BFSI CAPSTONE PROJECT FOR CREDX

|  |  |
| --- | --- |
| Image result for credX  BFSI CAPSTONE PROJECT FOR CREDX  : EDA ANALYSIS (MID-SUBMISSION) | ABSTRACT  Helping CredX identify the right customers using predictive models  COURSE  PGDDS from IIITB-UPGRAD  TEAM MEMBER  Deepak Kumar Chaubey(deepakchaubey07@gmail.com)  Maruthi Rao(maruthir123@yahoo.com)  Shanthakumara S N(shanthakumara11rvce@gmail.com)  Vivek Gusain(viveckgusain@gmail.com) |

Case Study: BFS Capstone Project

Business Understanding

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 objective of this study is to find the right customer to reduce the credit loss.

Data Understanding

The CredX company has provided us two data sets, demographic/application data and credit bureau data.

The demographic data is obtained from the information provided by the applicants at the time of credit card application, which includes customer level information on age, gender, income, marital status etc.

The credit bureau data contains variables such as number of times 30 dpd or worse in last 3/6/12 months, outstanding balance, number of trades etc.

The demographic data consists of 71295 observations with 12 variables including 1577 NA’s and 3 duplicates application id, the credit bureau data consists of 71295 observations with 19 variables including 3028 NA’s and 3 duplicates application id.

Data Cleaning

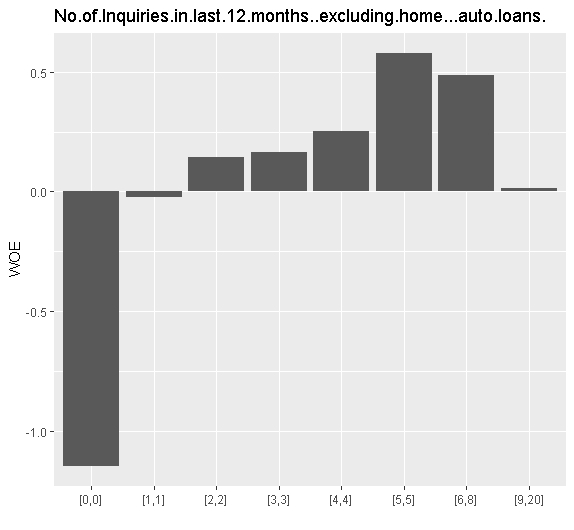
We noticed that there are 1425 NA’s in our dependent variablein our merged file, which indicates whether applicant has default on his or her credit card payment or not. Also, since the data has no information about default, theseapplicants are the rejected applicant (one who have not been given credit card by the company), so we will remove this data in our validation sets.

After removing validation data, we are left with 69867 observation including 1718 NA’s. Since NA’s is only 2.5% of the total observation in our master file. We will drop those NA’s from our file. The Application ID is also removed since it’s an Identity variable which we cannot use in our analysis.

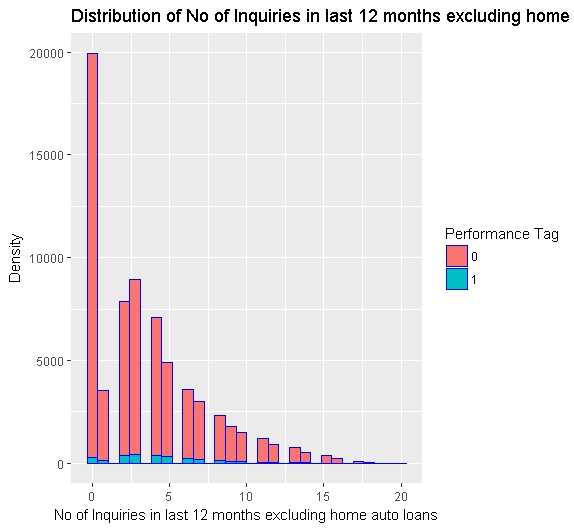
Exploratory Data Analysis

We will be using WOE and IV for EDA. We see that “No of Inquiries in last 12 months excluding home auto loans”, “Avgas CC Utilization in last 12 months”, “No of PL trades opened in last 12 months”, “No of trades opened in last 12 months” has IV more than 0.3 which indicates that these variable have Strong predictive Power where as “Outstanding Balance” and “Total No of Trade” has Medium predictive Power. Now let check the plot of these variables with respect to WOE.

1. **No of Inquiries in last 12 months excluding home auto loans**

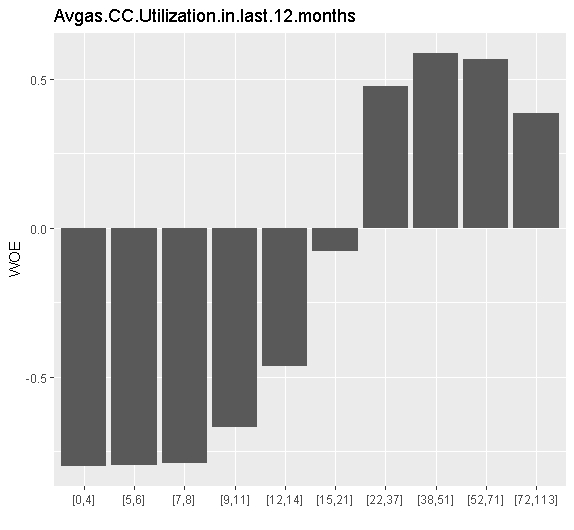
**

* We see that the applicant whose No of Inquiries in last 12 months excluding home auto loans is 0, has the higher chances of default, since their percentage of good customers were very less in comparison to their bad customers.

