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 time 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 3 duplicates application id.

Data Cleaning

We see that there are 1425 NA’s in our dependent variable in our merged file which indicates which applicant has default on his or her credit card payment or not. So, since the data has no information about default, this applicant are the one who have not been given credit card by the company, so we will keep this data in our validation sets.

After removing validation data, we are left with 69867 obs 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 a Identity variable which we cannot use in our analysis.

Exploratory Data Analysis

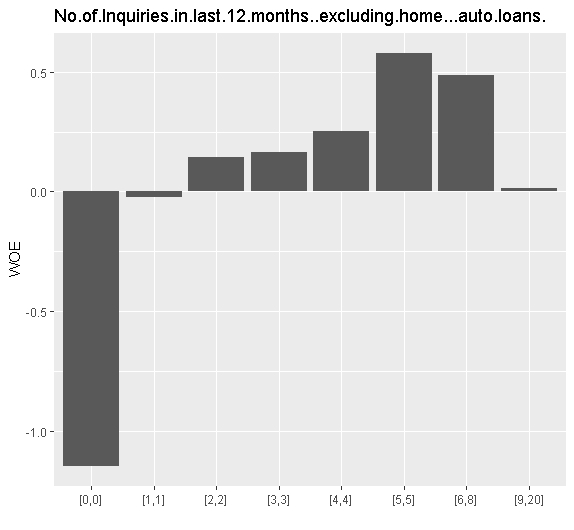
*We will be using WOE and IV for EDA.*

*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*

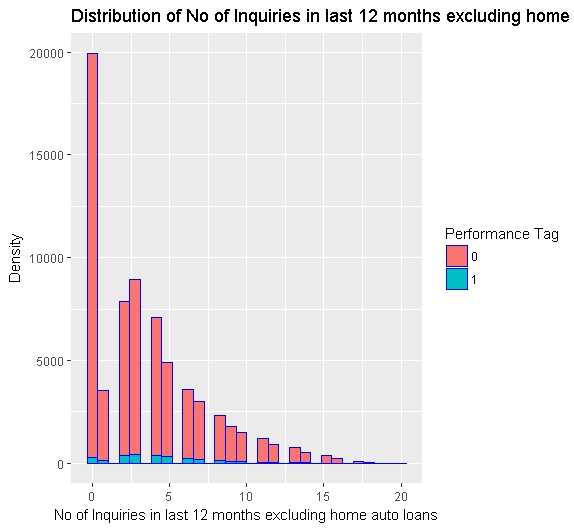
*Strong predictive Power where as “Outstanding Balance” and “Total No of Trade” has Medium predictive Power.*

*Now let check the plot of these variables with respect to WOE.*

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

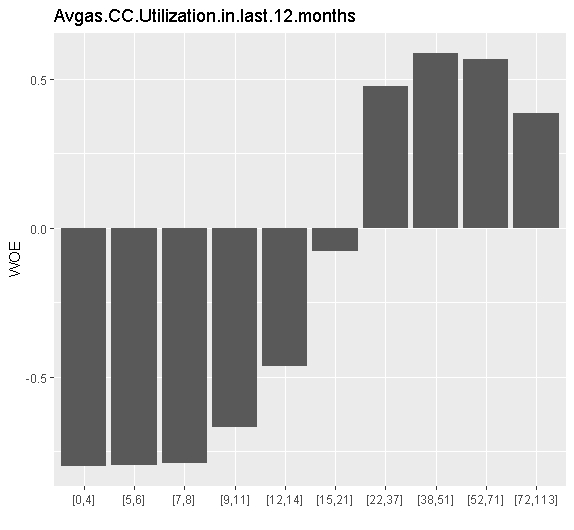
**

* 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.

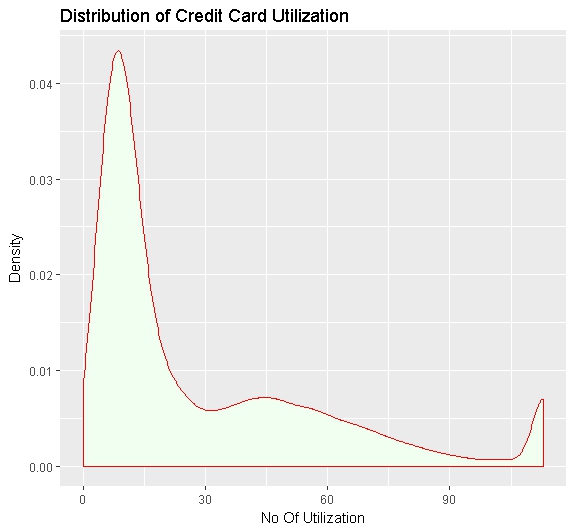


* We see that the distribution is right skewed, so we will use log transformation to make the distribution normal.

