BFS Capstone Project

Risk Analytics Case Study

Group Members

- Avineet
- Ritesh
- Akhilesh
- Anurag

Business Objective

Problem Statement:

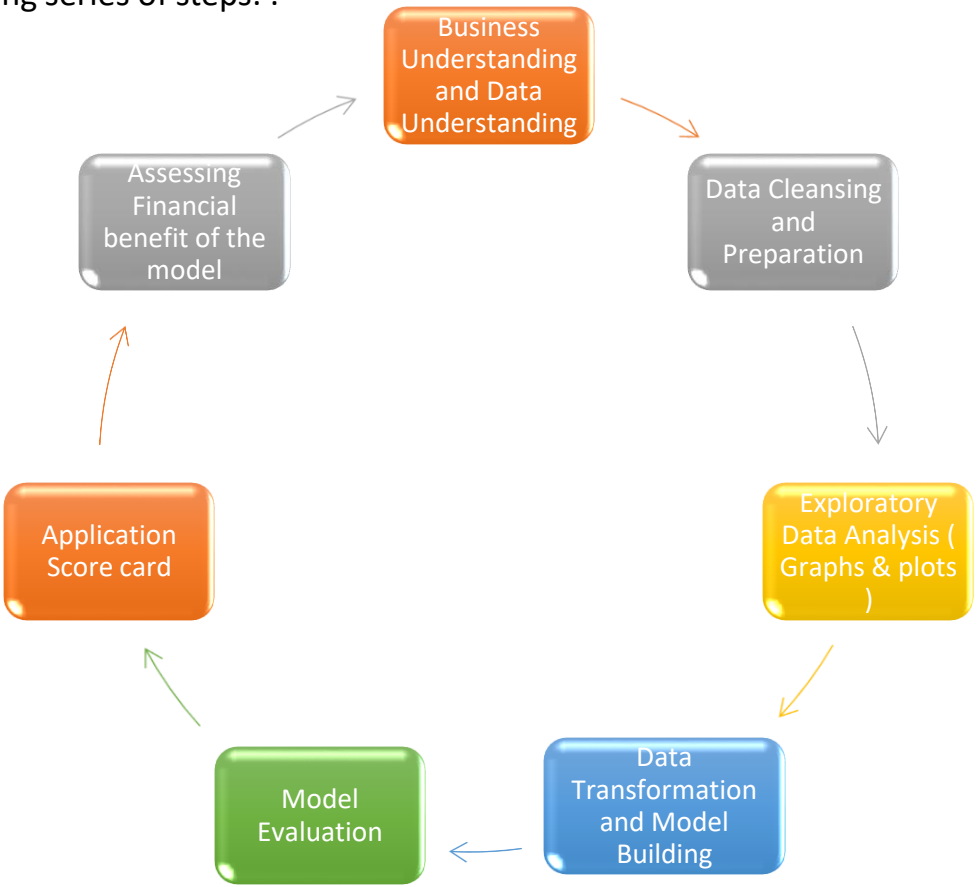
Credx is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss due to increase in defaults. The CEO believes that the best strategy to mitigate credit risk is to acquire "the right customers".

Objective:

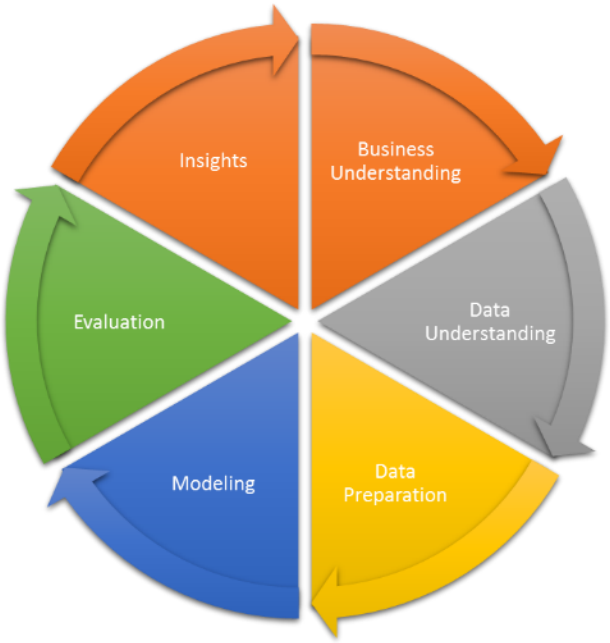
The objective of the project is to help CredX identify the right customers using predictive models, which are built using past data of the bank's applicants. We need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Approach

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: :



CRISP–DM framework



Data Understanding

The problem statement included following 2 data sets:

➢ Demographic Data

Includes 71,295 records and12 features

#	Variables	Data type
1	Application ID	Int64
2	Age	Int64
3	Gender	Object
4	Marital Status	Object
5	No of dependents	Float64
6	Income	Float64
7	Education	Object
8	Profession	Object
9	Type of residence	Object
10	No of months in current residence	Int64
11	No of months in current company	Int64
12	Performance Tag	Float64

➢ Credit Bureau data

Includes 71,295 records and19 features

#	Variables	Data Type
1	Application ID	int64
2	No of times 90 DPD or worse in last 6 months	int64
3	No of times 60 DPD or worse in last 6 months	int64
4	No of times 30 DPD or worse in last 6 months	int64
5	No of times 90 DPD or worse in last 12 months	int64
6	No of times 60 DPD or worse in last 12 months	int64
7	Avgas CC Utilization in last 12 months	float64
8	No of times 30 DPD or worse in last 12 months	float64
9	No of trades opened in last 6 months	int64
10	No of trades opened in last 12 months	int64
11	No of PL trades opened in last 6 months	int64
12	No of PL trades opened in last 12 months	int64
13	No of Inquiries in last 6 months (excluding home & auto loans)	int64
14	No of Inquiries in last 12 months (excluding home & auto loans)	int64
15	Presence of open home loan	float64
16	Outstanding Balance	float64
17	Total No of Trades	int64
18	Presence of open auto loan	int64
19	Performance Tag	float64