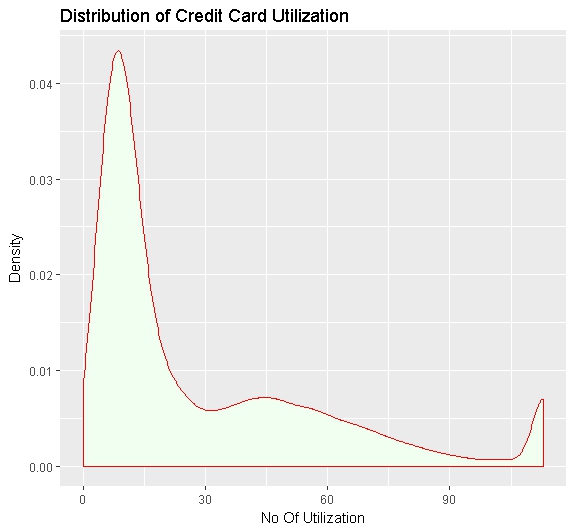


* We see that the distribution is right skewed, so we will use log transformation to make the distribution normal.

1. **Avgas.CC.Utilization.in.last.12.months**

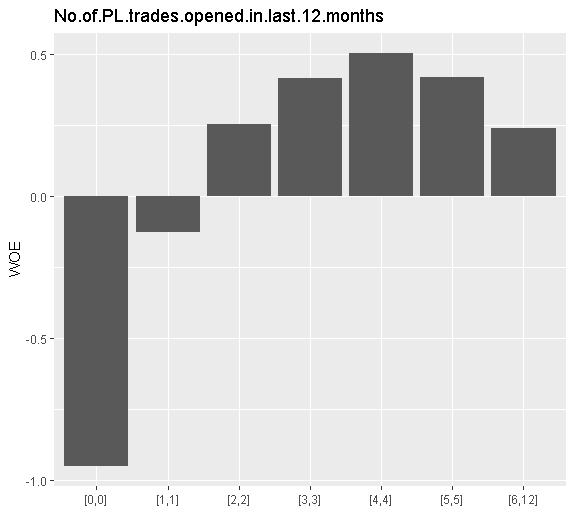


* We see that the applicant whose Avgas.CC.Utilization.in.last.12.months is between 0 to 14 has the higher chances of default.

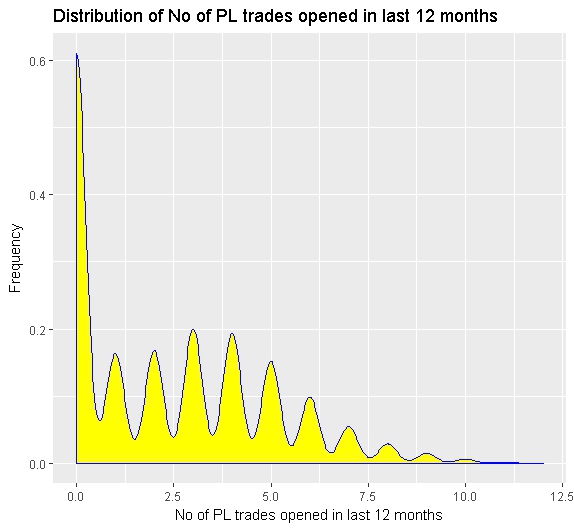


* We also see that the credit card utilization has outliers beyond 94% of the data, so we have capped the data to 94% and we also see that distribution is skewed towards right, so we have also transformed the variable using log transformation.
* Above graphs depicts the distribution of the utilization of credit card across customers and from the graph we can infer that, most of the customers max average utilization is 28. And there are few customers with high average spending on credit card, but most of the customers have average utilization between 0-25.

1. **No of PL trades opened in last 12 months**

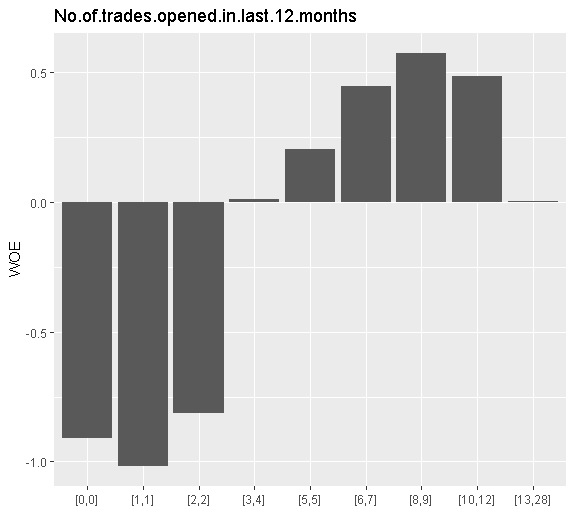
**

* We see that the applicant whose No.of.PL.trades.opened.in.last.12.months is 0 has the higher chances of default.