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

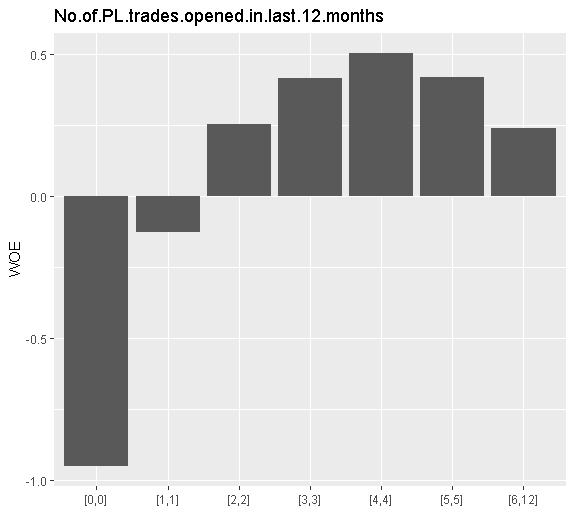


* We see that the applicant whose Avgas.CC.Utilization.in.last.12.months is betw een 0 to 14 has the higher chances of default.

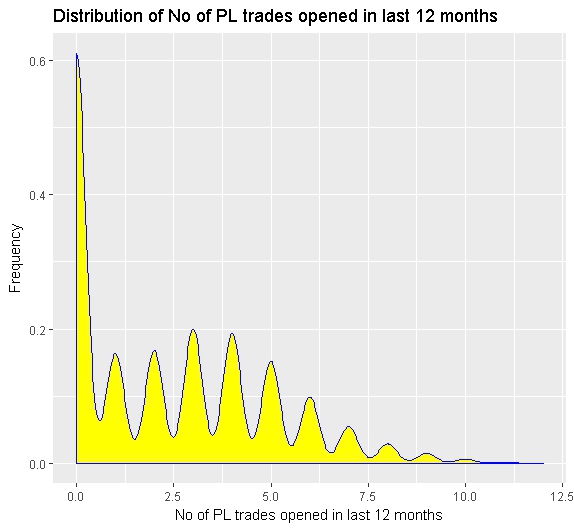


* We also see that the credit card utilization has outliers beyond 94% of the data, so we have cap the data to 94% and we also see that distribution is skewed towards right, so we have also transformed the variable using log transformation.
* Above graphs depicts the distribution of the utilization of 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*

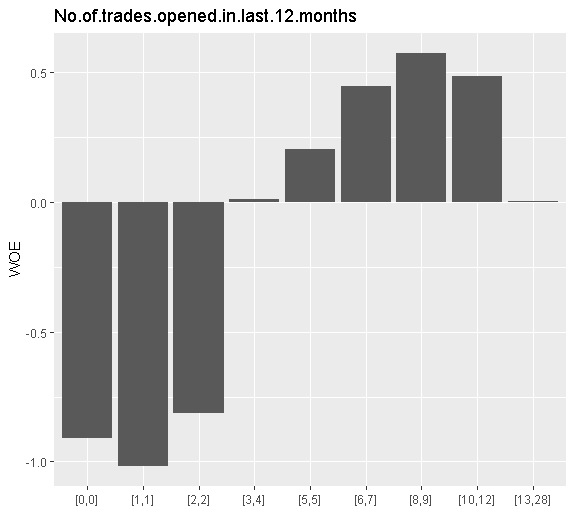
**

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

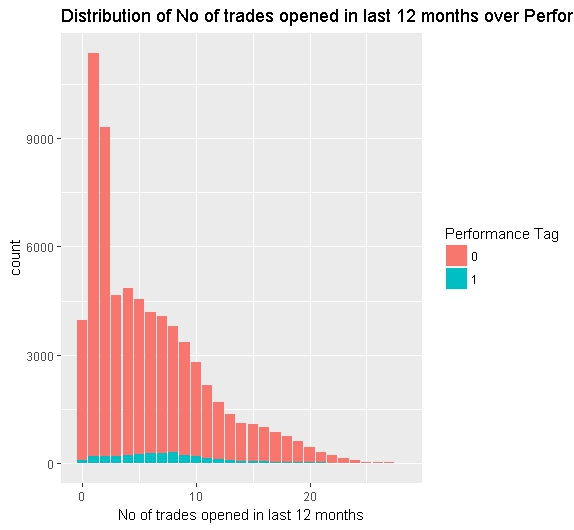


* We see that the distribution is right skewed and there are two outliers at 11 and 12, so we have log transformation to take care of both the problem.

