Acquire the Right Credit Card Customers By using Machine Learning Algorithms

Team Members

Shobhana Halapeti

Ankit Jha

Rohit Garg

Jun 2019

Project Description

Problem Description :-

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 CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’

Objective : -

To help CredX identify the right customers using past data of the bank’s applicants.
To determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Datasets Provided :- We have been provided 2 data sets.

- A dataset with customer’s demographic details.
- A dataset with Credit history of the customer.

Solution Approach

The high-level Steps for arriving at the solution are below:

- Data cleaning and preparation(Distribution of data, missing values, Outliers)
- Perform Exploratory Data Analysis (EDA) and derive Insights(Univariate analysis, Bi Variate Analysis)
- Apply Weight of Evidence (WOE) & Information Value Analysis
- Identifying the most impacting attributes of the dataset
- Model Selection(Based on type of data)
- Model Building (Logistic , Random Forest , Support Vector Machines)
- Model Evaluation
- Model Fine Tuning
- Build Application Scorecard
- Assessing the financial benefit of your project

Data cleaning and preparation

1. In Demographic data we have **65 records where Customer's age is less than 18. Customer with Age < 18 are omitted** from both the datasets because they are not eligible for the credit card. Some of the records even show they are married which is inconsistent with the society norms.

2. **Assumption:**- As per the data dictionary, **"No of dependents"** variable in Demographic Dataset represents "No of Children". We have observed that the customers with Single marital status have children which cannot be the case. However, **we have considered this variable for model building assuming that it represents no of dependents** (not "no of children").

3. Negative & zero values for **"Income"** variable are considered as **Missing** and **imputed** using **regression**.

4. It has been observed that there are customers with **missing credit card utilization**. In some of these cases, the other variables in the Credit Bureau data are zero. As per the available information on record, these are the cases where either there is no hit in the credit bureau or the applicant does not have any other credit card. Hence, **the missing "Credit Card Utilization "in the Credit Card data is replaced with 0** considering the fact that these customers have the relevant demographic data.

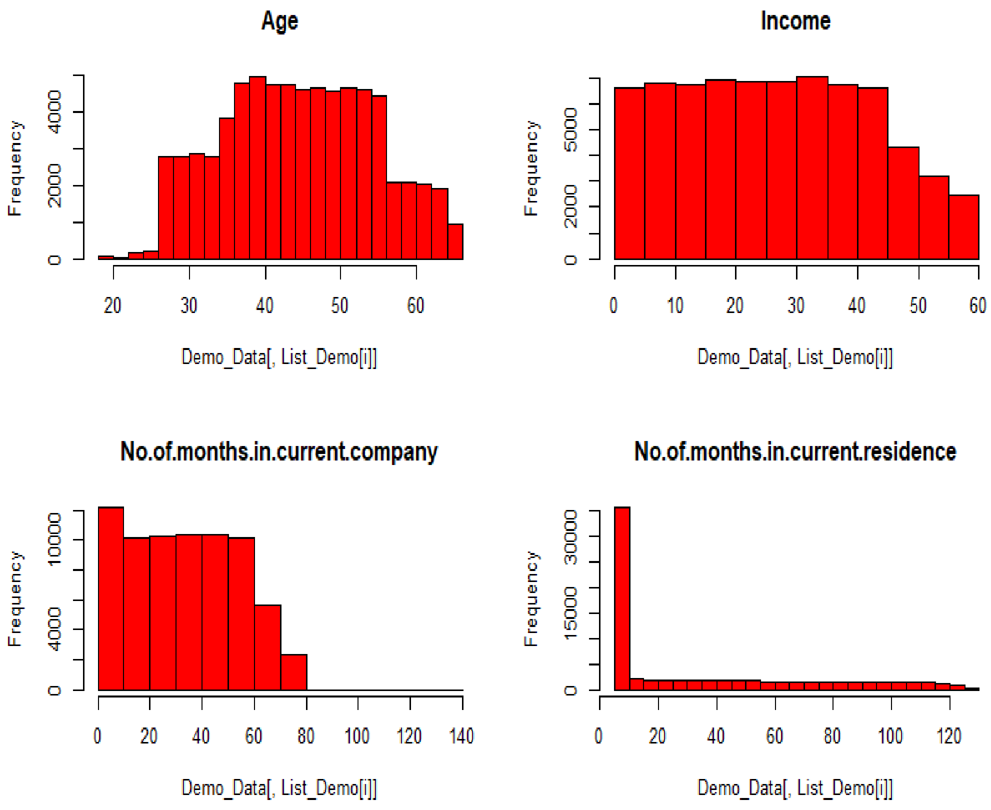
5. Customers with No Performance tag are considered as rejected population and this subset of customers will be used for model evaluation.

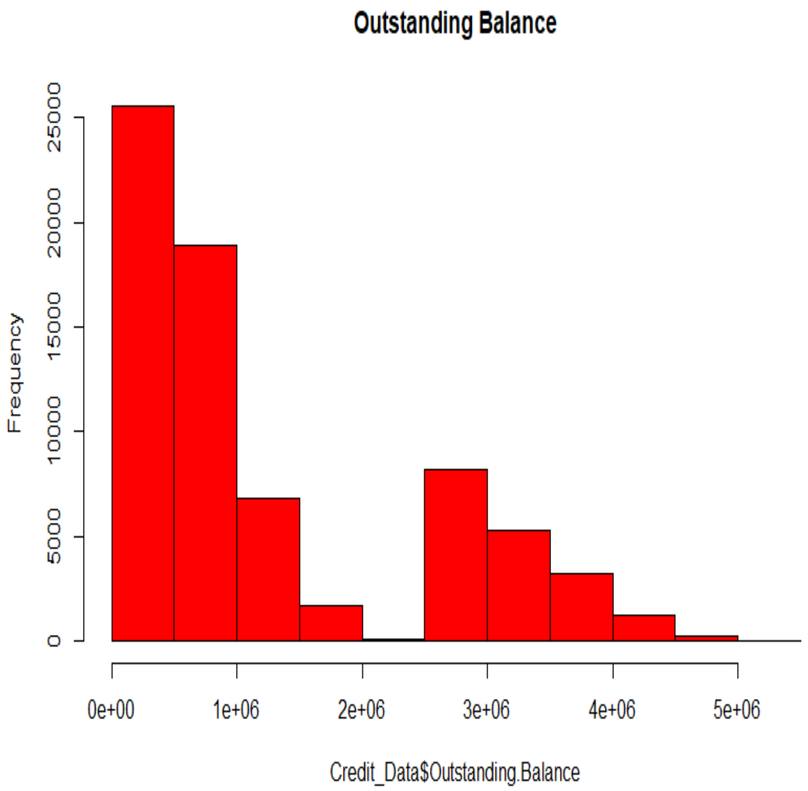
6. One record of "No.of.trades.opened.in.last.6. months" has missing value which is imputed with the mean value.

7. Customers with missing "Presence of home loan" & "Outstanding balance" variable are replaced with 0, considering the fact that the other variables from the credit bureau dataset for these customers are zero and they have relevant demographic data.