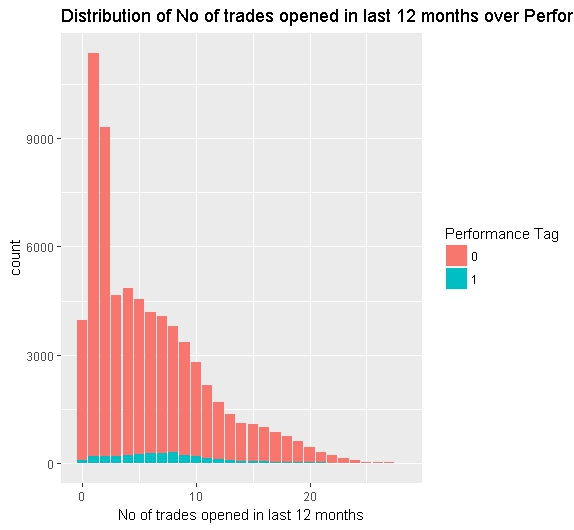


* We see that the distribution is right skewed and there are two outliers at 11 and 12, so we have log transformation to take care of both the problem.

1. **No of trades opened in last 12 months**

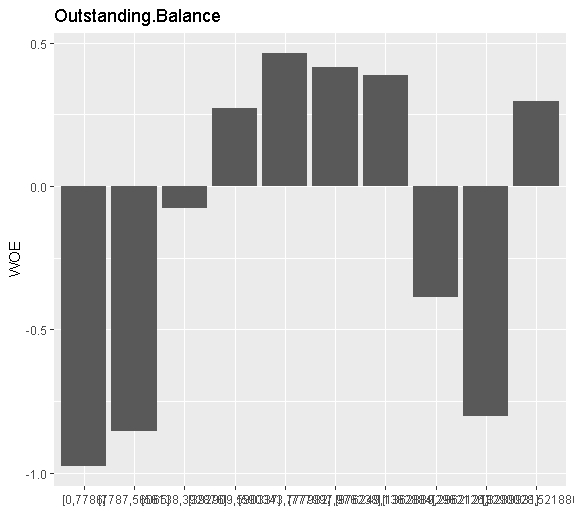
**

* We see that the applicant whose No.of.trades.opened.in.last.12.months is between 0 to 2 has the higher chances of default.

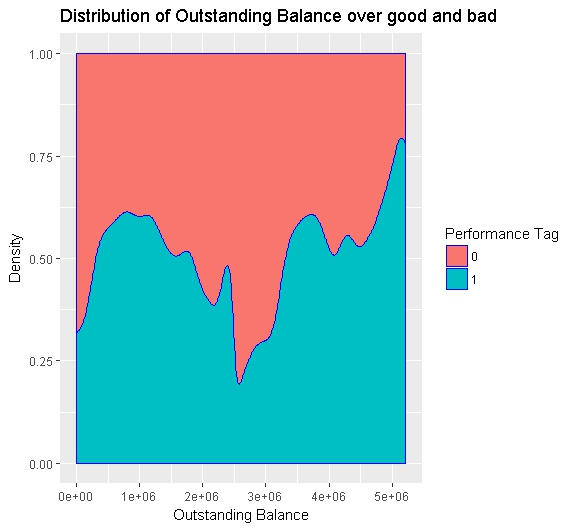


* We also see that the distribution is slightly skewed towards right, so we have use log transformation.
* There are high numbers of defaulters with 8 trades opened.
* Now we will discuss two more variable which has Medium predictive Power as per information value.

1. **Outstanding Balance**

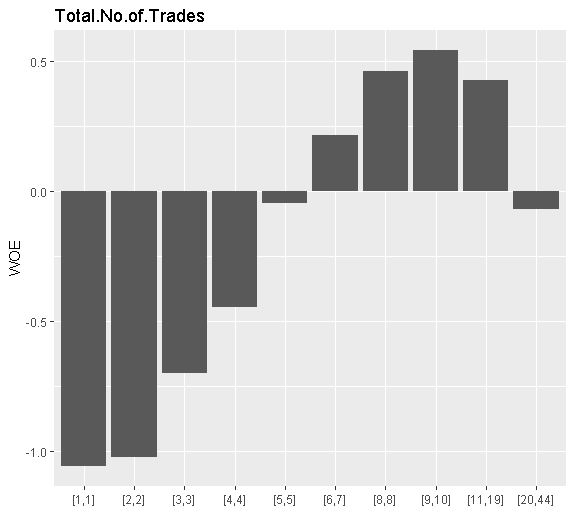


* We see that the applicant whose Outstanding Balance is between 0 to 56065 and between 1362889 to 3289931 has the higher chances of default and has the lowest WOE, which means this group of applicant has less number of good applicants in comparison to its bad customers.

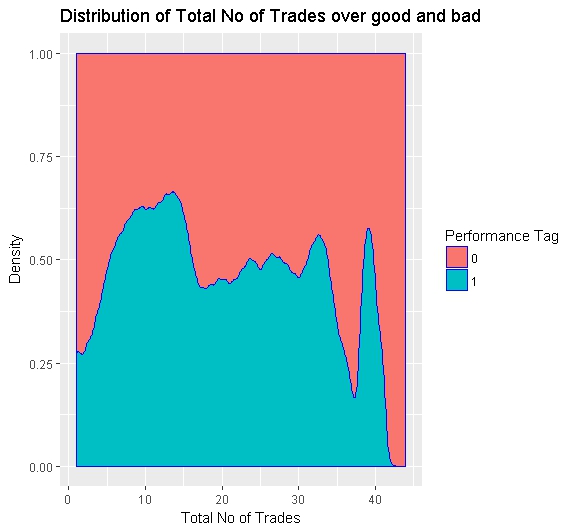


* We see beyond outstanding balance 5e+06 the percentage is default is more than 75%, which should be a concern for the company.