Approach followed for data cleaning

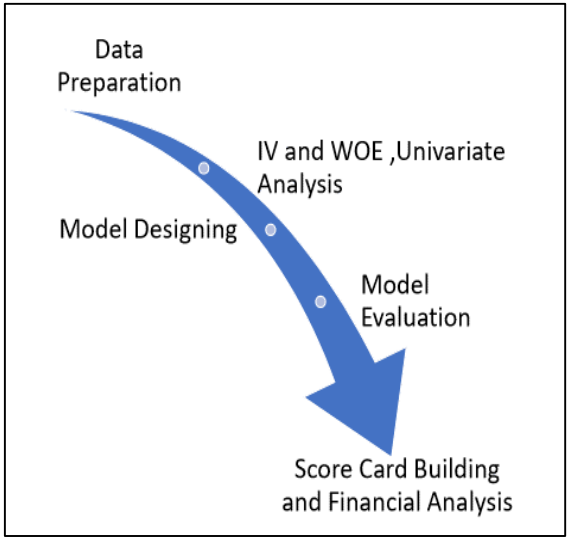
Data Preparation

- **Treatment of invalid entries**
  - Check the dataset for invalid entries eg negative income, age less than or equal to 0, duplicate application id etc. and delete such records.

- **Missing Value treatment**
  - Compute WoE for each variable and impute the missing values for those variables with that category of the variable to which the WoE of the null value is closest. For computing Woe, continuous variables will have to be converted into categories

- **Perform EDA**
  - Perform EDA and identify impact of each variable on default, which would be used for building model

- **Variable selection**
  - Compute Information value(IV) for each variable and select variables with high IV from both datasets for the purpose of building model.



Note:
For the purpose of easy computation, we have switched the indicators of Performance Tag, i.e. 1 for Good and 0 for Bad

Data Quality Issues

Based on our review of the dataset identified following data quality issues:

- 3 pair of duplicate application ids in both datasets
- 20 observations where age is less than 0
- 107 instances where income was less than 0
- 9 instance where CC utilization in last 12 months was zero, however performance tag indicated default. (the same is not possible as for default there has to be 90 or more DPD accordingly average utilization cannot be zero in such case).

All the above records were deleted and a merged data frame was created after removing records where performance tag was null.

WOE and IV Analysis

- WOE and IV values are calculated for each of the attributes using information
- Attributes for which WOE values that were made by using bin functions
- For 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
- 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.

	VAR_NAME	IV
8	No of months in current residence	0.052060
4	Income	0.037624
7	No of months in current company	0.012735
9	Profession	0.002217
10	Type of residence	0.000924
2	Education	0.000783
0	Age	0.000625
3	Gender	0.000325
5	Marital Status (at the time of application)	0.000095
6	No of dependents	0.000056
1	Application ID	0.000021