Univariate Analysis

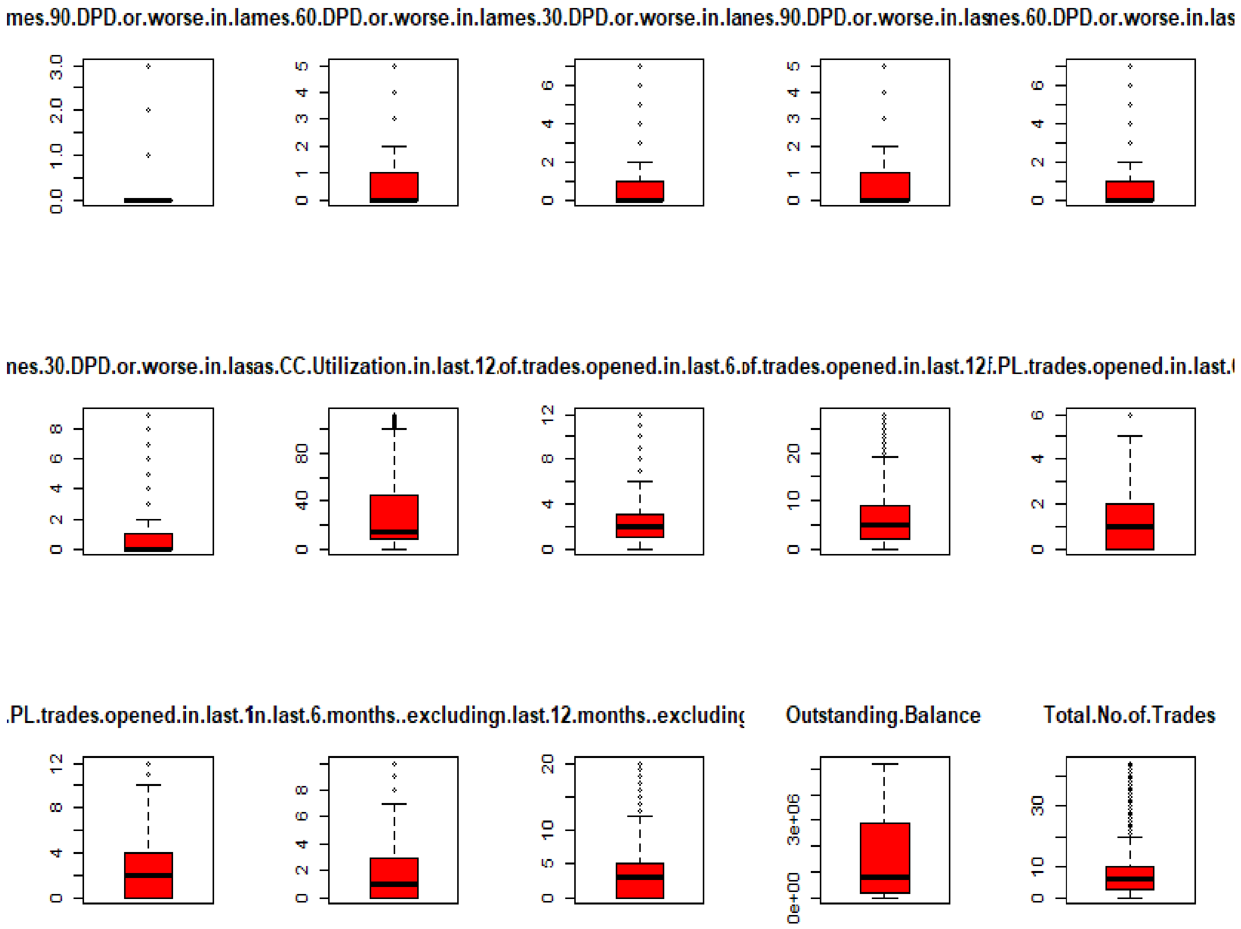
Data Distribution: No of months in Current Residence and Outstanding Balance has right skewed distribution

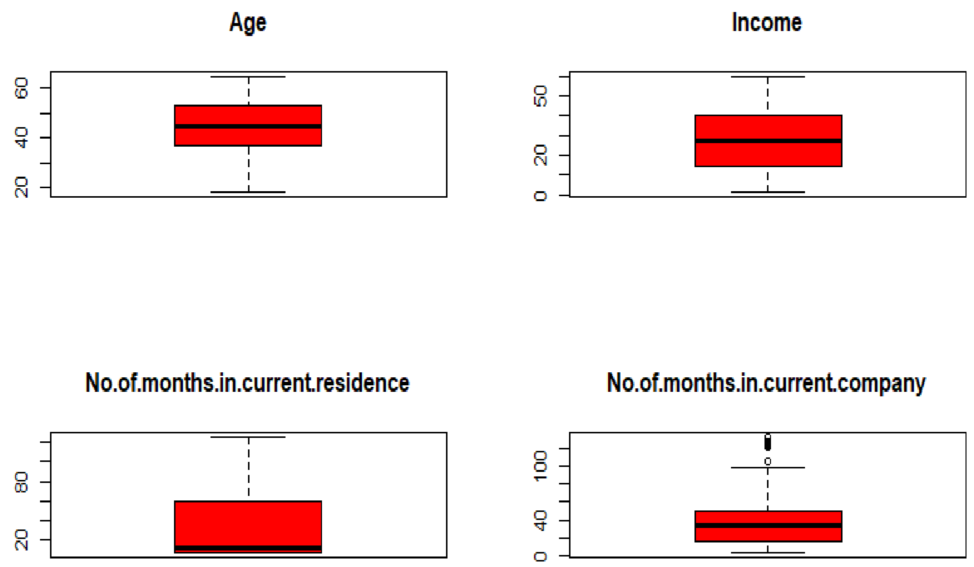




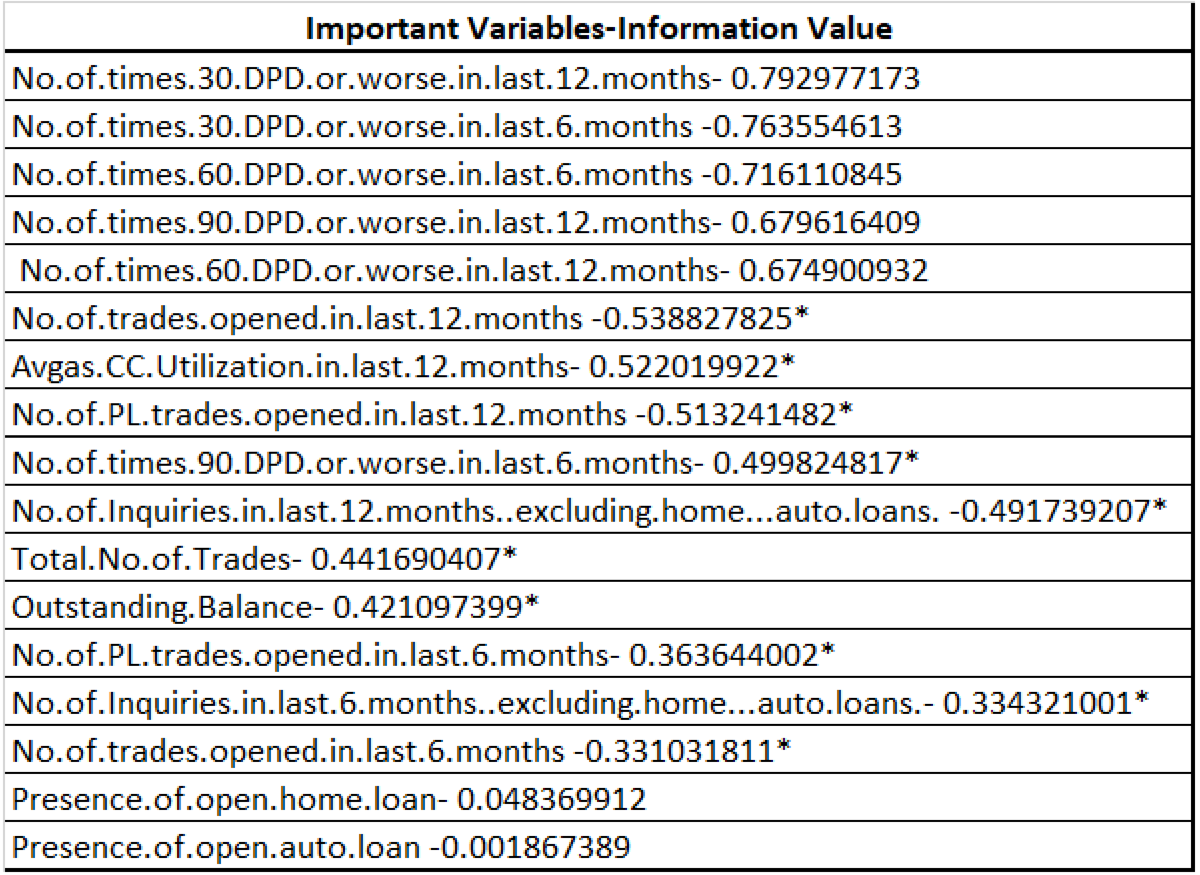
Univariate Analysis-Continued…….