1. **Total No of Trade**

**

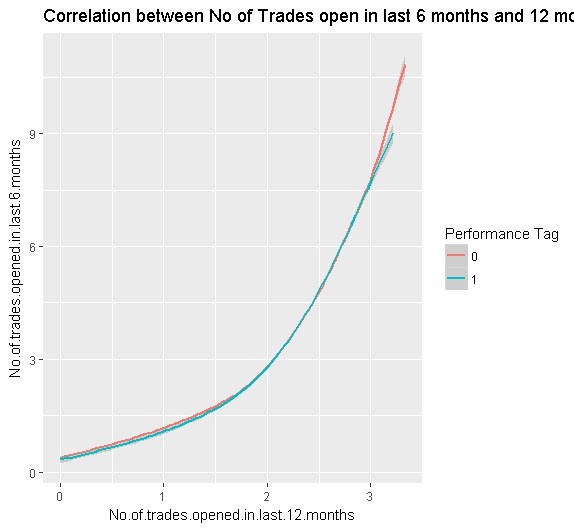
* We see that the applicant whose Total No of Trades is between 1 to 4 has the higher chances of default.



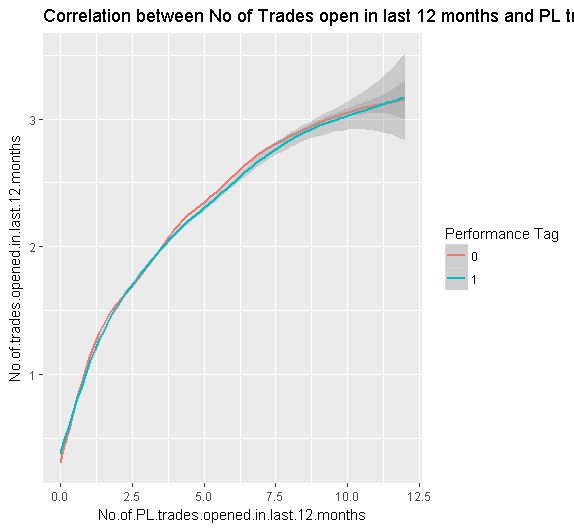
* We see that trades between 5 to 15 has default rate more than 50% and the default rate drops steeply after 40 transaction which may be an outliers.

# **EDA Plots and Descriptions of Insights for the rest of the variables**

1. **We see that there is correlation between "No of Trades open in last 6 months and 12 months".**

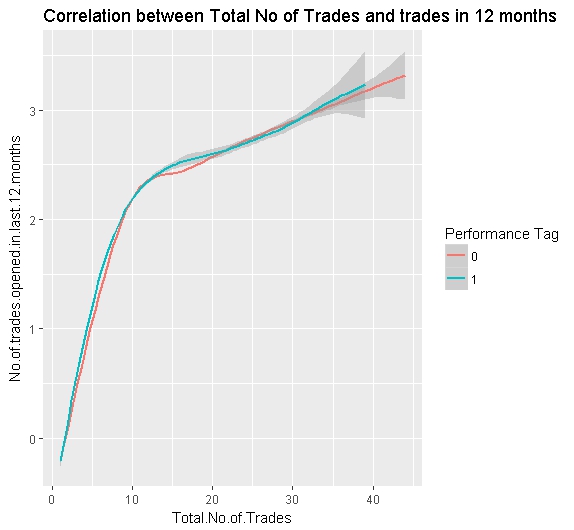
**

* we see that as the trades in last 12 month increases the number of trades in last 6 month also increases which is obvious