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

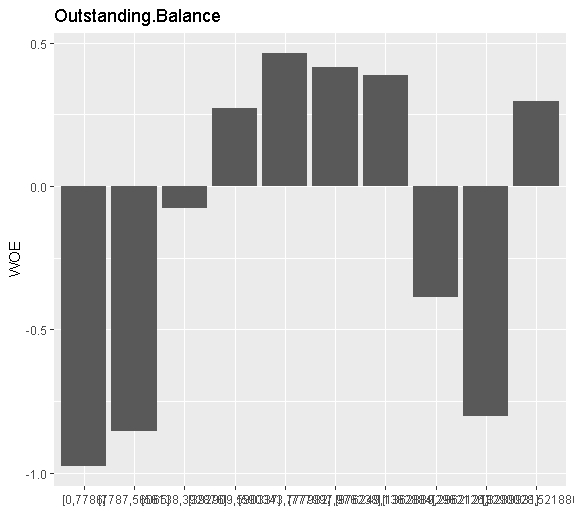
**

* 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

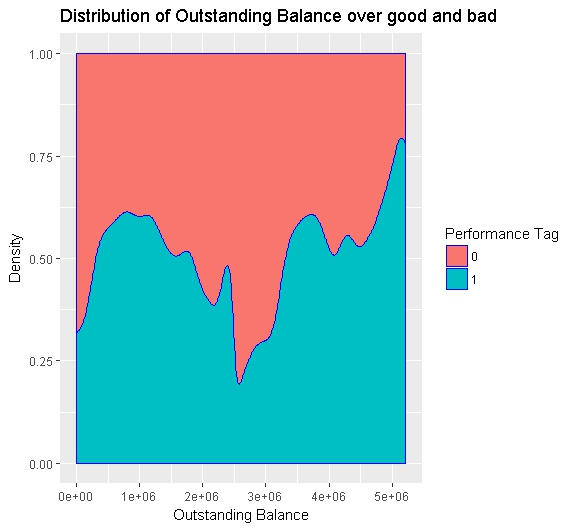


* We also see that the distribution is slightly skewed towards right, so we have use log transformation.
* There are high numbers of defaulters with 8 trades opened.
* Now we will discuss two more variable which has Medium predictive Power as per information value.

1. *Outstanding Balance*

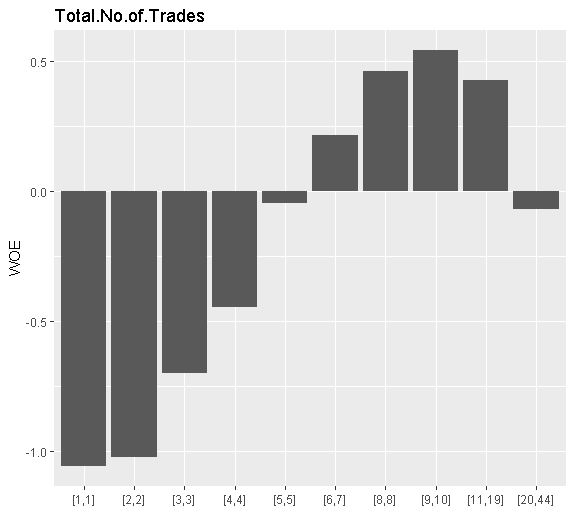


* We see that the applicant whose Outstanding Balance is between 0 to 56065 and between 1362889 to 3289931 has the higher chances of default and has the lowest WOE, which means this group of applicant has less number of good applicant in comparison to its bad customers.

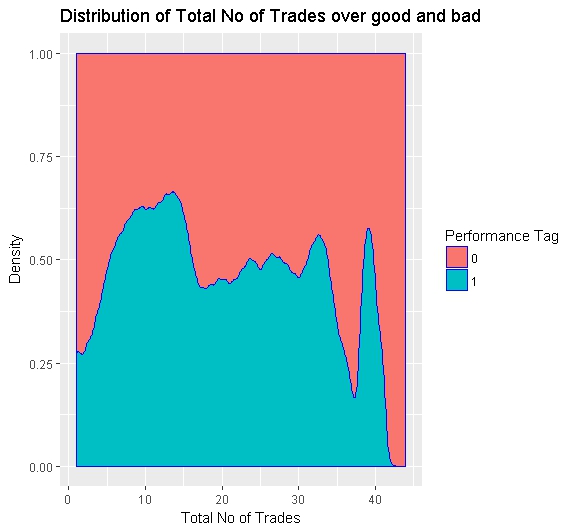


* We see beyond outstanding balance 5e+06 the percentage is default is more than 75%, which should be a concern for the company.

1. *Total No of Trade*

**

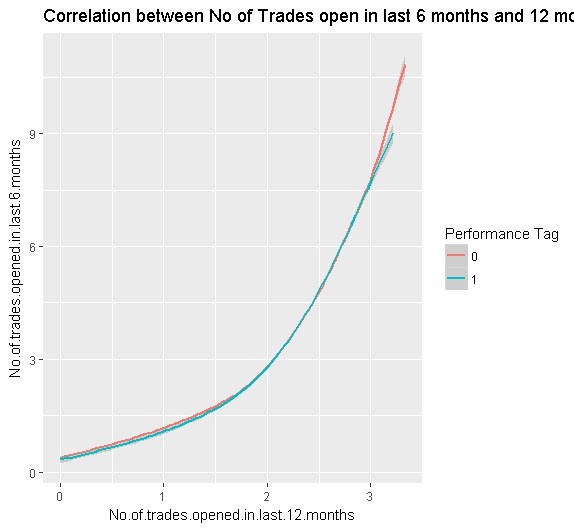
* We see that the applicant whose Total No of Trades is between 1 to 4 has the higher chances of default.



* 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.

