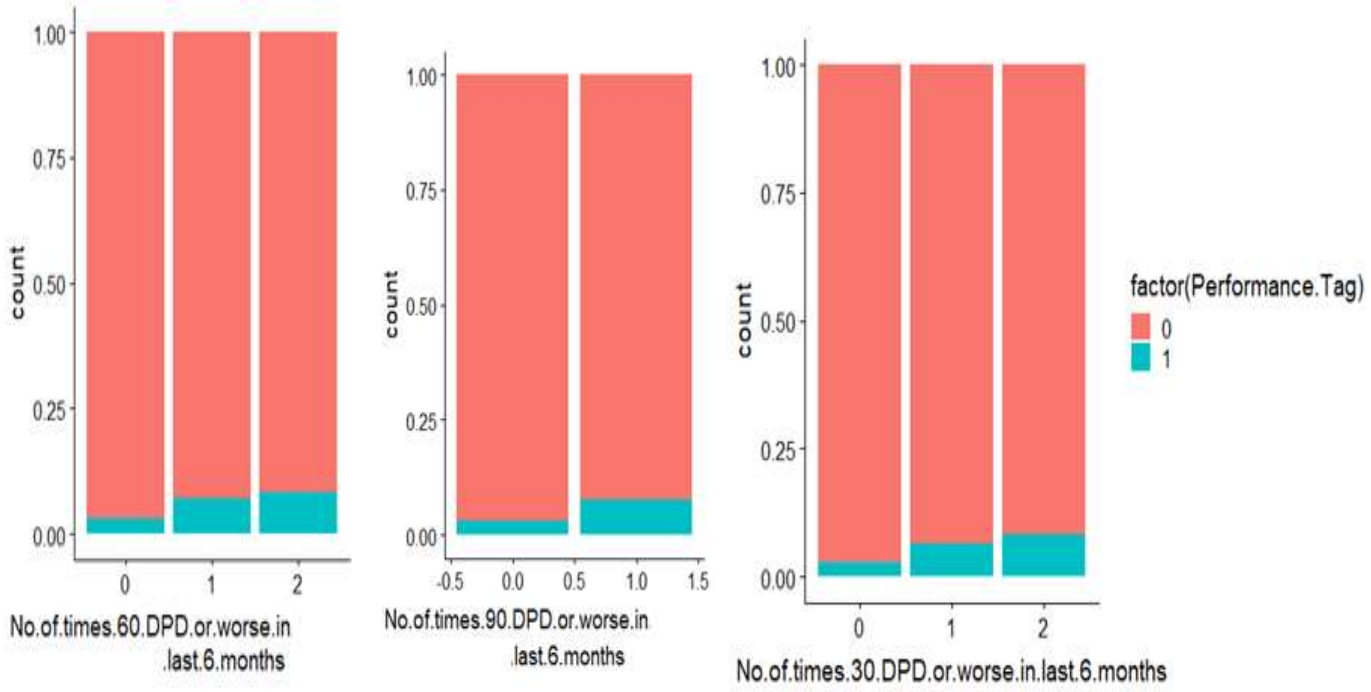
- There are more married applicants than single applicants but there is no difference in default rates.