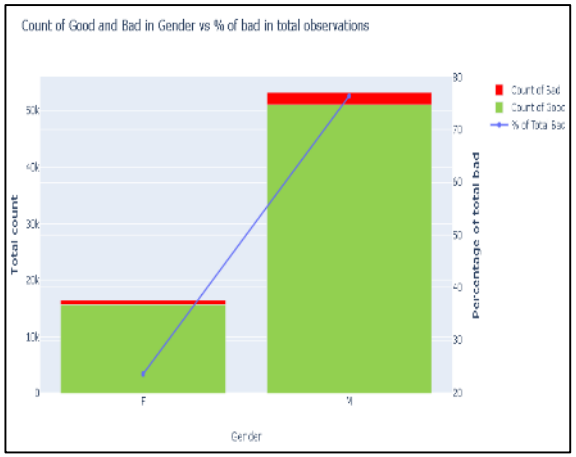
Exploratory Data Analysis

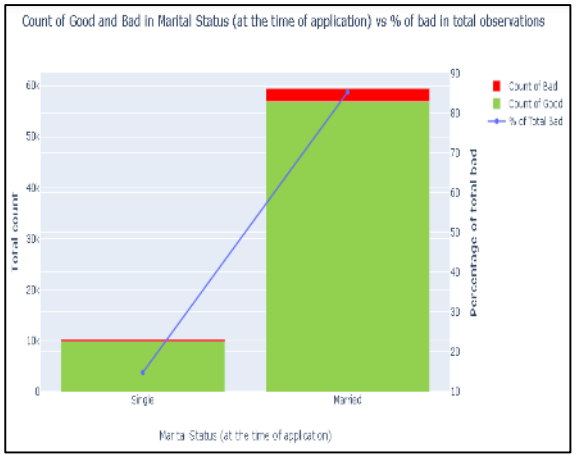
Gender

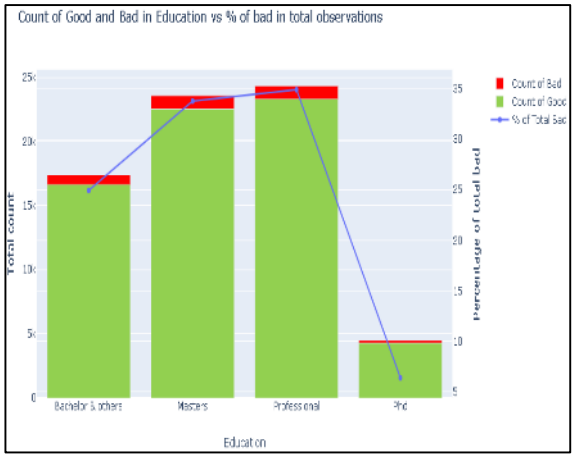
Marital Status

Education

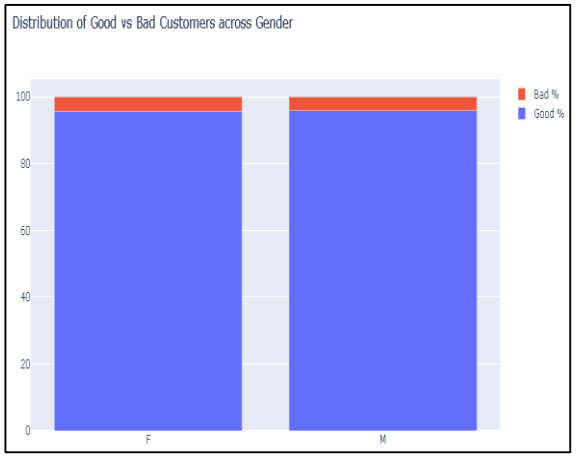
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

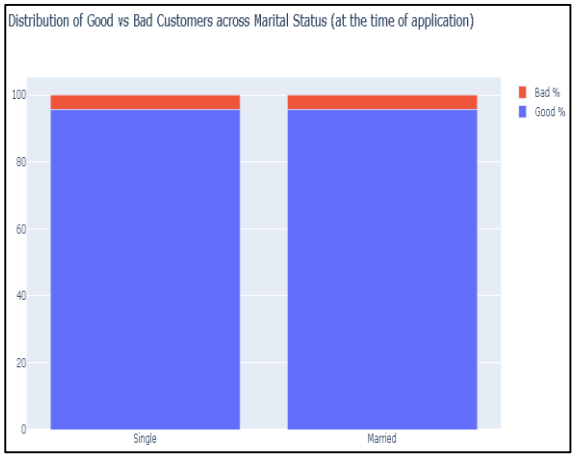


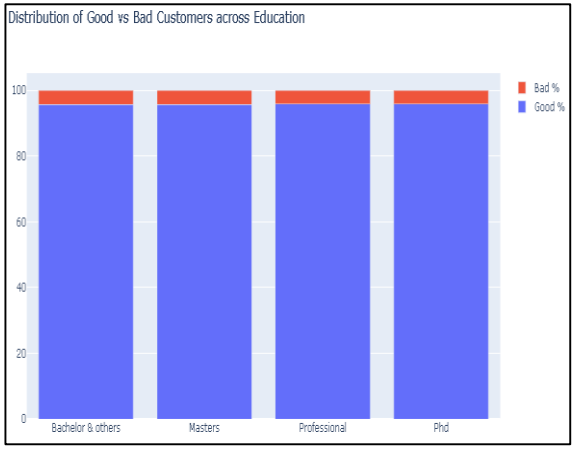




Graph showing percentage of good and bad customers in each of the categories of variable