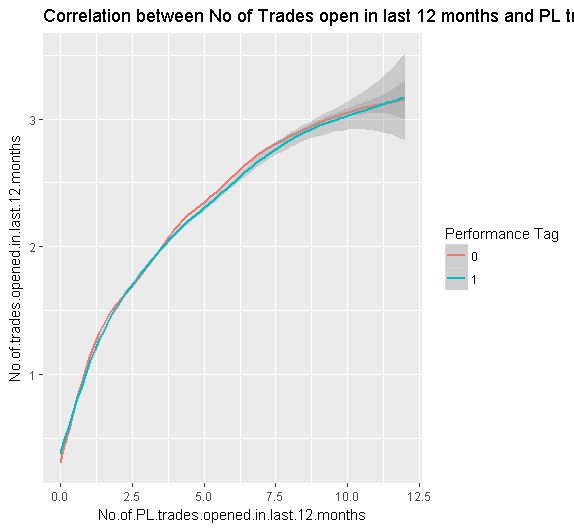
# EDA Plots and Descriptions of Insights for the rest of the variables

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

**

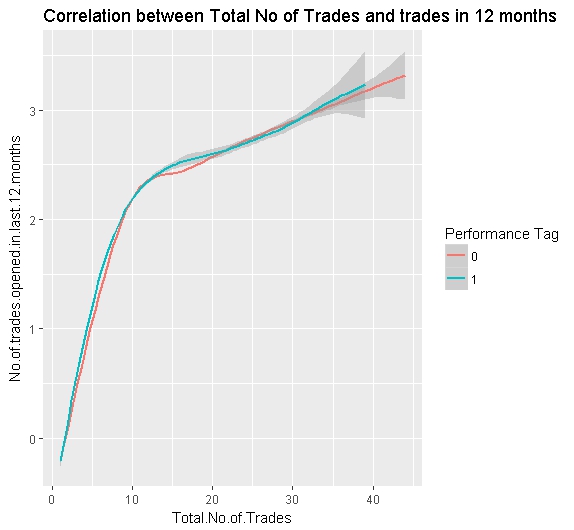
* we see that as the trades in last 12 month increases the number of trades in last 6 month also increases which is obvious

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

**

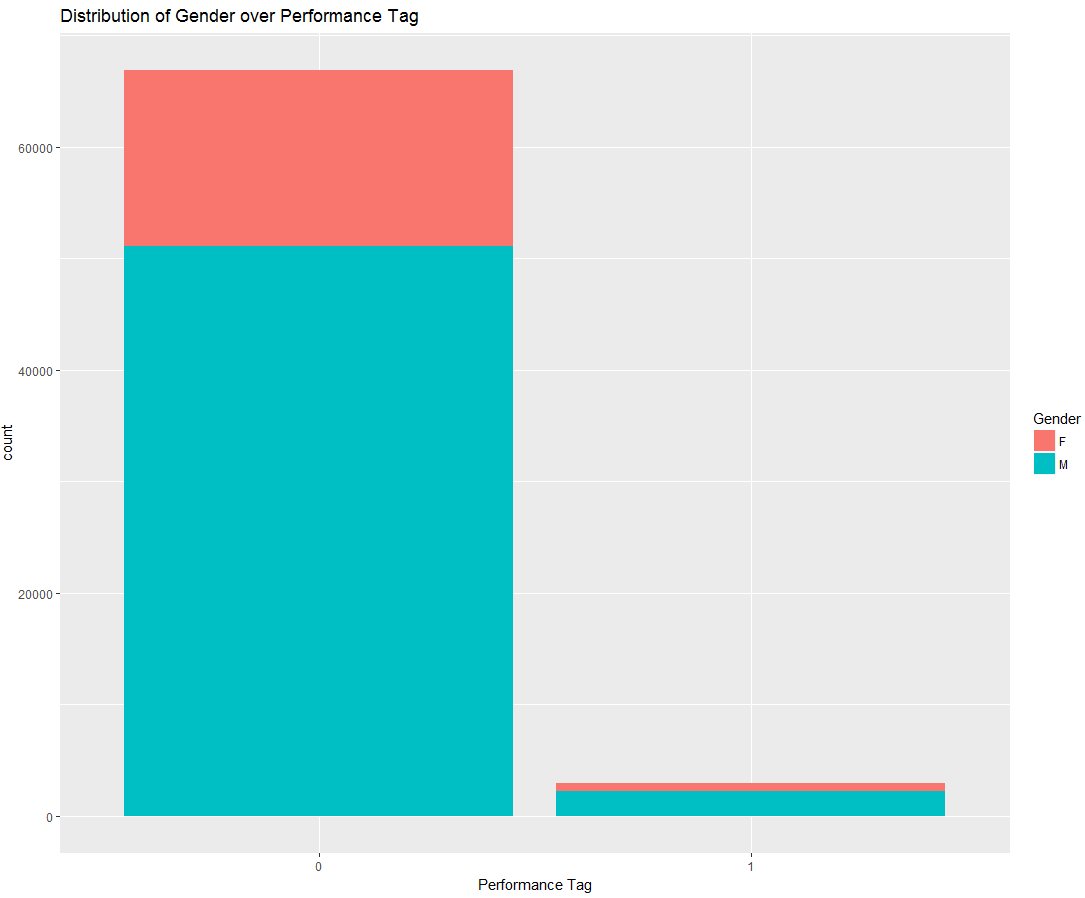
* We see that as no of pl trades in last 12 months increases no of trades in 12 months also increases.