Outliers- Variables with Outliers (majorly from Credit Bureau data) in the dataset are binned to eliminate the impact of outliers on model





Important Variables- Information Value

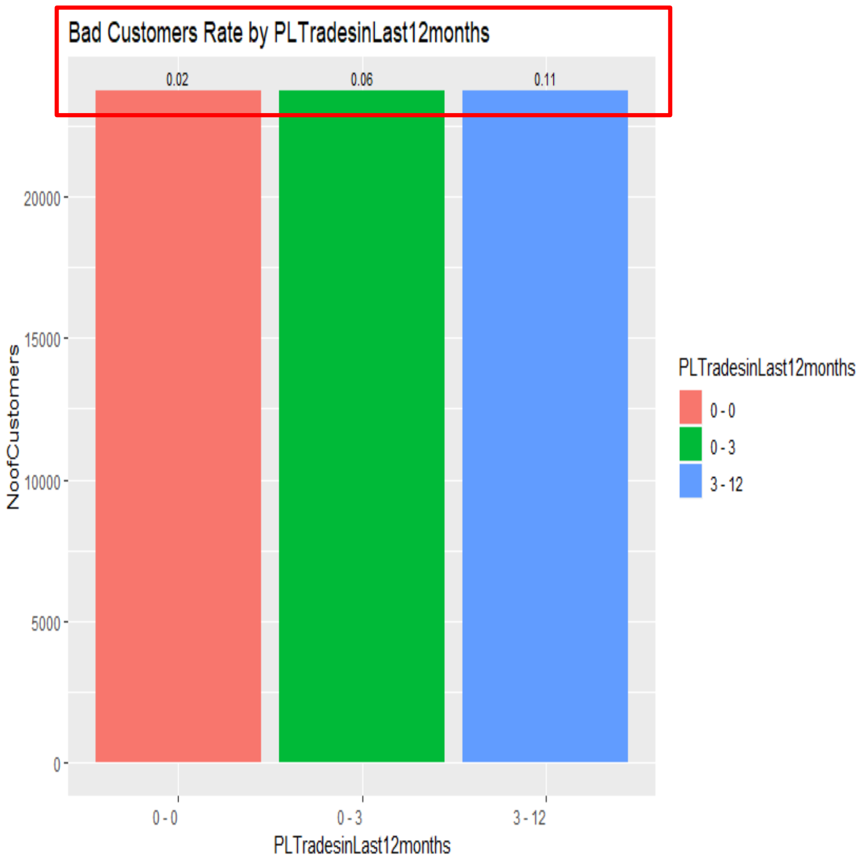


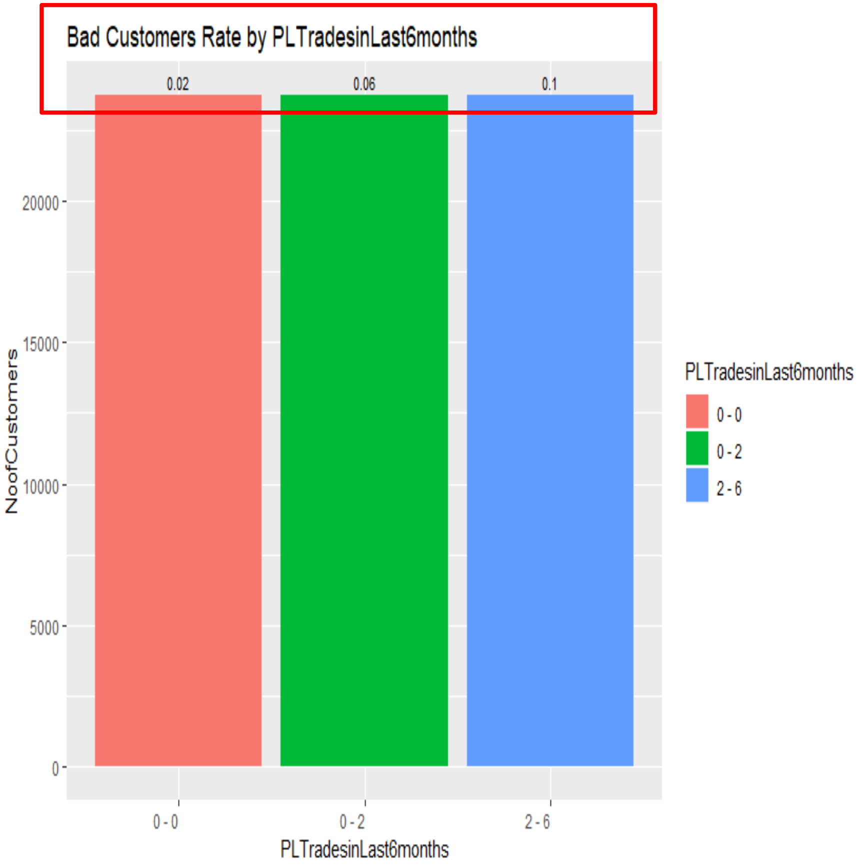
Important Variables-Information Value
No.of.times.30.DPD.or.worse.in.last.12.months- 0.792977173
No.of.times.30.DPD.or.worse.in.last.6.months -0.763554613
No.of.times.60.DPD.or.worse.in.last.6.months -0.716110845
No.of.times.90.DPD.or.worse.in.last.12.months- 0.679616409
No.of.times.60.DPD.or.worse.in.last.12.months- 0.674900932
No.of.trades.opened.in.last.12.months -0.538827825*
Avgas.CC.Utilization.in.last.12.months- 0.522019922*
No.of.PL.trades.opened.in.last.12.months -0.513241482*
No.of.times.90.DPD.or.worse.in.last.6.months- 0.499824817*
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. -0.491739207*
Total.No.of.Trades- 0.441690407*
Outstanding.Balance- 0.421097399*
No.of.PL.trades.opened.in.last.6.months- 0.363644002*
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.- 0.334321001*
No.of.trades.opened.in.last.6.months -0.331031811*
Presence.of.open.home.loan- 0.048369912
Presence.of.open.auto.loan -0.001867389

Data Exploration

Exploratory Analysis to find relation between Important Predictors and Dependent Variable(Performance Tag)