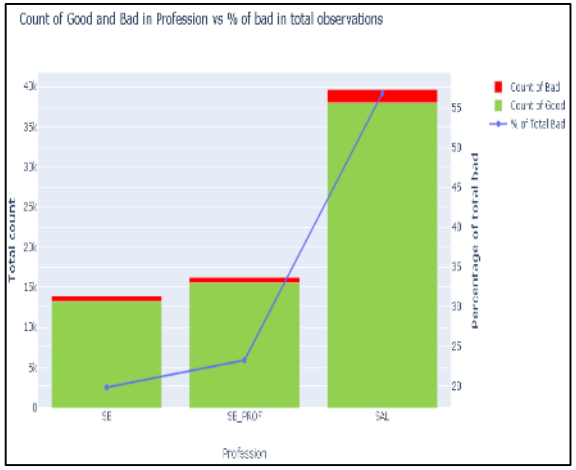
Exploratory Data Analysis

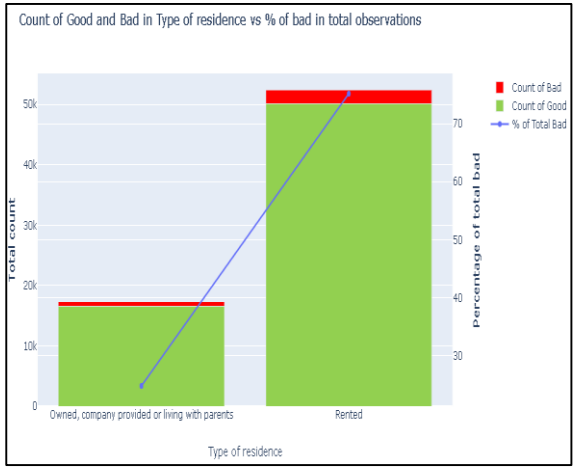
Profession

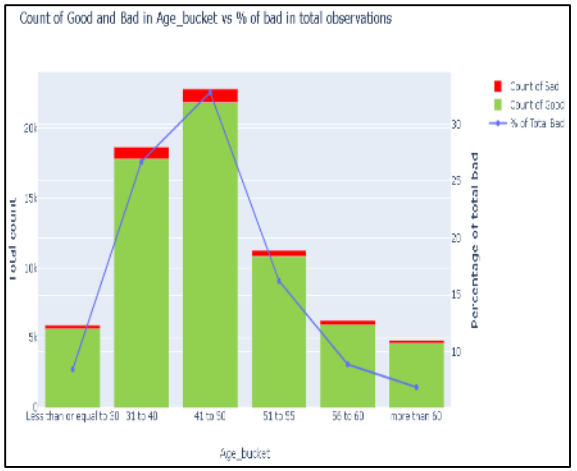
Type of residence

Age

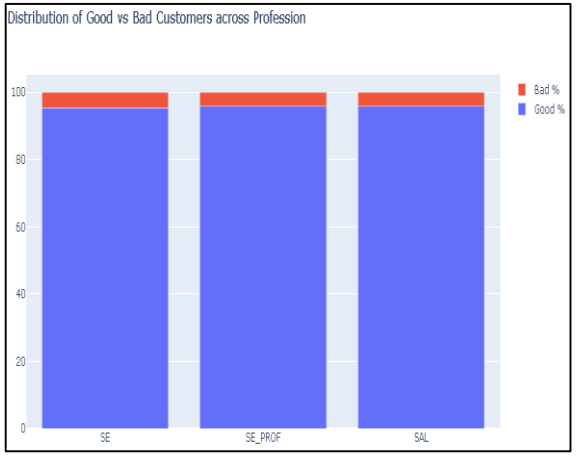
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

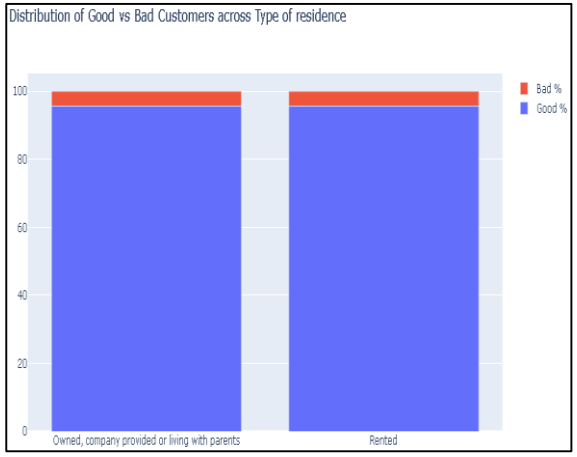


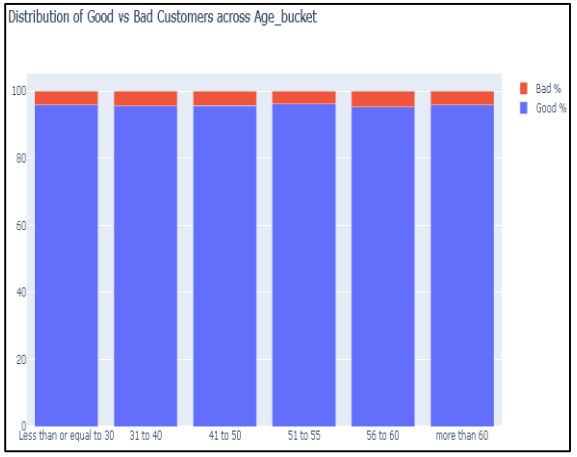




Graph showing percentage of good and bad customers in each of the categories of variable