1. We see that there is correlation between “No of Trades open in last 12 months and PL trades in 12 months”.**

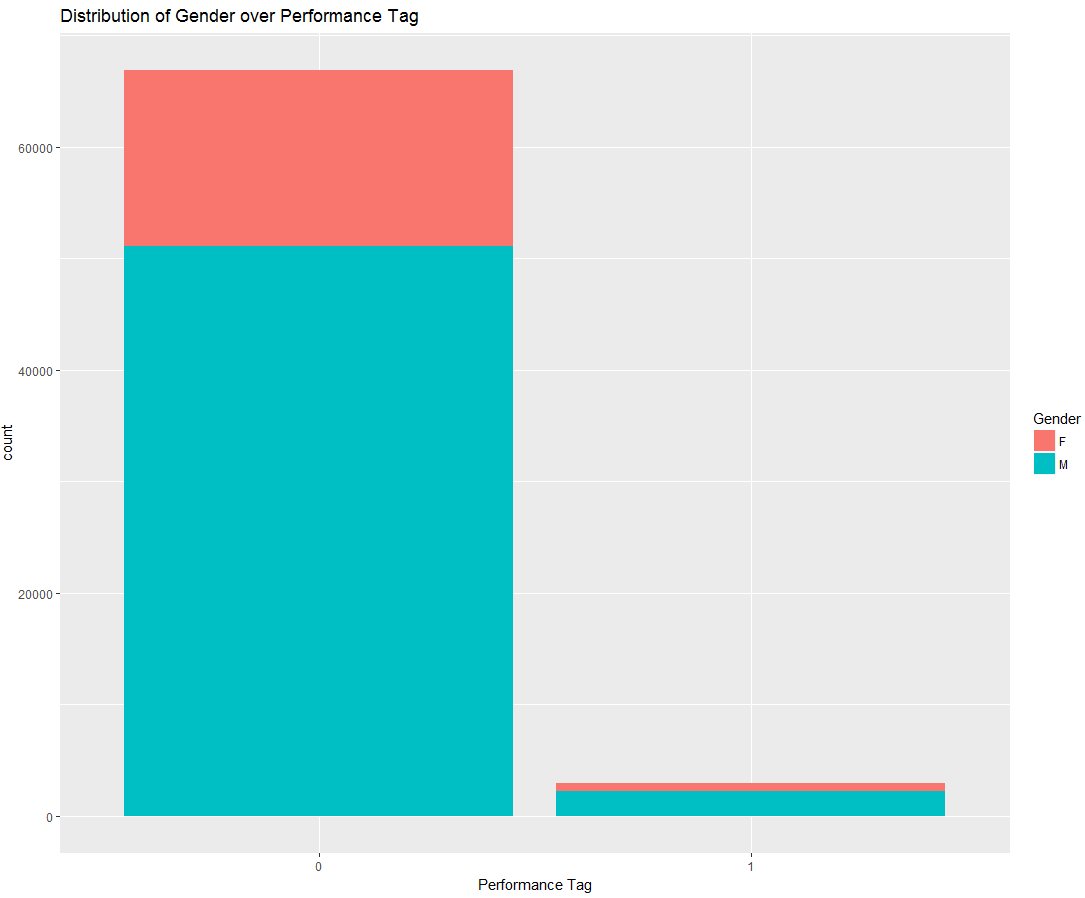
* We see that as no of pl trades in last 12 months increases no of trades in 12 months also increases.

1. We see that there is correlation between “Total No of Trades and trades in 12 months”.

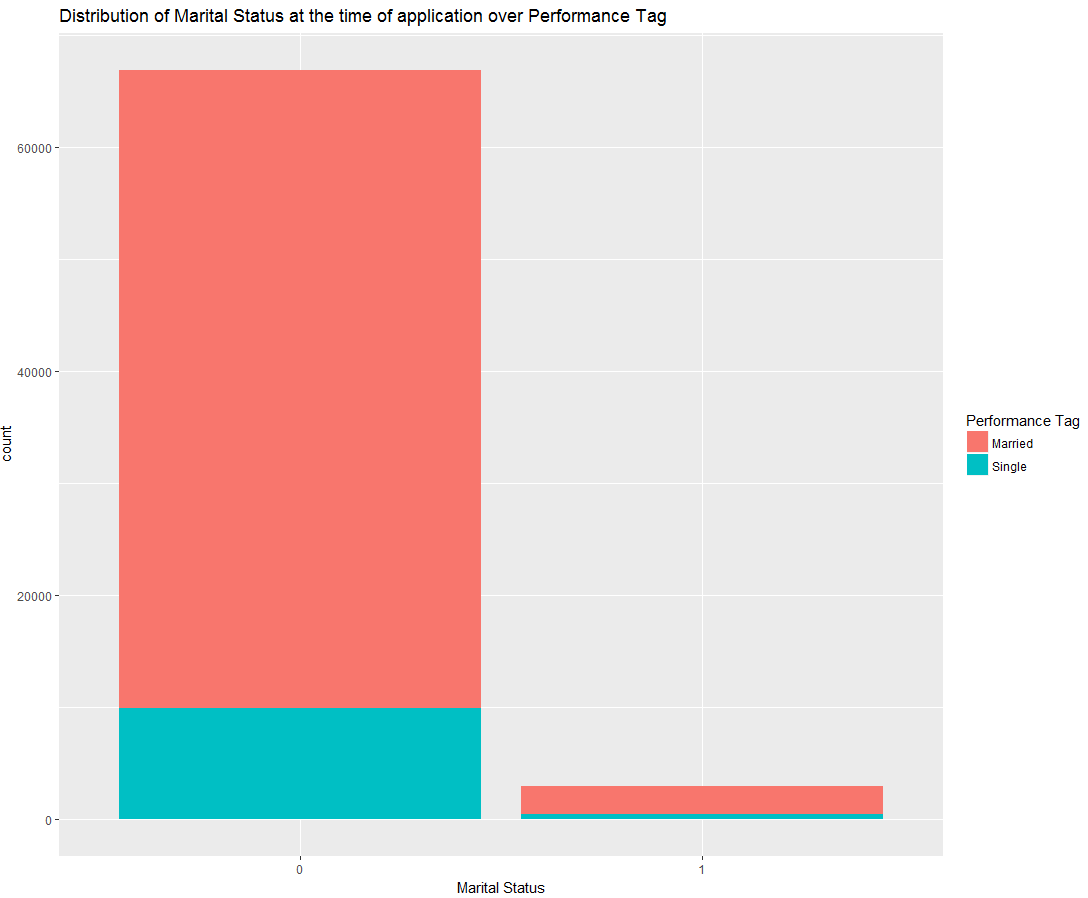
**

* We see that in the beginning months applicant uses the card very often till they have used 10 transaction and then they slow down their card usage.