Increase in Bad Customers Rate with the Increase in PL Trades

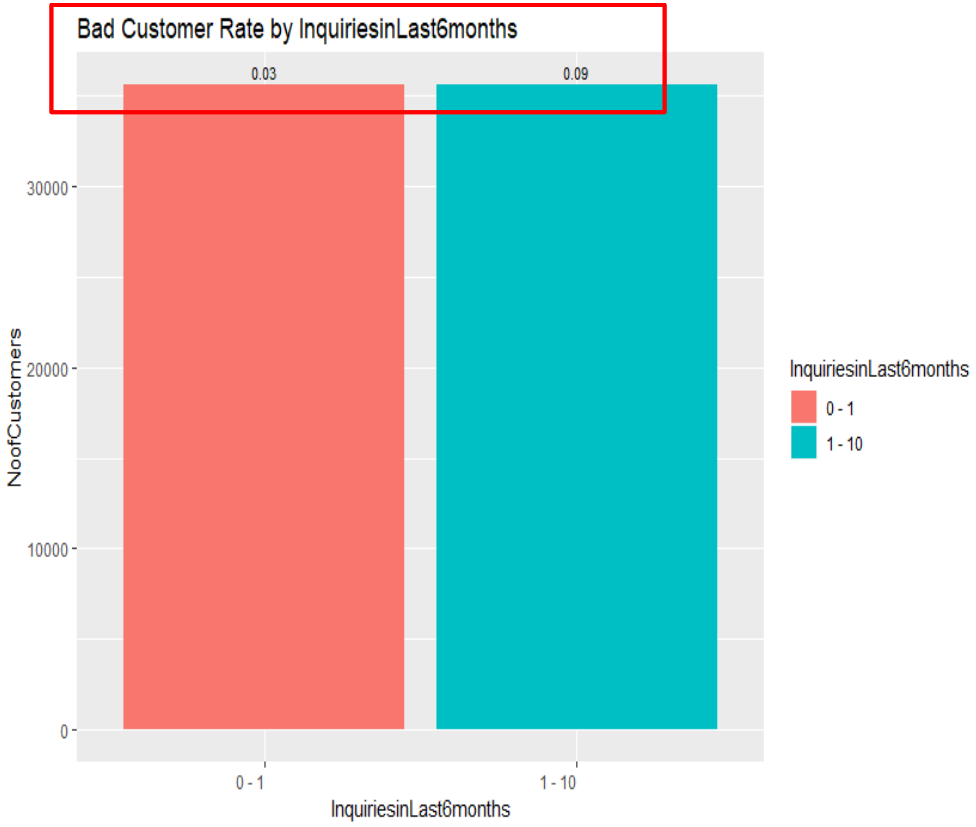




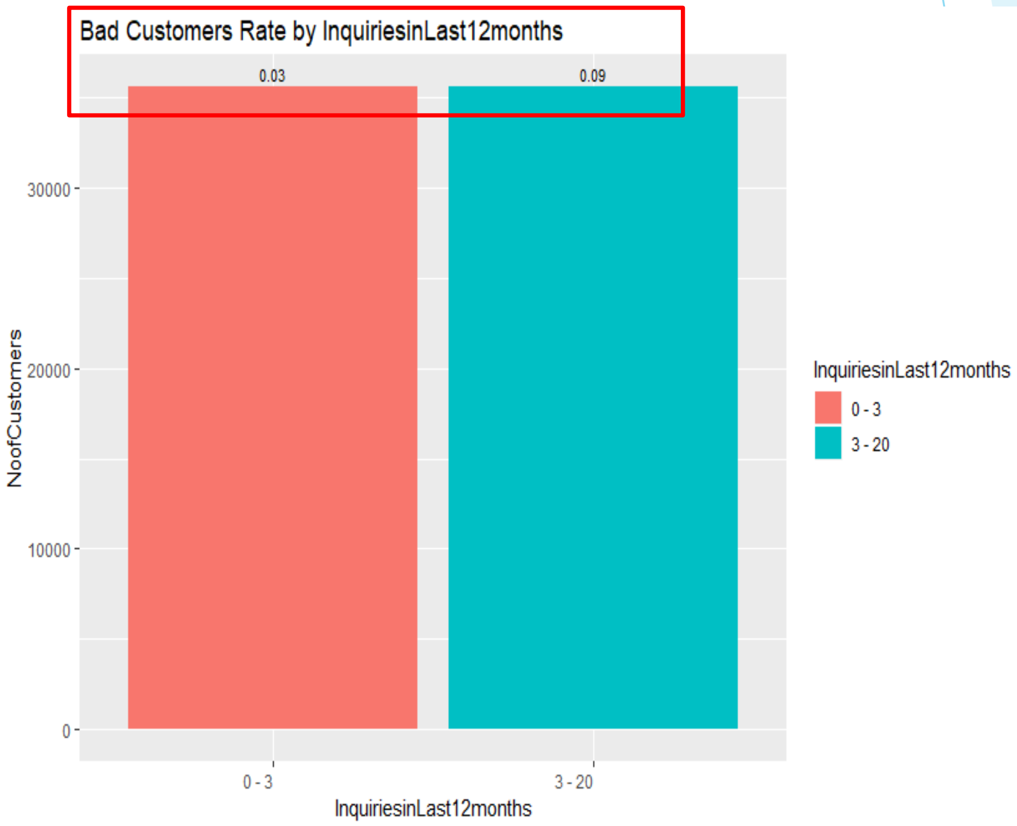
Data Exploration

Exploratory Analysis to find relation between Important Predictors and Dependent Variable(Performance Tag)