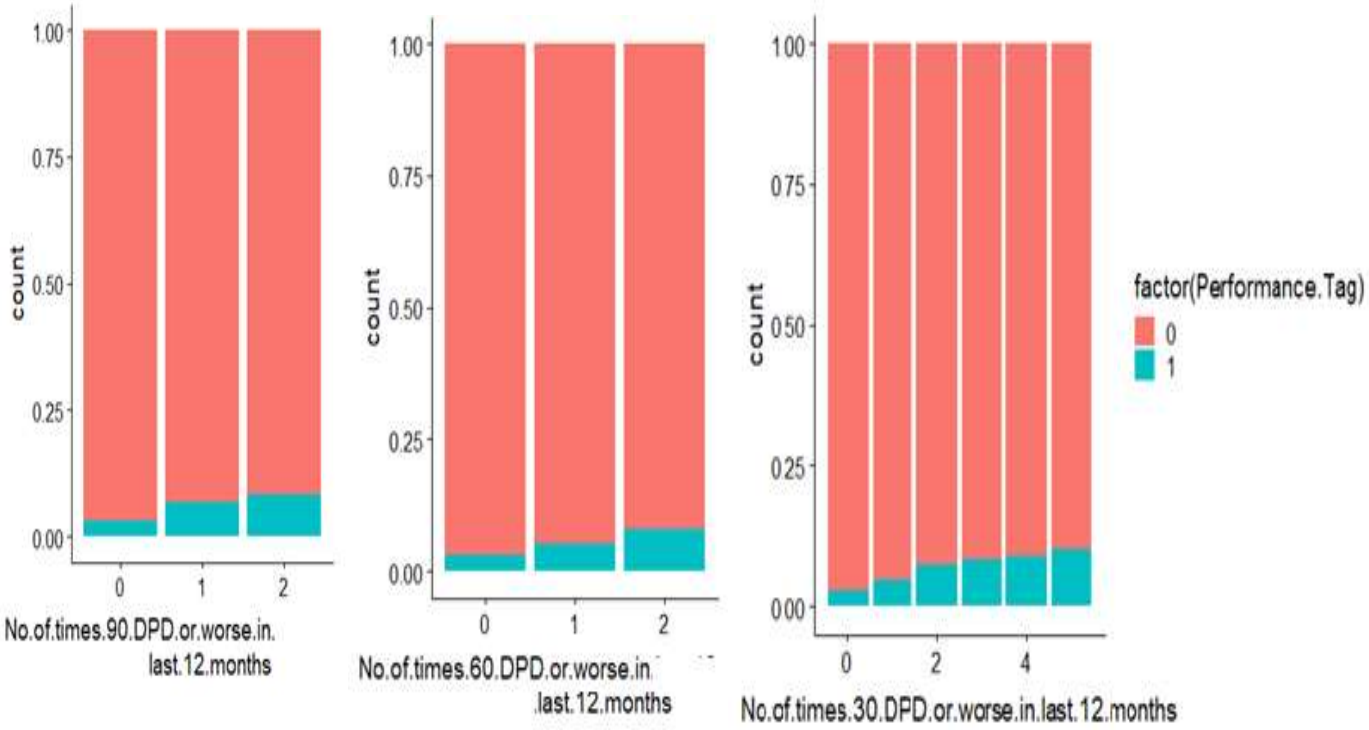
- There are more SAL profession applicants but there is no difference in default rates.



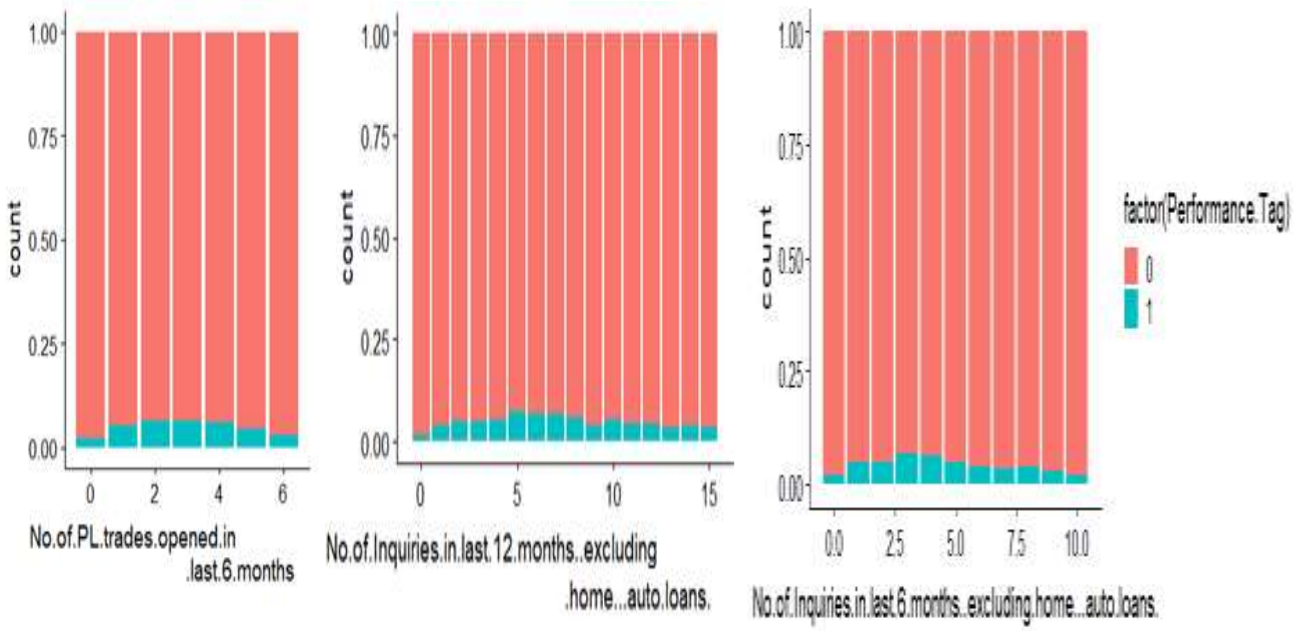
- Number of defaulters are increasing with increase in Number of 30/60/90 DPD or worse in last 6 months variable values. Hence these variables can be important predictors.