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

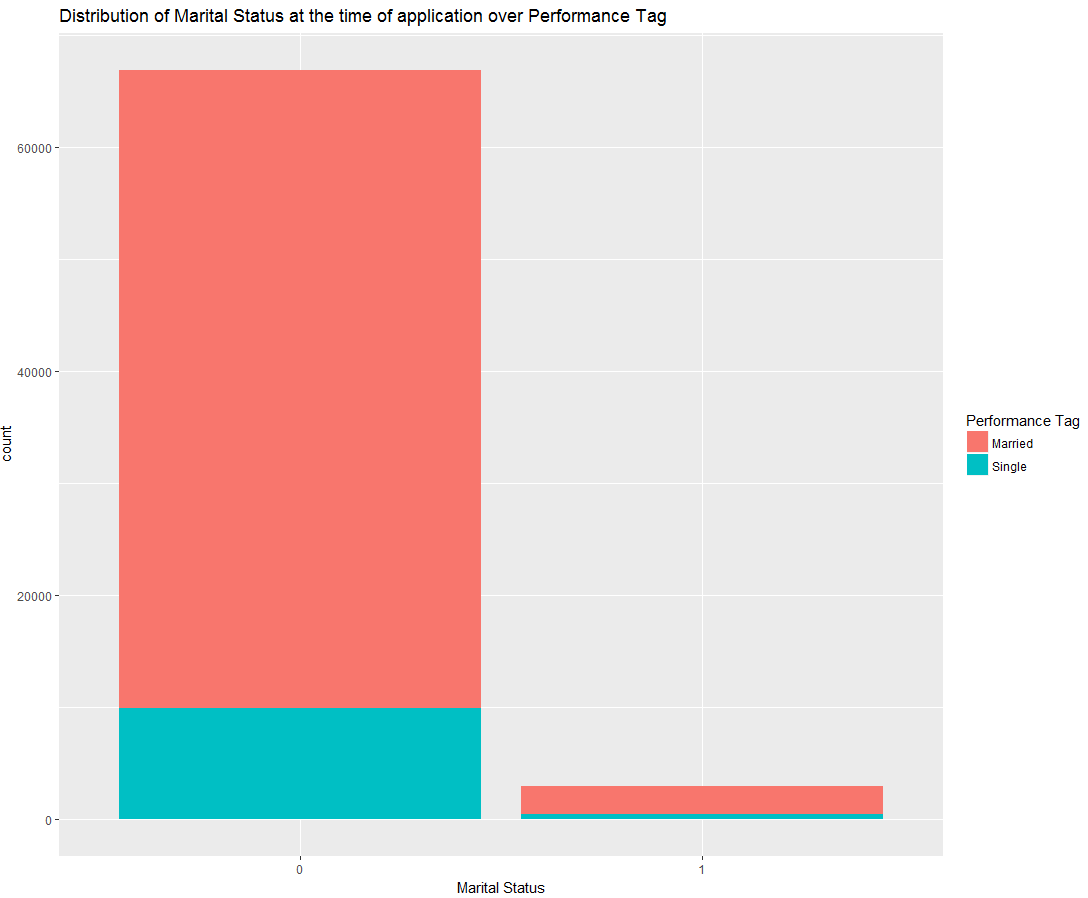
**

* 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.

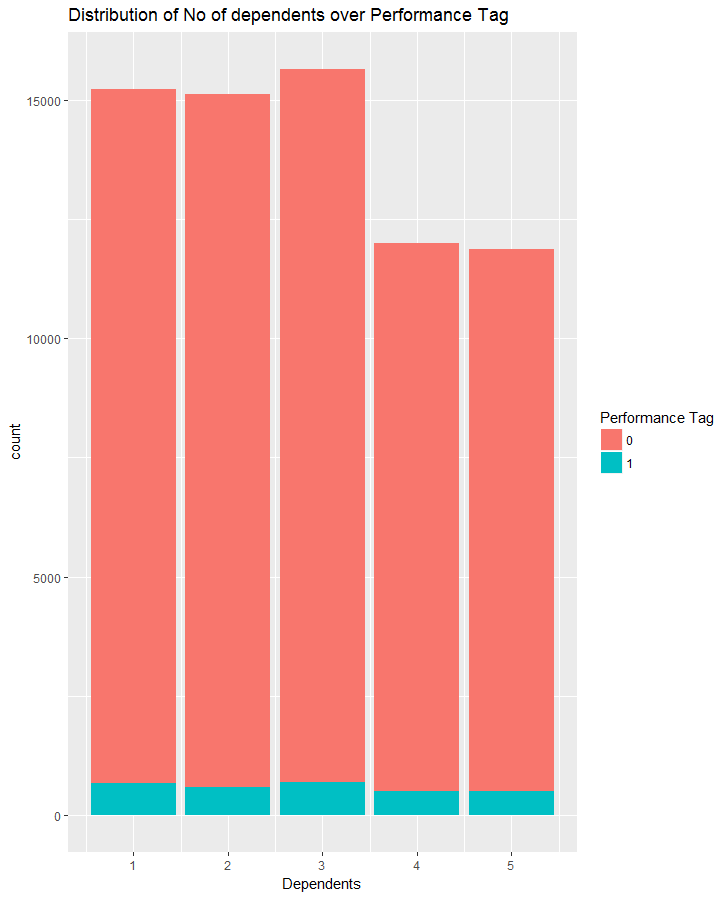
1. *Distribution of Gender Vs. Performance Tag - From the below graph we can infer that, there are more number of males who defaulted.*