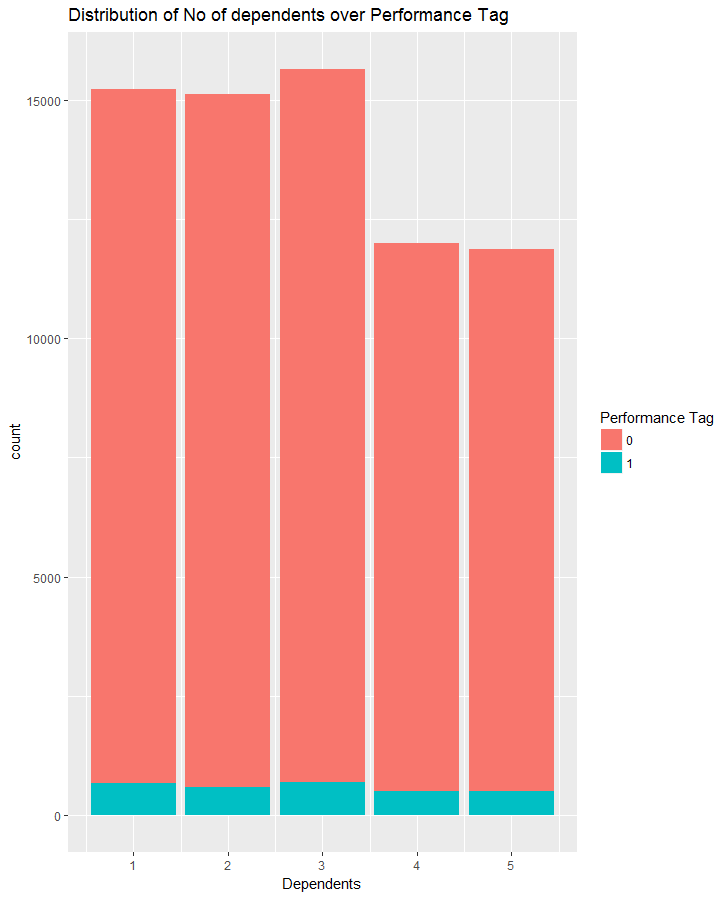
1. Distribution of Gender Vs. Performance Tag - From the below graph we can infer that, there are more number of males who defaulted.



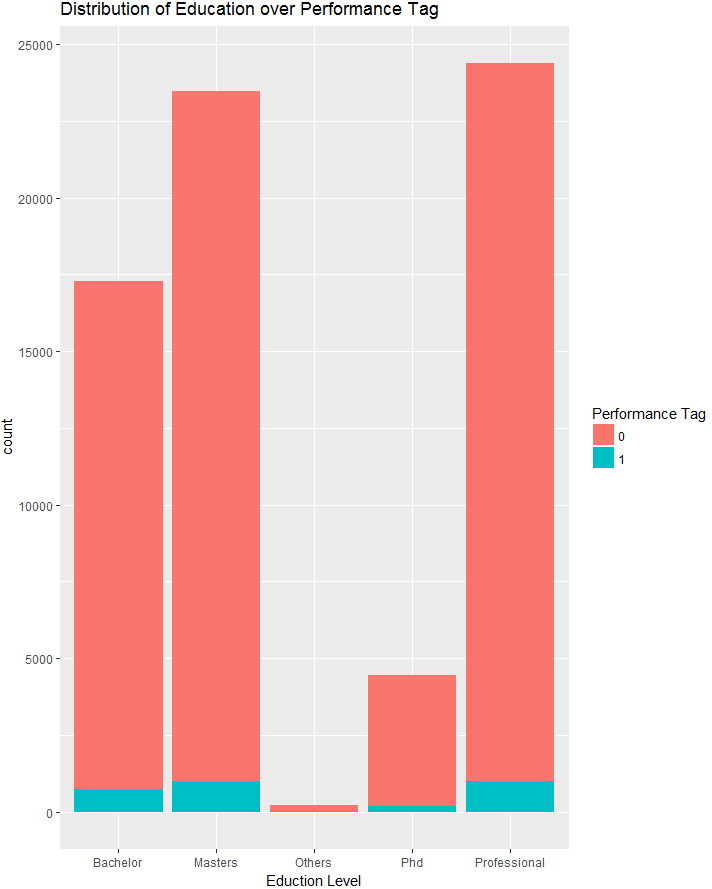
1. Distribution of Marital status Vs. Performance Tag - From the below graph we can infer that, there are more number of married customers who defaulted.



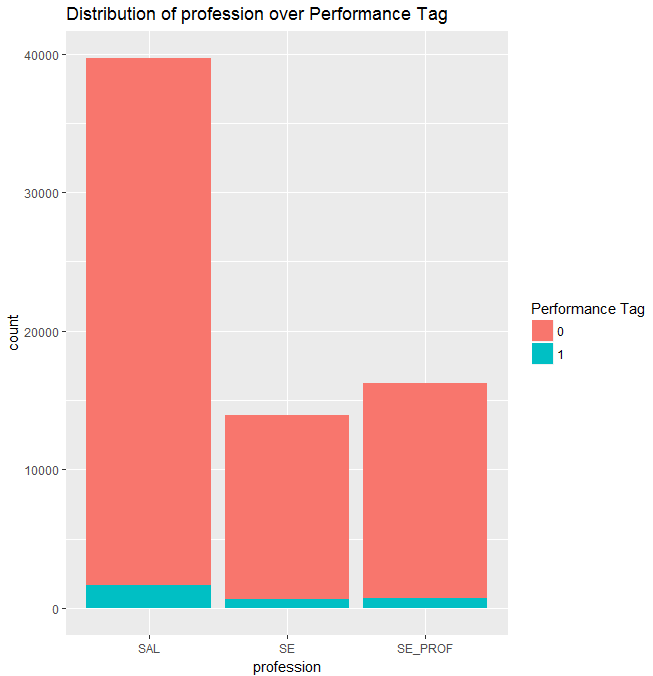
1. Distribution of Dependents Vs. Performance Tag - From the below graph we can infer that, customers with dependents 1,2,3 have higher number of defaults than 4,5.