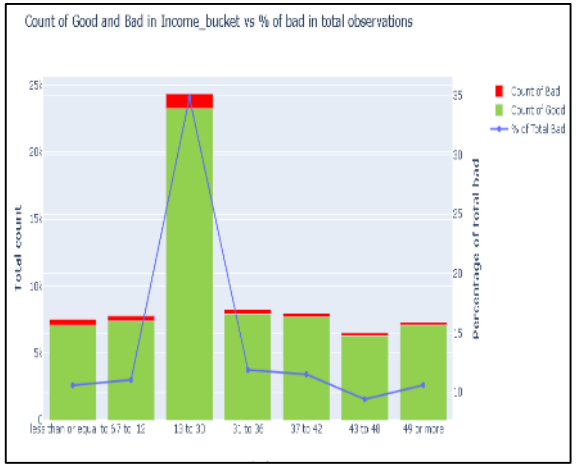
Exploratory Data Analysis

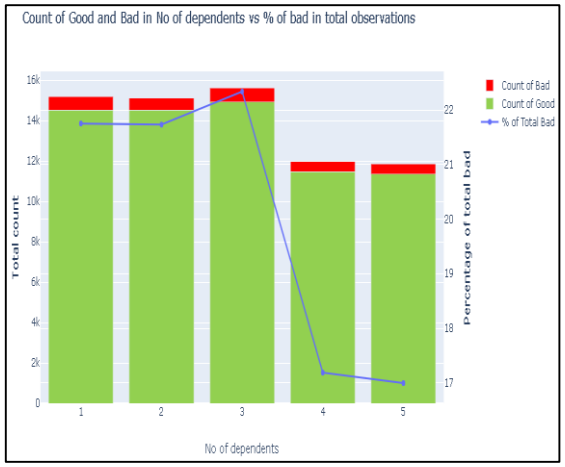
Income

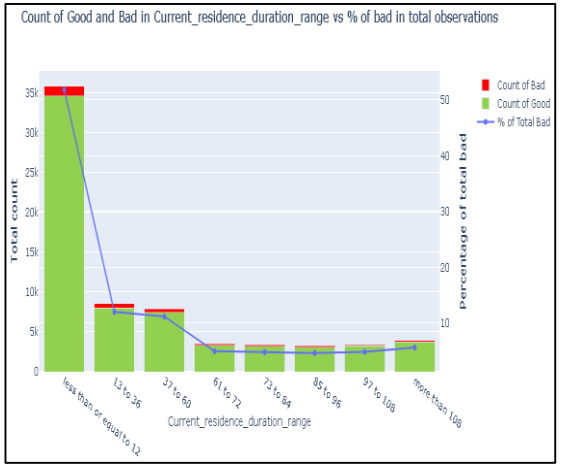
No of dependents

No of months in current residence

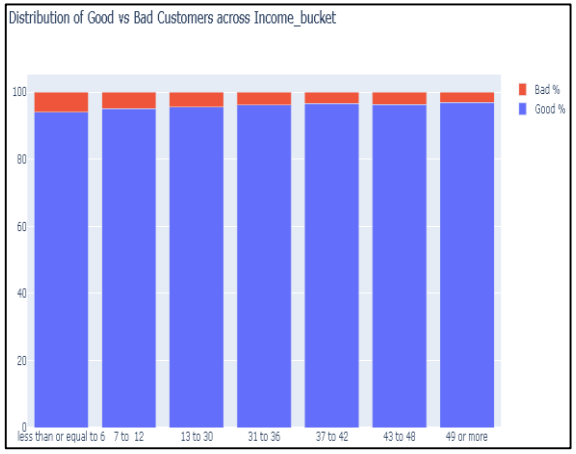
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

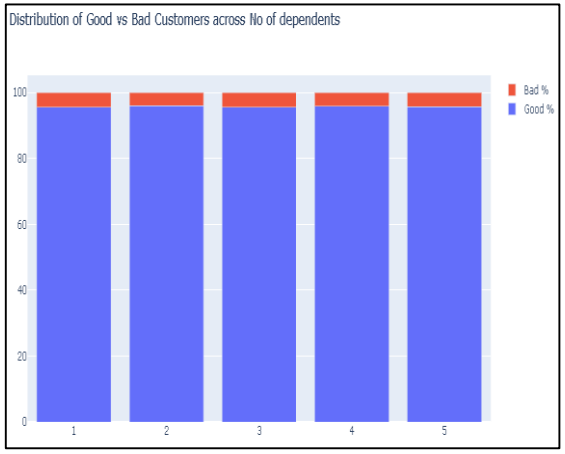


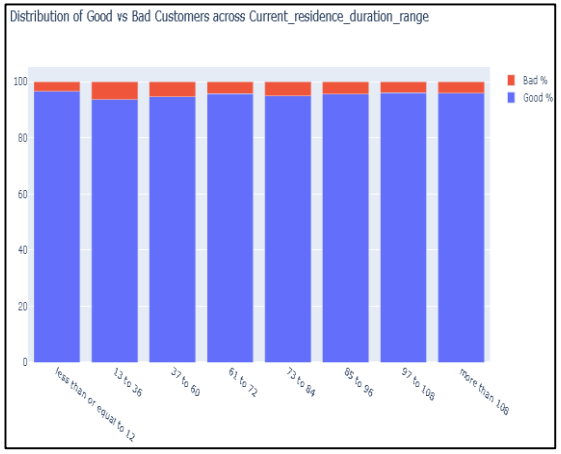




Graph showing percentage of good and bad customers in each of the categories of variable







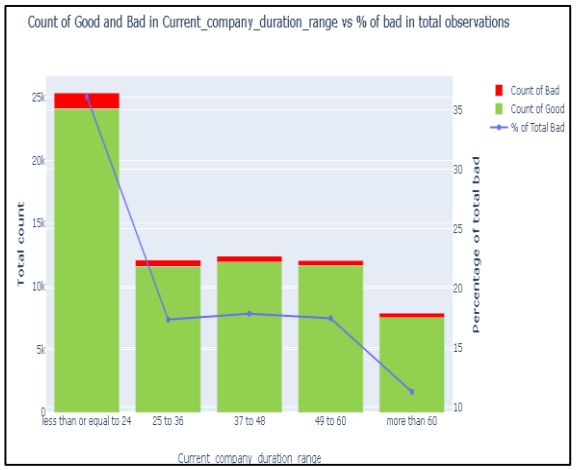
Exploratory Data Analysis

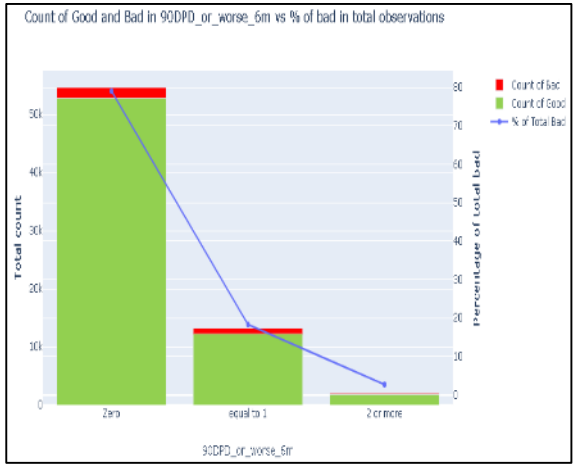
No of months in current company

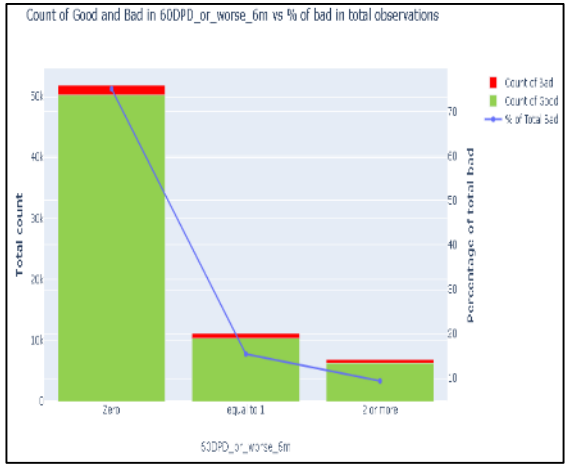
No of times 90 DPD or worse in last 6 months