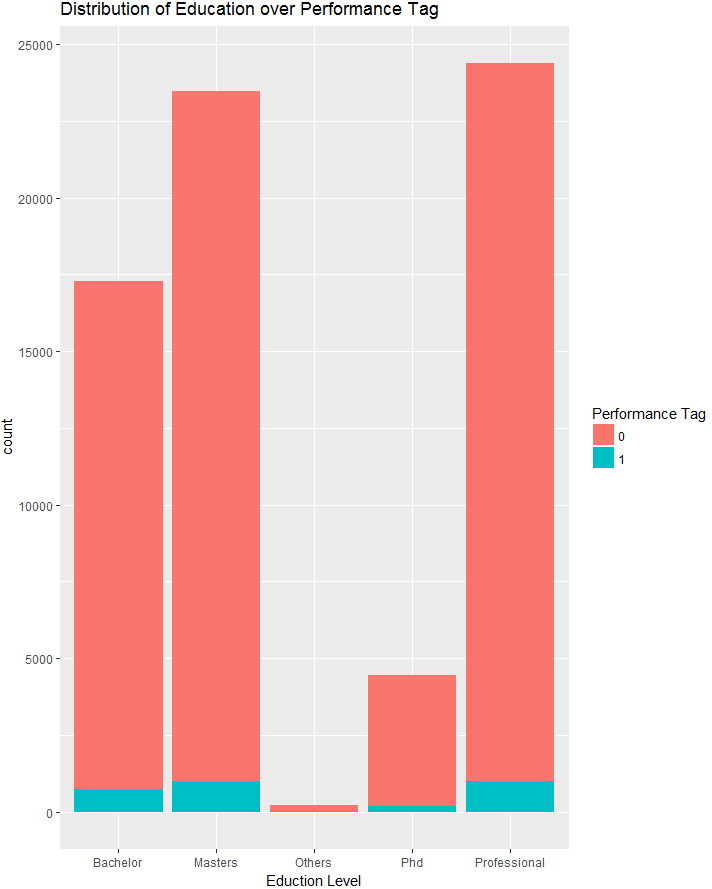
1. *Distribution of Marital status Vs. Performance Tag - From the below graph we can infer that, there are more number of married customers who defaulted.*



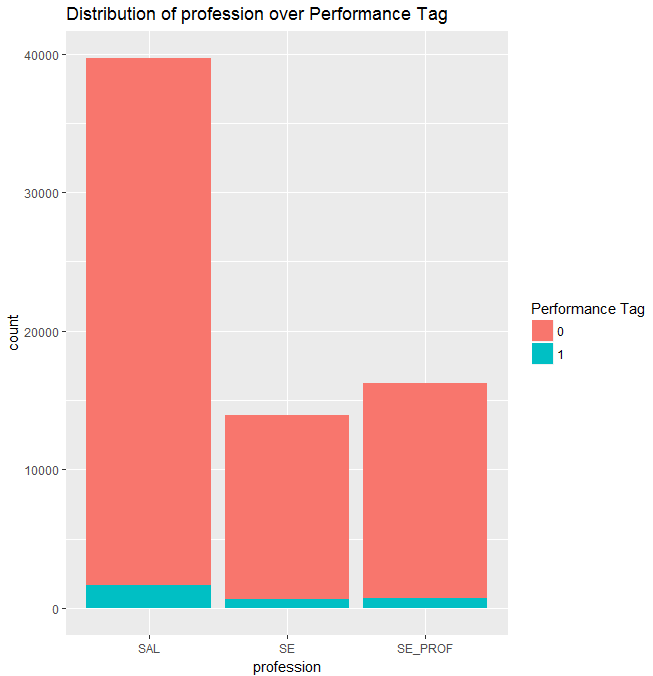
1. *Distribution of Dependents Vs. Performance Tag - From the below graph we can infer that, customers with dependents 1,2,3 have higher number of defaults than 4,5.*