Increase in Bad Customers Rate with the Increase in No of Inquiries



Bad Customer Rate by InquiriesinLast6months

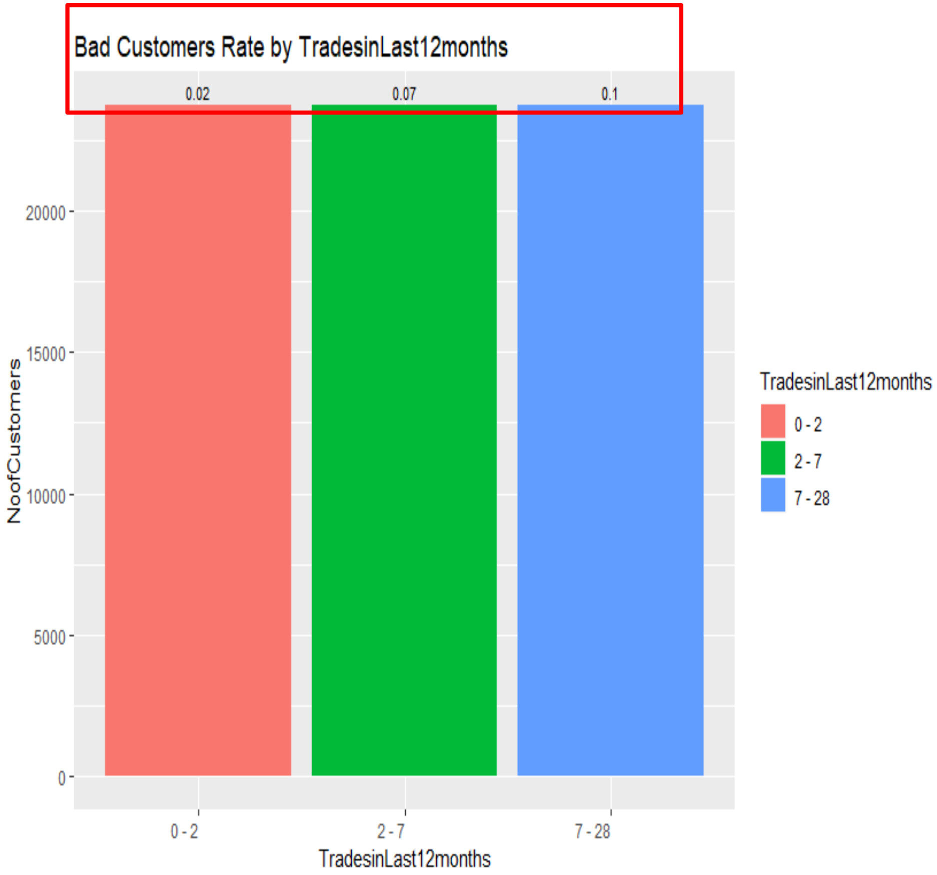


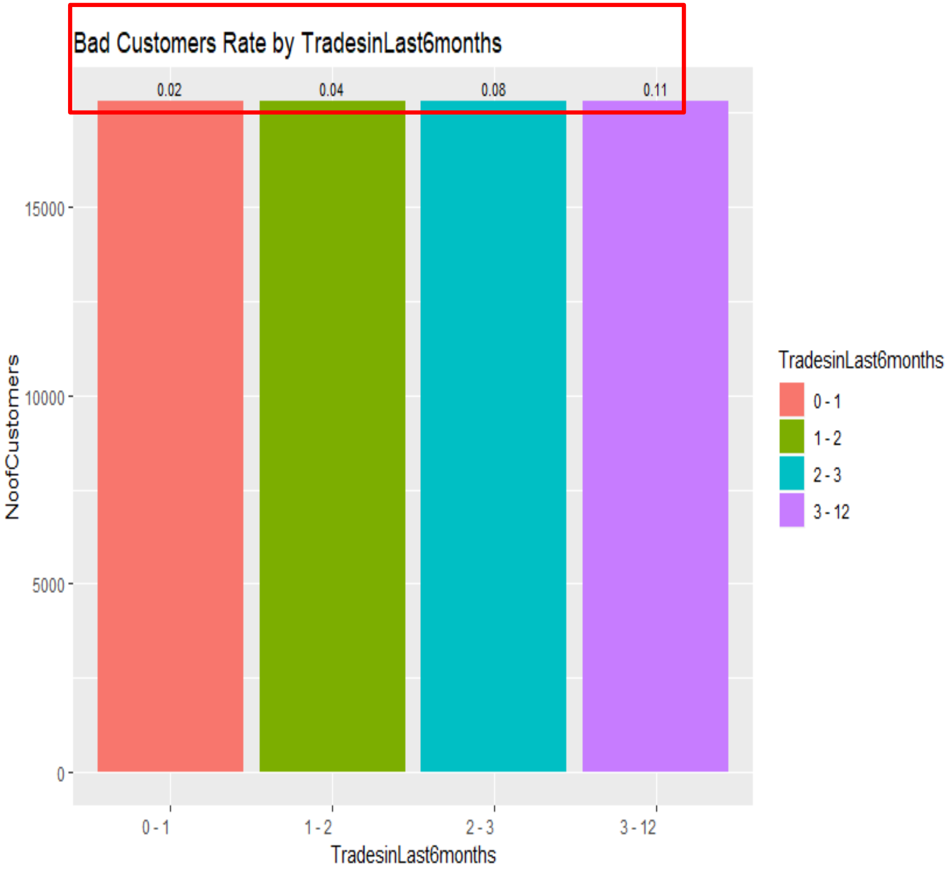
Bad Customers Rate by InquiriesinLast12months

Data Exploration

Exploratory Analysis to find relation between Important Predictors and Dependent Variable(Performance Tag)