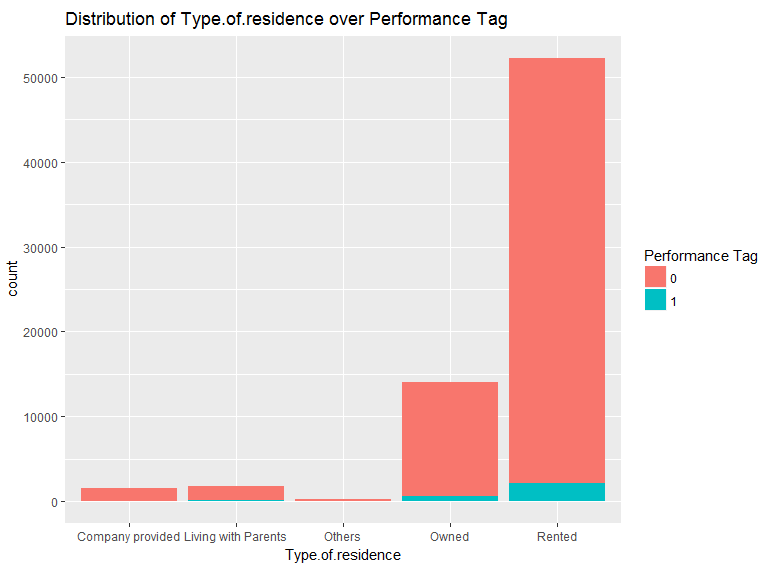
1. Distribution of Education Vs. Performance Tag - From the below graph we can infer that, there are more number of customers with professional degree/education who defaulted. And next category of education which has high number of defaults is Customers with Master degree*.*



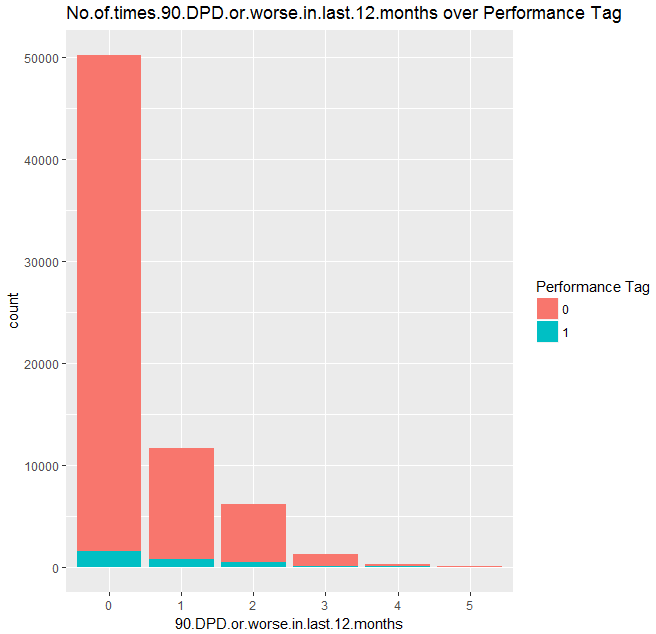
1. Distribution of Profession Vs. Performance Tag - From the below graph we can infer that, there are more number of customers under salaried profession who defaulted.



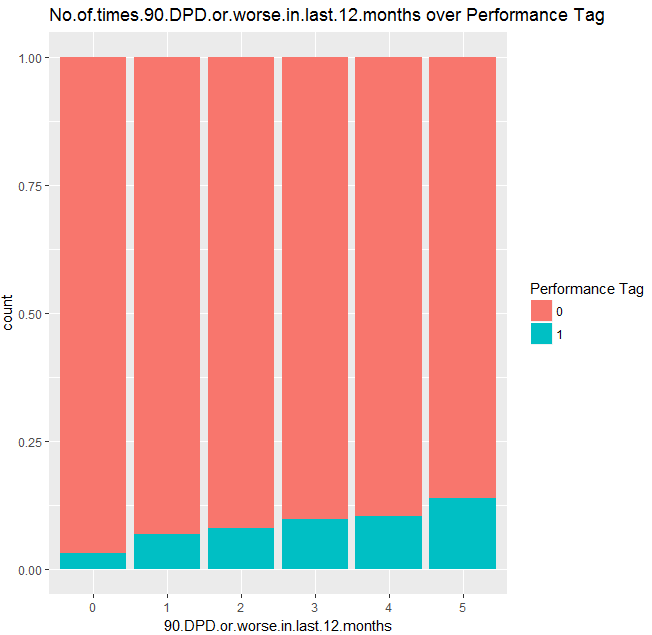
1. Distribution of Residence Vs. Performance Tag - From the below graph we can infer that, Customers living in rented houses, have more defaulters, than other type of residence.