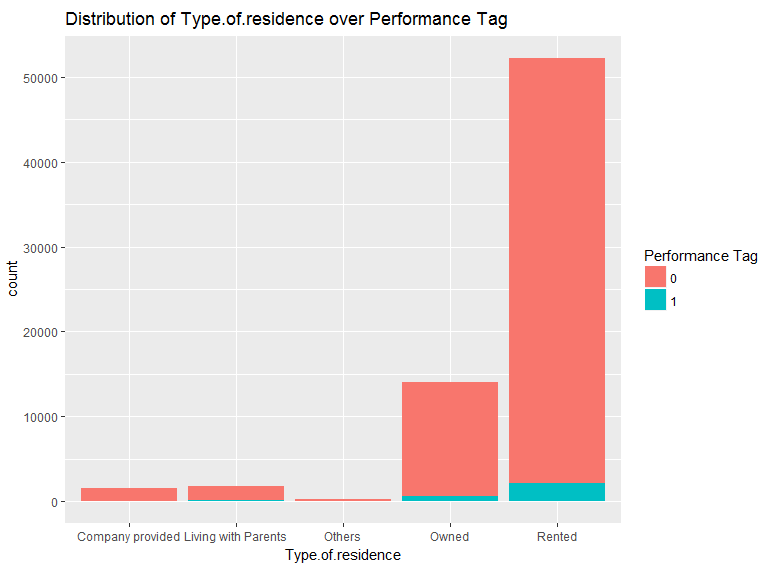
1. *Distribution of Education Vs. Performance Tag - From the below graph we can infer that, There are more number of customers with professional degree/education who defaulted. And next category of education which has high number of defaults is Customers with Master degree.*



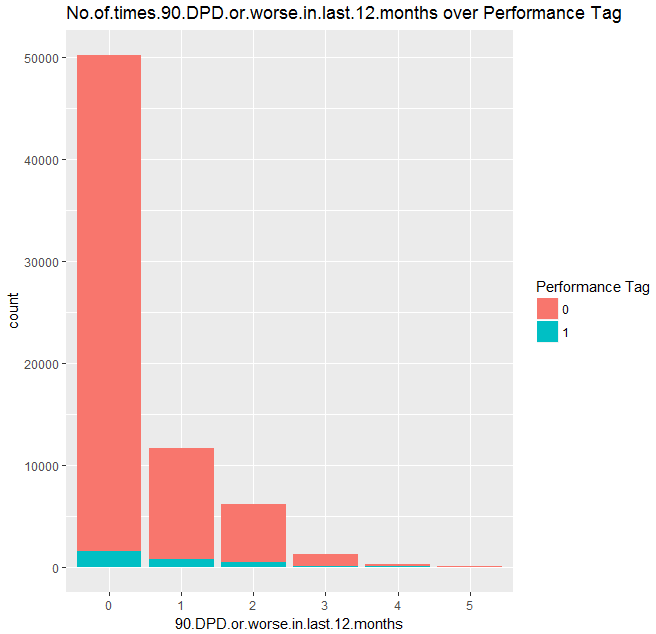
1. *Distribution of Profession Vs. Performance Tag - From the below graph we can infer that, There are more number of customers under salaried profession who defaulted.*



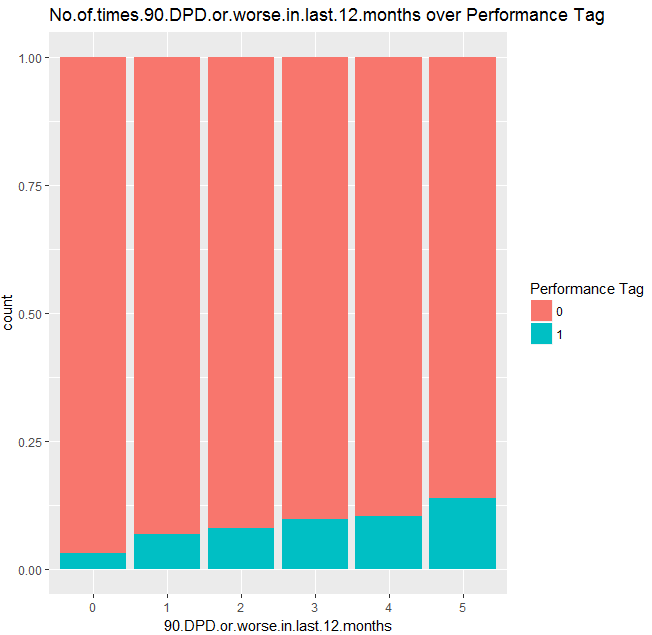
1. *Distribution of Residence Vs. Performance Tag - From the below graph we can infer that, Customers living in rented houses, have more defaulters, than other type of residence.*



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 remove 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 k-1 dummy variable for all the categorical data using package AtConP form <https://github.com/prk327/AtConP.git> through function “df.matrix”.
* **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.
* ***Splitting the Dataset into train and test***
* We have two data sets, one original raw data and another woe values data. We then have created a subset of demographic data from both raw and woe data sets.
* We have divided our four data into 80:20 ratio. The data sets is imbalance since 96% of the dependent variable are for good and only 4% represent bad.
* 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 & Evaluation***
* **Demographic data model –**