No of times 60 DPD or worse in last 6 months

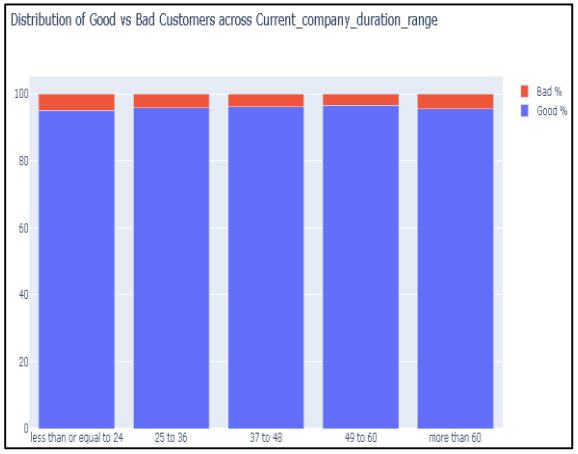
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

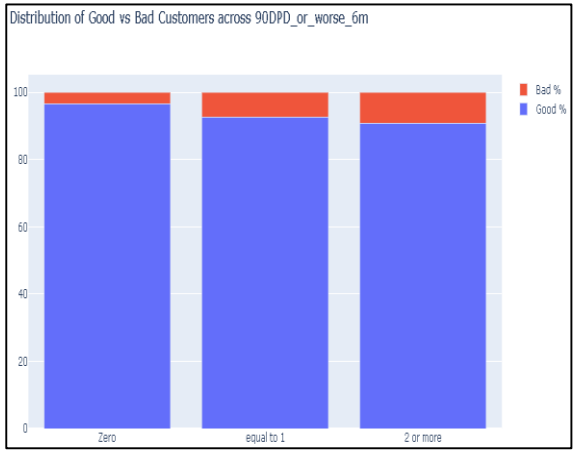


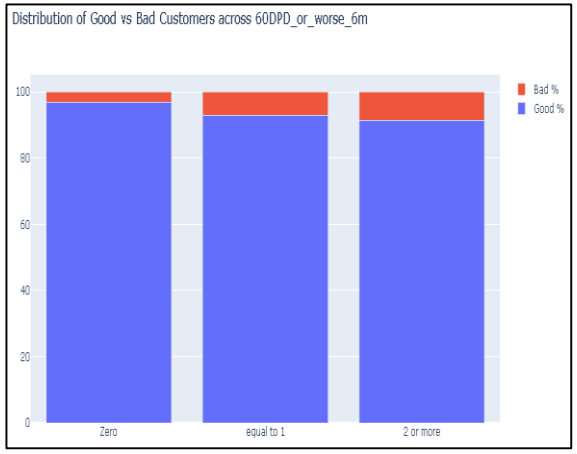




Graph showing percentage of good and bad customers in each of the categories of variable







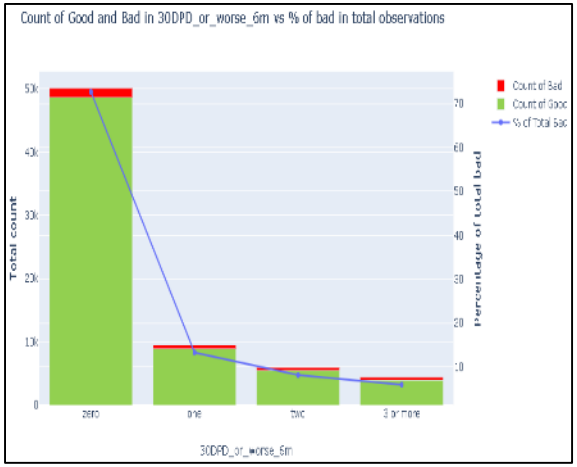
Exploratory Data Analysis

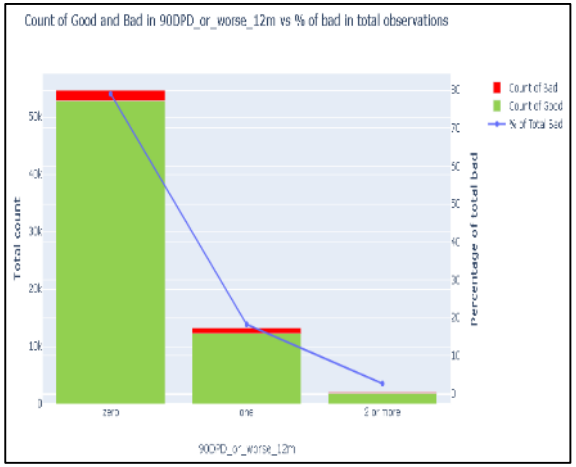
No of times 30 DPD or worse in last 6 months

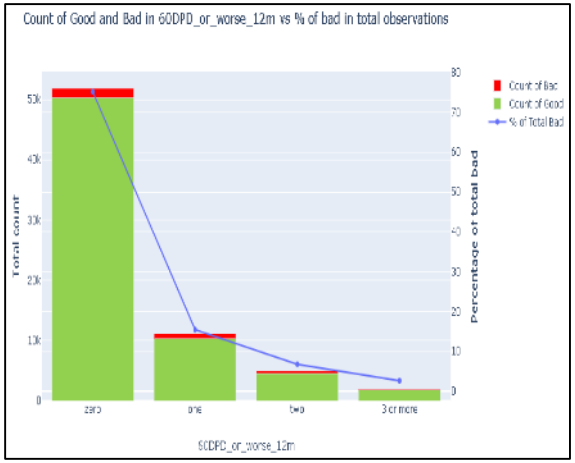
No of times 90 DPD or worse in last 12 months