1. Distribution of 90 DPD in 12 Months Vs. Performance Tag - From the below graph we can infer that, there are only few customers having 90 DPD more than 2 times.



1. Distribution of 90 DPD in 12 Months Vs. Performance Tag - From the below graph we can infer that, that customers having 90 DPD for 5 times, have higher number of defaulters. We can say that customers with 90 DPD more than or equal to 3 times have higher chances of default. And customers in 5th bucket should be considered as defaulters.



Since we are now done with our EDA analysis, and we know that which variable are significant and which variable are highly correlated, so we will use this knowledge to better understand the applicant and once we are done with our modelling we can check if our hypothesis assumption were correct or not.

We will first build a logistic regression model on demographic data and evaluate the prediction using the confusion matrix, ROC curve and see whether only demo data will help us in acquiring good customers. Then we will combine the credit data with demo data using application ID and again build logistic regression model and few other classification models like decision tree and compare the accuracy, sensitivity and specificity of both the models. Finally, we will create an application score card and basis on that we will decide on customers who are beneficial for the bank.

* **Data Cleaning and Transformation**
* **Missing value treatment** - We have removed the missing values from our data sets since it is only 2.5% of the total observation.
* **Outlier treatment**–We have use quartile and log transformation for outliers and skewed distribution.
* **Dummy Creation** – We have created dummy variable for all the categorical data using package AtConP.
* **Binning Variable** – We have created binning variable through “information package” and then impute the variable through “DF.Replace.Bin” of AtConP package.
* **WOE Variable** – We have created WOE values through “create\_infotables” of “information package” and then use “DF.Replace.WOE” of “AtConP package” to impute the woe values into original data sets.
* **Module Building and Evaluation:**
* First observation is that the data is imbalanced. Only 4% of data has defaulters. So this has been found to have performance problems with models such as Random Forest and SVM. However performance was found to be descent with Logistic regression models.

**So the approach taken is as follows:**

1. Regular logistic regression modeling with data with an 80 - 20 split.

2. Random forest data model with down sampling of the negative data to 10% so that we can get a 50/50 distribution of positive and negative data.

3. SVM Model with RBF and Polynomial kernel model with down sampling of the negative data to 10% so that we can get a 50/50 distribution of positive and negative data.

**Evaluate the data model using**

* Confusion matrix
* KS Statistic
* Lift chart

The model will then be able to output a probability of default. This probability can then be converted into a score.

* **Application Scorecard Implementation:**

Our current problem of identifying the defaults will come under "Operational Credit Risk Analytics", because we are identifying the risk at the individual level.

Now "Operational Credit Risk Analytics" is applicable at all the stages of the customer lifecycle.

a. During Acquisition -> New customers or Prospects will be a good customer or not

b. On Existing customers -> Monitor Credit health of the customer to identify delinquency

c. Collection and Recovery -> Recovery from delinquent customers and predict how much will they pay back

Application Scorecard can be used during Acquisition (for both rejected or accepted customers) or on the Existing customers.

**Steps to be followed for the building an Application Scorecard for individual customers:**

Step 1: Identify the customers as good or bad.

Though, we have tried different models (as mentioned in the Modeling section) we will use logistic regression model to predict the odds of being good for each applicant (logit equation).

As per the instructions in the project:

Good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

**Explanation and understanding**

Good/Bad = 10/1. It means for 10 Good there is 1 Bad

Now we know Good/Bad = ODDs

Based on the above information Good to Bad Ration will be:

**Score Good:Bad**

400 10:1

420 20:1

440 40:1

460 80:1

. .

. .

. .

Step 2: Creating Application Scorecard. Once we have the odds, we will sort the applicants from high to low odds (i.e. good to bad). As mentioned in the basics of BFSI domain, banking industry prefers to talk in terms of scores (e.g. between 200 to 900) rather than odds. Thus, we will calibrate the odds to a scale of scores between 200 to 900 for all the Applicants.

In simple terms we will create an Application Scorecard.

Note: Scores are just odds calibrated to a different scale.

Each, row in the Application scorecard represents

-----------------------------------------------------------------------------------------------------------------------

Applicant ID | Predicted P(good) by the regression model | Corresponding odds(good) | Log of odds(good) | Final score |

-----------------------------------------------------------------------------------------------------------------------

The higher the score, the better the customer is from a risk perspective.

Step 3: Identify the cut-off (threshold) score to label applicants as "good" or" bad", below which credit cards will not be granter to applicants. We will use probability threshold in a logistic regression model to decide the threshold score.

Step 4: Validation of this will be done by keeping in mind the below given aspects of the model validation (1 or 2 will be used):

a. Discriminatory Power -> Can be measured through KS Statistics/GINI Index/Sensitivity and Specificity (measured by Confusion Matrix).

b. Calibration or Predictive Accuracy.

c. Stability (Since for this we need to keep an eye on the variable stability, Population Stability Index and the predictive patter, which we will know after Model is deployed and running for some time, therefore we will not use this aspect.).

**Optional (If time permits):**

1. Use of Reject Inference to reduce/eliminate biased sampling.

The problem with the data available at acquisition is that you only have the data of the approved candidates and not the ones who were declined credit (since they discontinued their relationship with the bank after rejection). If you use this data to make decisions for all the applicants, it will be a case of biased sampling. To avoid this bias, any of the reject inference techniques can be used.

In our case we will be using Credit Bureau data for Reject inference.