We have use Logistic Regression to train our model and below are the metrics:

|  |  |
| --- | --- |
| **Significant variables in final model** | **Coefficients value (Numeric)** |
| AgeWOE | 0.030159 |
| NoofdependentsWOE | 0.037780 |
| IncomeWOE | 0.140696 |
| ProfessionWOE | 0.043194 |
| NoofmonthsincurrentresidenceWOE | 0.203507 |
| NoofmonthsincurrentcompanyWOE | 0.123066 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 74676 |
| Null deviance | 76188 |
| Residual Deviance | 74662 |

We have use StepAIC to select the model based on AIC value, then we have use VIF to check for any multicollinearity,

Then we have reduce the variable based on its P-Value, we have consider only those variables which have significant p-value. Below is our Model for demographic data

log(odds) = -0.047434 + 0.030159 (AgeWOE) + 0.037780 (NoofdependentsWOE) + 0.140696 (IncomeWOE) + 0.043194 (ProfessionWOE) + 0.203507 (NoofmonthsincurrentresidenceWOE) + 0.123066 (NoofmonthsincurrentcompanyWOE)

We have evaluated our model based on accuracy, Sensitivity, Specificity, C-statistic, KS-statistic and AUC:

|  |  |
| --- | --- |
| **Metrics** | **Values (Numeric)** |
| C-statistic | 5.98 |
| KS-statistic | 0.27 |
| Overall Accuracy | 56% |
| Sensitivity | 56% |
| Specificity | 57% |
| Area under the curve (AUC) | 56% |

* **Combined data final model:**

We have use Logistic Regression on master\_woe\_train\_rose data to train our model and below are the metrics:

|  |  |
| --- | --- |
| Significant variables in final model | Coefficients value (Numeric) |
| MaritalStatusatthetimeofapplicationWOE | -0.025522 |
| NoofdependentsWOE | 0.032028 |
| IncomeWOE | 0.042891 |
| ProfessionWOE | 0.025999 |
| NoofmonthsincurrentcompanyWOE | 0.035899 |
| Nooftimes30DPDorworseinlast6monthsWOE | 0.081673 |
| Nooftimes90DPDorworseinlast12monthsWOE | 0.044641 |
| Nooftimes30DPDorworseinlast12monthsWOE | 0.066104 |
| AvgasCCUtilizationinlast12monthsWOE | 0.12487 |
| Nooftradesopenedinlast12monthsWOE | 0.072854 |
| NoofPLtradesopenedinlast12monthsWOE | 0.07824 |
| NoofInquiriesinlast6monthsexcludinghomeautoloansWOE | 0.062323 |
| NoofInquiriesinlast12monthsexcludinghomeautoloansWOE | 0.127238 |
| OutstandingBalanceWOE | 0.06989 |
| TotalNoofTradesWOE | 0.038813 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 70932 |
| Null deviance | 76188 |
| Residual Deviance | 70900 |

We have evaluated our model based on accuracy, Sensitivity, Specificity, C-statistic, KS-statistic and AUC:

|  |  |
| --- | --- |
| **Metrics** | **Values (Numeric)** |
| C-statistic | 6.88 |
| KS-statistic | 5th Decile |
| Overall Accuracy | 64% |
| Sensitivity | 64% |
| Specificity | 64% |
| Area under the curve (AUC) | 64% |
| CutOff | 0.5247475 |

* **Reject Inference –**

We have use the above model to predict the good and bad for the validation data sets which we have removed in the beginning.

We have built the final model from the final data sets as below:

|  |  |
| --- | --- |
| Significant variables in final model | Coefficients value (Numeric) |
| (Intercept) | -0.2028 |
| NoofdependentsWOE | 0.04116 |
| NoofmonthsincurrentcompanyWOE | 0.04761 |
| Nooftimes60DPDorworseinlast6monthsWOE | 0.04195 |
| Nooftimes30DPDorworseinlast6monthsWOE | 0.04614 |
| Nooftimes90DPDorworseinlast12monthsWOE | 0.0493 |
| Nooftimes30DPDorworseinlast12monthsWOE | 0.05215 |
| AvgasCCUtilizationinlast12monthsWOE | 0.13532 |
| Nooftradesopenedinlast12monthsWOE | 0.09748 |
| NoofPLtradesopenedinlast12monthsWOE | 0.07697 |
| NoofInquiriesinlast6monthsexcludinghomeautoloansWOE | 0.08263 |
| NoofInquiriesinlast12monthsexcludinghomeautoloansWOE | 0.17659 |
| OutstandingBalanceWOE | 0.06403 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 72048 |
| Null deviance | 77768 |
| Residual Deviance | 72022 |

We have evaluated our model based on accuracy, Sensitivity, Specificity, C-statistic, KS-statistic and AUC:

|  |  |
| --- | --- |
| **Metrics** | **Values (Numeric)** |
| C-statistic | 6.832209e-01 |
| KS-statistic | 4th Decile |
| Overall Accuracy | 65% |
| Sensitivity | 64% |
| Specificity | 65% |
| Area under the curve (AUC) | 64% |
| CutOff | 0.5346465 |

* ***Application Score Card***

We have use the above model to create a application score card, we have create a custom function to generate a score card. The Cutoff score value were selected as 108, through which we can approve an applicant which are greater than this threshold value.

**Thank You**