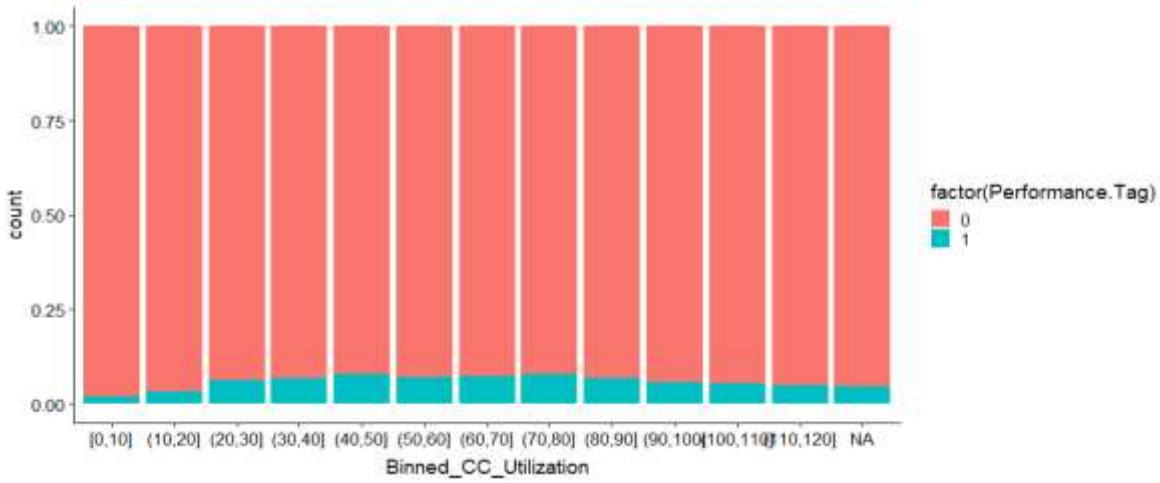
- Number of defaulters are increasing with increase in Number of 30/60/90 DPD or worse in last 12 months variable values. Hence these variables can be important predictors.