No of times 60 DPD or worse in last 12 months

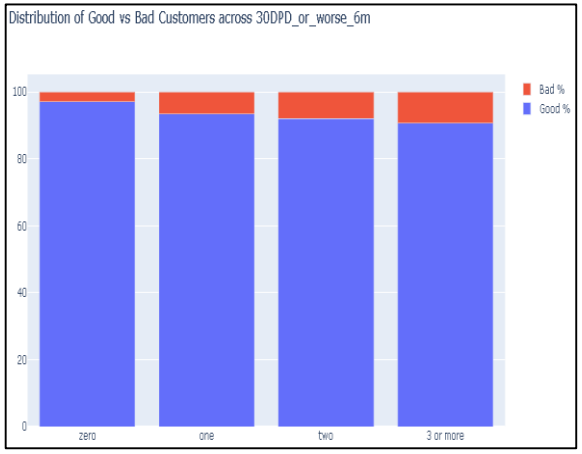
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

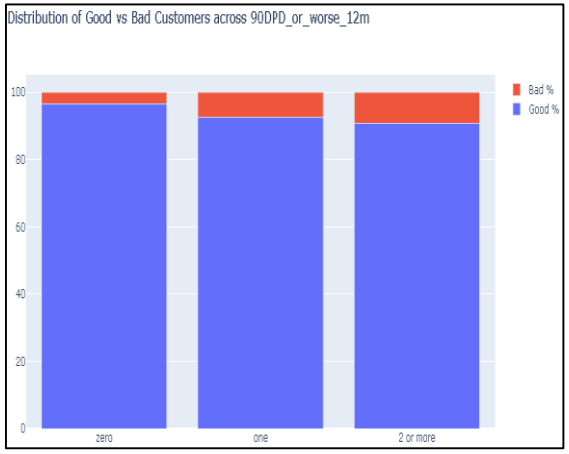


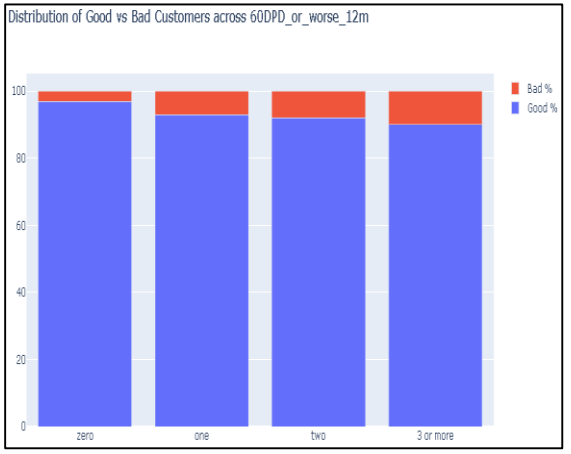




Graph showing percentage of good and bad customers in each of the categories of variable







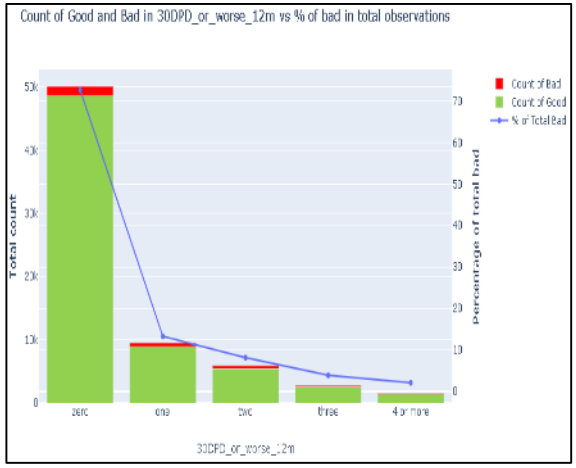
Exploratory Data Analysis

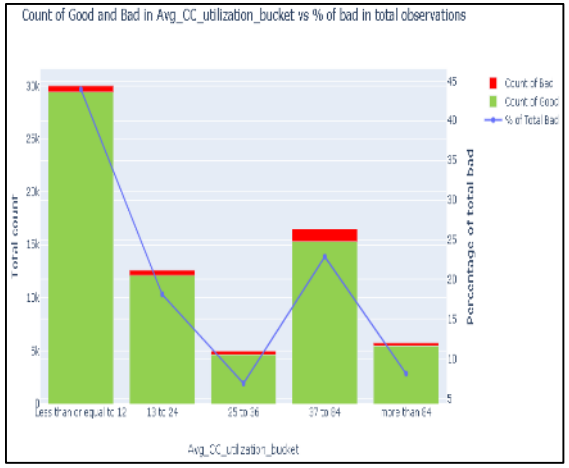
No of times 30 DPD or worse in last 12 months