Increase in Bad Customers Rate with the Increase in Trades

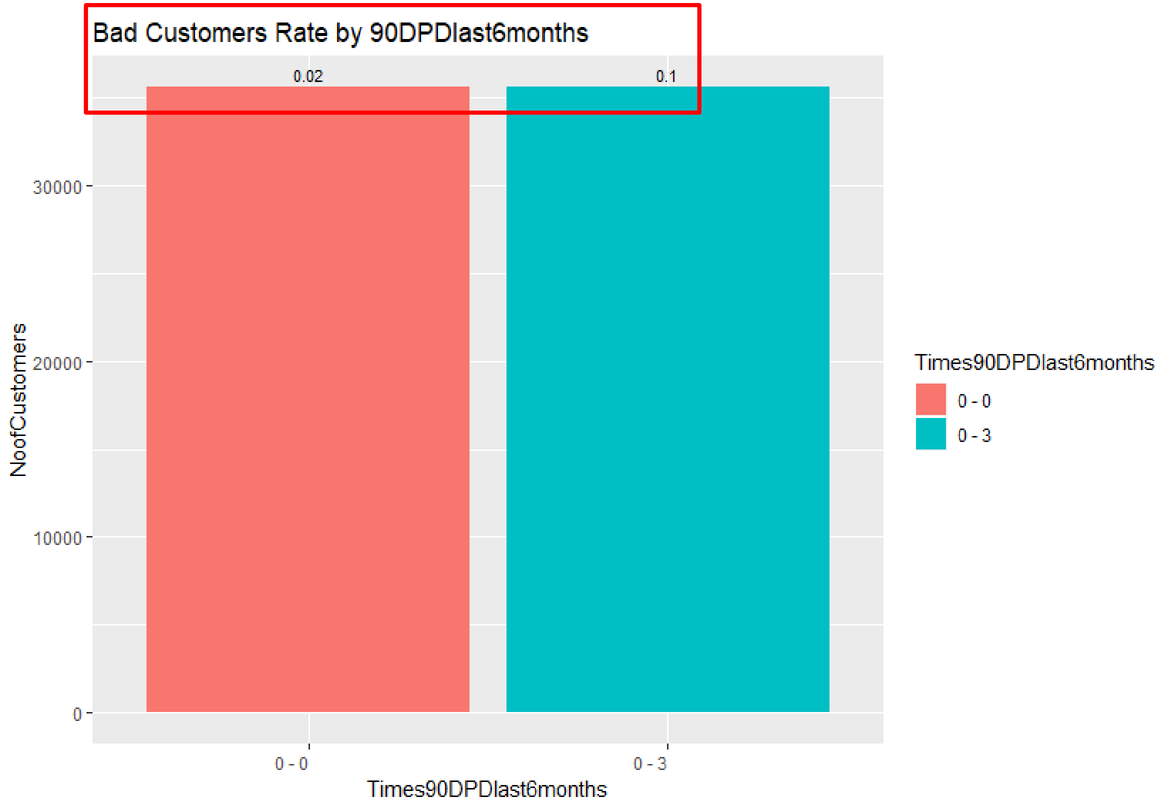




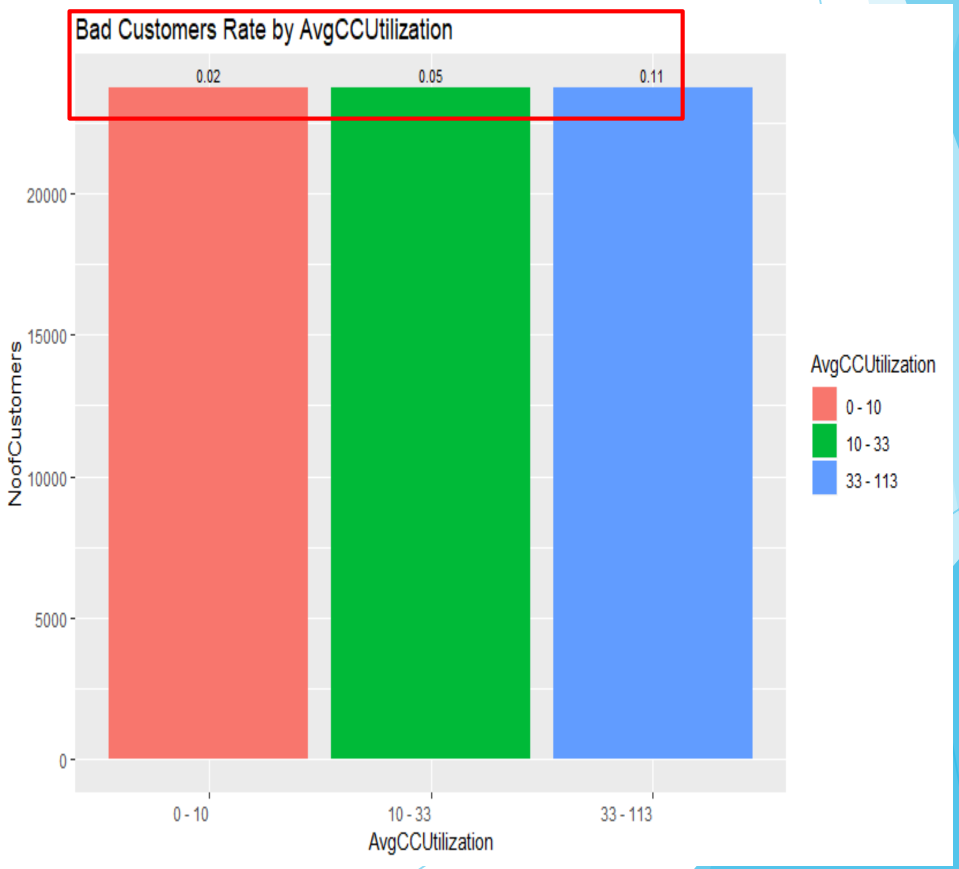
Data Exploration

Exploratory Analysis to find relation between Important Predictors and Dependent Variable(Performance Tag)

Increase in Bad Customers Rate with the Increase in number of times 90 days past due and Average CC Utilization



Bad Customers Rate by 90DPDlast6months

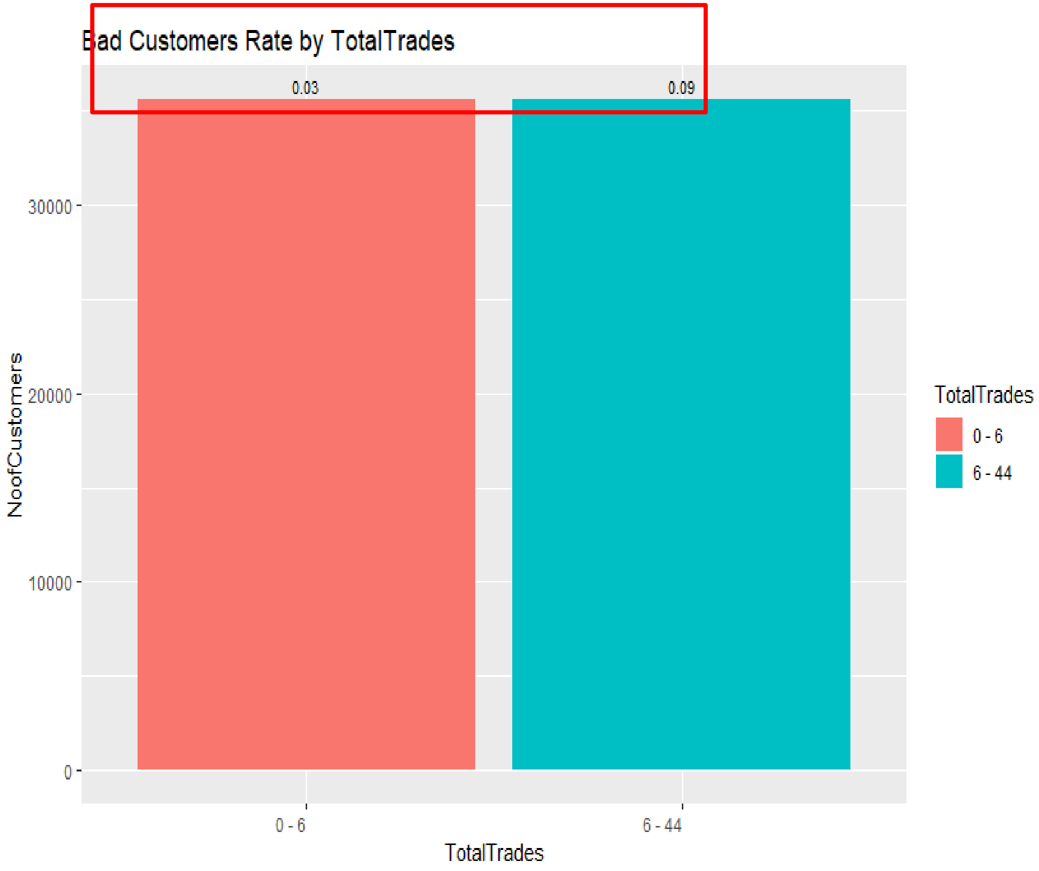


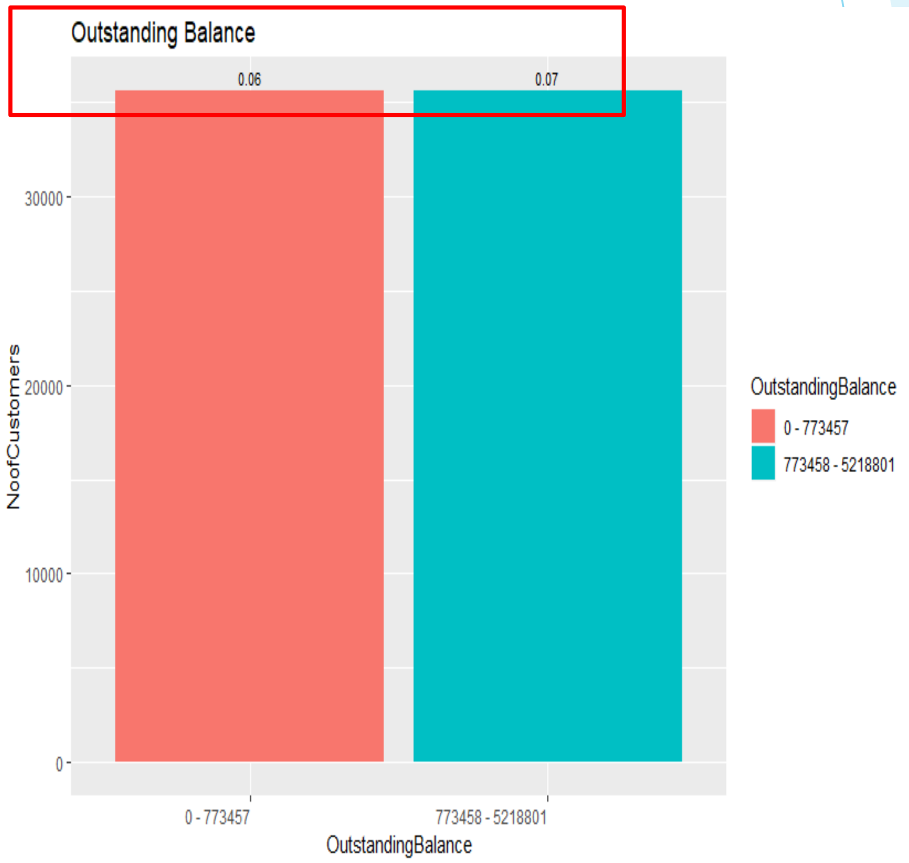
Bad Customers Rate by AvgCCUtilization

Data Exploration

Exploratory Analysis to find relation between Important Predictors and Dependent Variable(Performance Tag)

Increase in Bad Customers Rate with the Increase Total Trades and Outstanding Balance





Model Building

Logistic Model on Demographic Data vs Logistic Model on Demographic and Credit Bureau Data : -

As per Information Value, variables from Demographic data are not significant with respect to the predictive power for Customer Performance. Hence, Final Model is built on Demographic and Credit Data together.

Three models are built on Demographic and Credit Bureau data

1. Logistic Regression-Logistic Model is built using Weight of Evidences of Important Variables identified by Information Value Analysis.