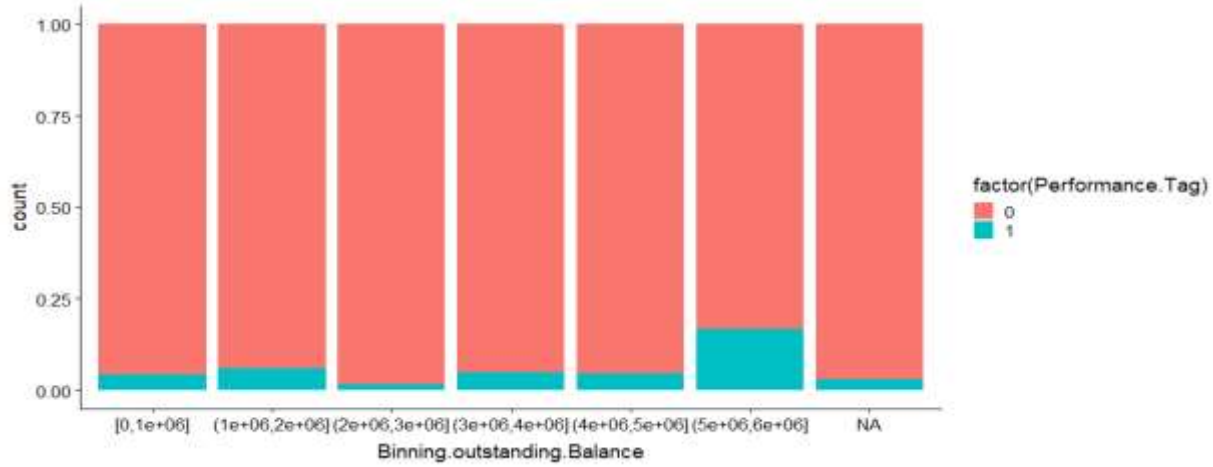
- Number of enquiries and numbers of enquiries fields don't show any pattern.



- 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. We have used ROSE package for balancing our data sets. It helps to generate artificial data based on sampling methods and smoothed bootstrap approach.

Model Building and Evaluation

1. **Demographic data model:**
   - Applied Logistic Regression and Random Forest to iteratively build model by removing insignificant variables and variables exhibiting high levels of multicollinearity.
   - In Logistic Regression, we eliminated insignificant variable through drill down approach.
   - In Random Forest, we varied the hyper parameters and created the final model.
   - We have evaluated the model based on accuracy, Sensitivity and Specificity values.
   - Logistic regression model performed better compared to Random forest, but the overall performance of model seemed low.

Logistic Regression	
Metrics	**Values**
Cut-off	0.5
Overall Accuracy	55%
Sensitivity	55%
Specificity	55%

Random Forest	
Metrics	**Values**
Kappa	0.0132
Overall Accuracy	54%
Sensitivity	54%
Specificity	53%

2. **Combined data model:**

   Applied Logistic regression as initial model for the combined data.

   Evaluated the model using confusion matrix and KS-Statistics.

Logistic Regression	
Metrics	**Values**
KS-statistic	0.28 and lies in 5th Decile
Overall Accuracy	63%
Sensitivity	63%
Specificity	63%
Area under the curve (AUC)	0.67
Cut-Off	0.51

APPLICATION SCORE CARD

- The application score for each applicant calculated using the logistic regression model, ranges from **297.6 to 362.3**. Score increases by 20 points for doubling odds for good customers. Application score for odds of 10 to 1 is 400. We are yet to finalize the final Cut-off score. Higher the scores indicate lesser risk for defaulting.

Method used for computation of application scorecard:

a) Computed the probabilities of default for the entire population of applicants using the model.
b) Computed the odds for the good. Since the probability computed is for rejection (bad customers), Odd(good) = (1-P(bad))/P(bad)
c) Used the following formula for computing application score
Application score = 400 + slope * (ln(odd(good)) - ln(10)) , where slope is 20/(ln(20)-ln(10))

RoadMap for final submission

- We have plans to build more classification models like random forest, xgboost, svm on the final merged data to finalize the model with better discriminative power. We also propose to use class weights to balance the two imbalanced classes and see if it improves the discriminative power of the models.
- We propose to evaluate the model using different techniques like Confusion matrix, K-fold cross validation techniques, KS-Statistics, and based on that we would decide on the best model for our case.
- We will re-build the application scorecard on the final model to find the cut-off score and predict the potential financial benefits for the company.
- Predict the likelihood of default for the rejected candidates using the model.