2. Random Forest- Random Forest Model is improved by tuning the parameters like mtry = 6.
   Important Variables identified by Randome Forest- Age,Income, No.of.months.in.current.residence, Outstanding.Balance, Education, Profession, No.of.months.in.current.company, Type.of.residence, No.of.dependents,Gender, AvgCCUtilization, Marital Status, InquiriesinLast12months, TradesinLast6months.

3. Support Vector Machine-Support Vector Machine Linear Model is tuned using cost parameter at 0.05. This Cost parameter is identified using grid search and cross validation technique.

4. Stability for all the models is assessed using Cross validation technique on top of model evaluation for train and test data.

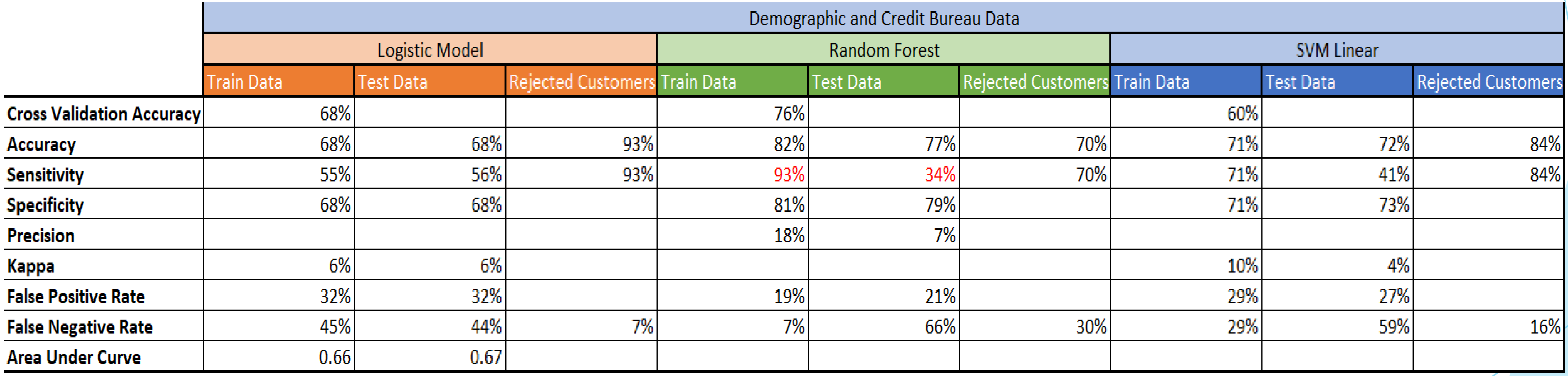
Model Evaluation

Models built are evaluated based on below metrics in order to choose the final model

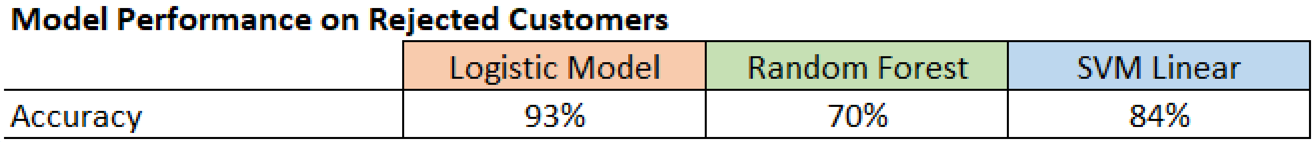
1. Stability-The cross-validation technique is used to understand the stability of the model. The model is evaluated on the rejected population in order to assess the results whether they correspond to the expectations.

2. Accuracy- Accuracy of the model when compared with reality

3. Discriminatory Power- Understand the discriminatory power of the model using metrics/ parameters like confusion metrics, Sensitivity, Specificity, Precision, Kappa metrics



	Demographic and Credit Bureau Data								
	Logistic Model			Random Forest			SVM Linear		
	Train Data	Test Data	Rejected Customers	Train Data	Test Data	Rejected Customers	Train Data	Test Data	Rejected Customers
Cross Validation Accuracy	68%			76%			60%		
Accuracy	68%	68%	93%	82%	77%	70%	71%	72%	84%
Sensitivity	55%	56%	93%	93%	34%	70%	71%	41%	84%
Specificity	68%	68%		81%	79%		71%	73%	
Precision				18%	7%				
Kappa	6%	6%					10%	4%	
False Positive Rate	32%	32%		19%	21%		29%	27%	
False Negative Rate	45%	44%	7%	7%	66%	30%	29%	59%	16%
Area Under Curve	0.66	0.67							



Model Performance on Rejected Customers

	Logistic Model	Random Forest	SVM Linear
Accuracy	93%	70%	84%

Model Fine tuning

Logistic Model- Final Logistic Model on Demographic and Credit Bureau data shows Accuracy and Specificity @ 68%. Sensitivity is @ 56%. The model average accuracy using cross validation technique is 68%. This model has stability going by the performance parameters of the model during the testing of the model across train, test data and using cross validation technique. Logistic Model has performed really well on Rejected population. It classifies 93% of Rejected Customers correctly

Random Forest Model- Random Forest Model shows Accuracy @77% and Specificity @ 79%. However, the model is bias on Train data for sensitivity (93% sensitivity on train data vs 34% sensitivity on test data). This Model average accuracy using cross validation technique is 76%. Model is underperforming to classify rejected customers correctly when compared with logistic model (70% accuracy).

SVM Model- SVM Model shows accuracy @72% and Specificity @ 73%. However, the model is bias on Train Data for sensitivity(71% sensitivity on train data vs 41% Sensitivity on test data). Model is underperforming to classify the rejected customers when compared with logistic regression(84% accuracy)

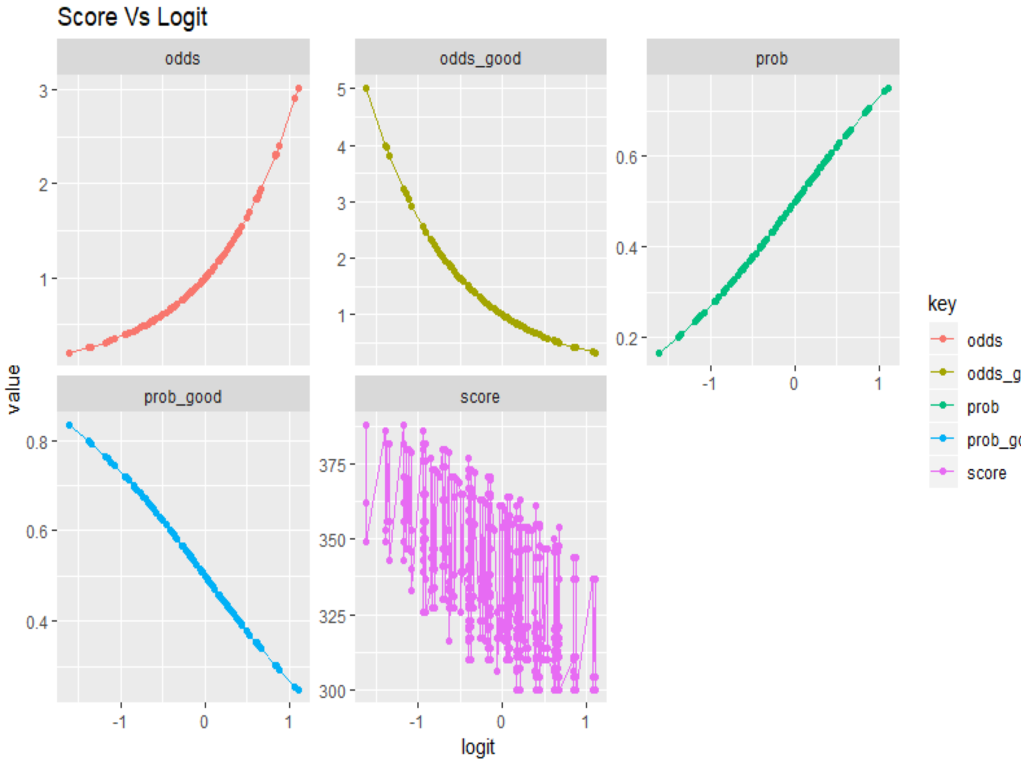
Hence, going by the metrics like Stability, Accuracy and Discriminatory power, Logistic Model meets all the required parameters compared to Random Forest and SVM Model. SVM and Random Forest performs well on data for accuracy. However, these models are bias on train data for sensitivity. Hence, **Logistic Model is considered as Final Model**.

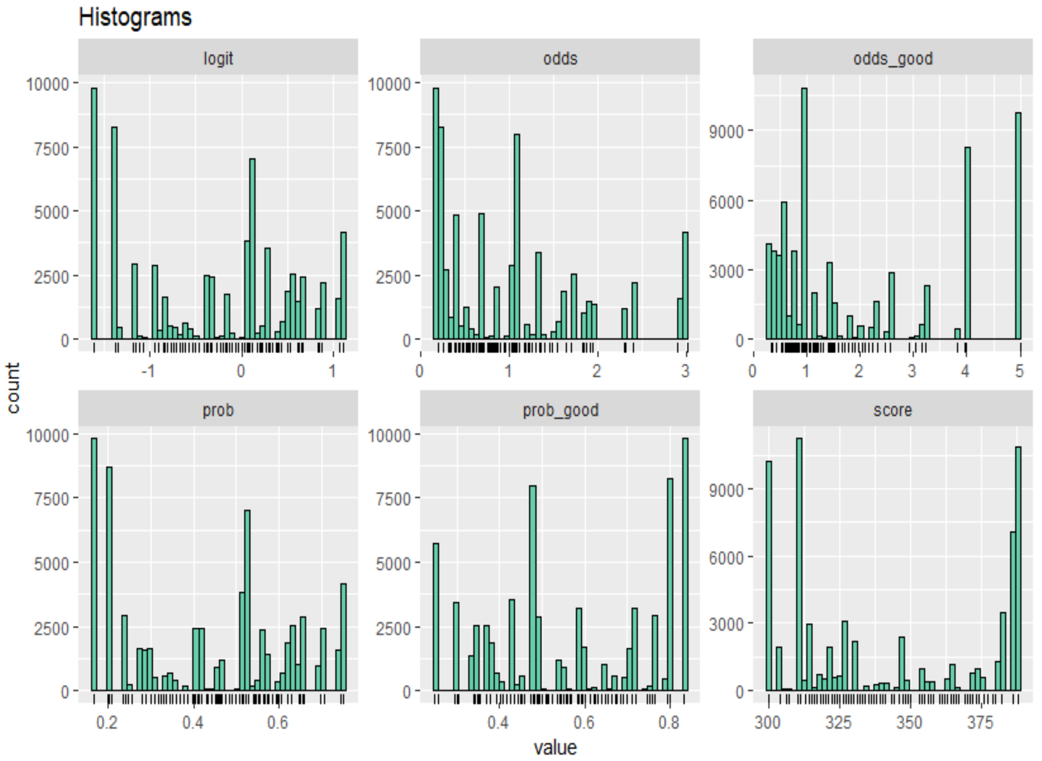
Application Scorecard

The application scorecard is built with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points in order to identify the Good Customers for business and reduce the potential credit loss

- **The cut-off score below which credit cards should not be granted to applicants:-**The cut off score is identified as 330 score where 50% of the customers are classified as good customers in the same way as on the balanced data with 50% Bad Customers rate.

- **Score Vs Odds of Good:-**We can see that the score is perfectly linearly correlated with the logit. With the increase in odds for Good, we can also see the increase in Score.

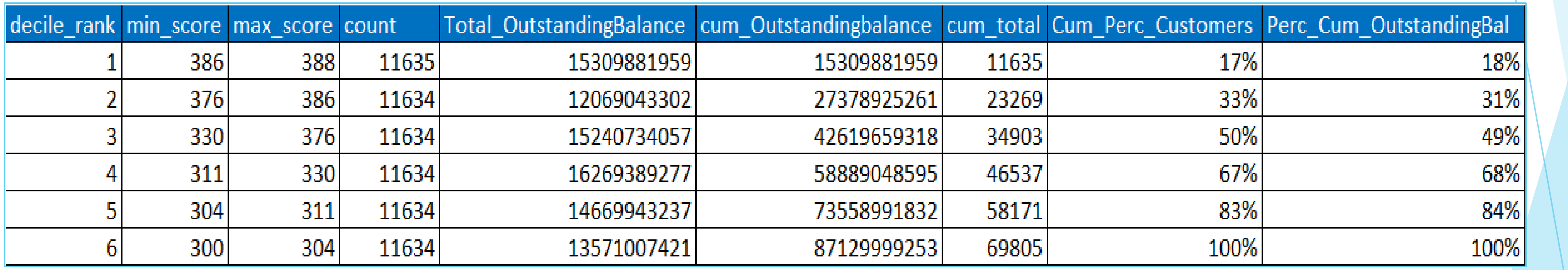

Score Vs Logit


Histograms

Application Scorecard(Continued…..)

- If we filter/classify the customers with 330 or higher score, we classify 50% of the overall customers as good customers . The model is able to classify 51% of Good Customers and 75% of the Bad Customers correctly.
- However, model is classifying 49% of Good Customers as Bad Customers.

No of Customers and Outstanding Balance by Score grouping



decile_rank	min_score	max_score	count	Total_OutstandingBalance	cum_Outstandingbalance	cum_total	Cum_Perc_Customers	Perc_Cum_OutstandingBal
1	386	388	11635	15309881959	15309881959	11635	17%	18%
2	376	386	11634	12069043302	27378925261	23269	33%	31%
3	330	376	11634	15240734057	42619659318	34903	50%	49%
4	311	330	11634	16269389277	58889048595	46537	67%	68%
5	304	311	11634	14669943237	73558991832	58171	83%	84%
6	300	304	11634	13571007421	87129999253	69805	100%	100%

Assessing the financial benefit of the project

The financial benefit of the model:

1. **The implications of using the model for auto approval or rejection, i.e. how many applicants on an average would the model automatically approve or reject :–**If 330 is used as minimum threshold score for Good customers then the Model will approve 50% of applications automatically. Model classify 51% of Good Customers and 75% of the Bad Customers correctly(75% Sensitivity and 51% specificity)

2. **The potential credit loss avoided with the help of the model :-** Since the model automatically identifies 75% of the bad customers pre hand, it will help to save 75% of the potential loss.

3. **Potential Loss of revenue due to the rejection of good customers:–** Since the model classifies 48% of Good Customers as Bad Customers, there is a possibility of 50% potential revenue loss . User of the model has to keep this in mind and try to identify the potential good customers from the slot of customers rejected by model on the basis of domain expertise or stand alone analysis for these customers using the important/significant predictors of Customer performance.

4. **Assumptions based on which the model has been built :**
   1. Final Model (Logistic Model) is selected based on the performance metrics like Accuracy, Discriminatory Power, Stability and performance of the model on Rejected Customers
   2. Model is built on WOEs of Important variables identified using Information Value analysis
   3. The data provided is imbalanced. Model built on imbalanced data can be biased. Hence, the data is balanced using SMOTE function. Balanced data has 50% Good customers and 50% Bad Customers.
   4. Score threshold is selected based on the trade off between Actual Bad Customers vs Identified Bad Customers using the Score threshold. The business objective to avoid/reduce credit loss has been given the weightage to select the Score threshold.