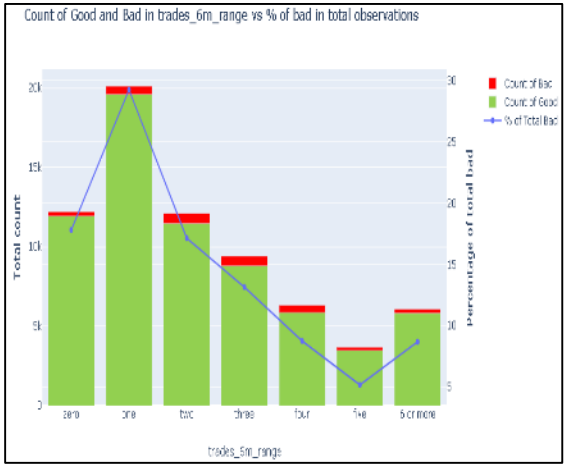
Avgas CC Utilization in last 12 months

No of trades opened in last 6 months

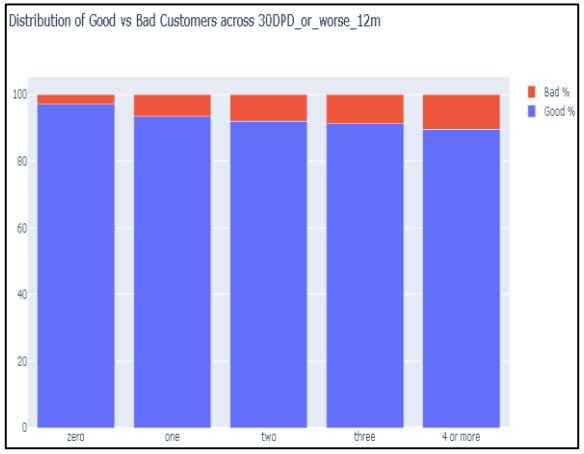
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

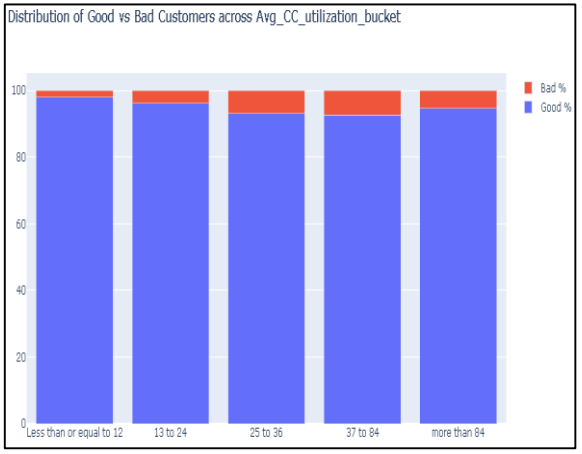


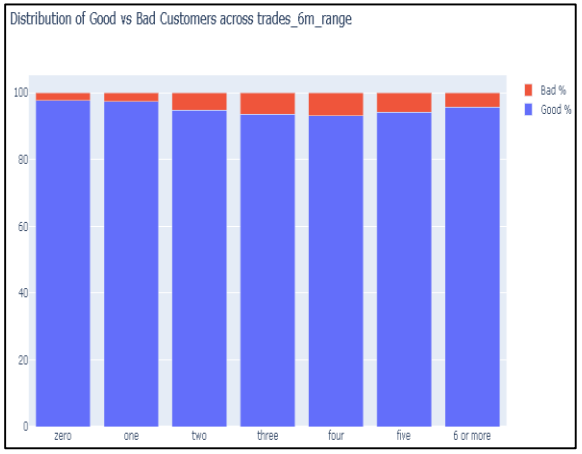




Graph showing percentage of good and bad customers in each of the categories of variable







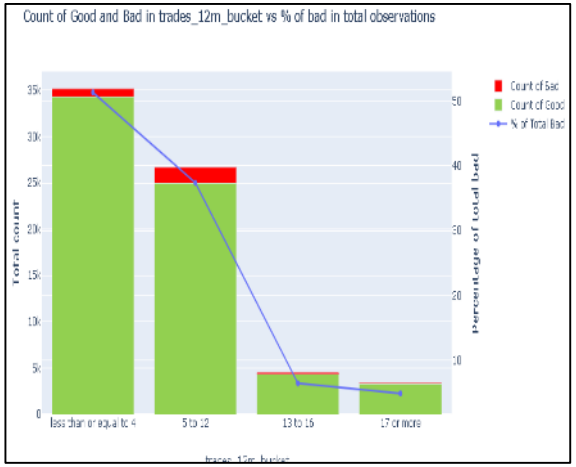
Exploratory Data Analysis

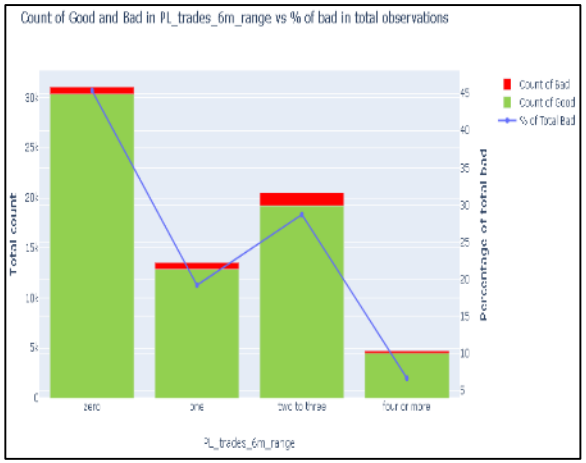
No of trades opened in last 12 months

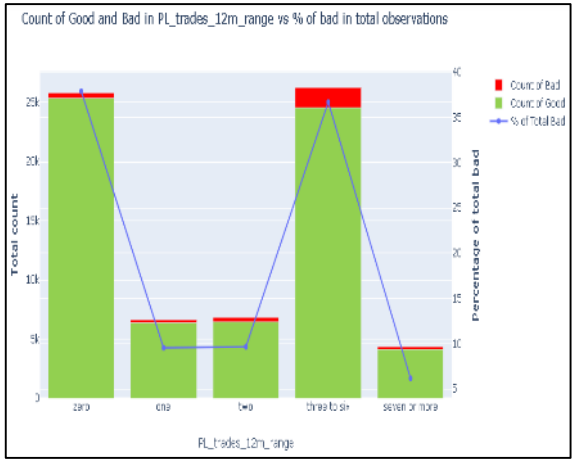
No of PL trades opened in last 6 months

No of PL trades opened in last 12 months

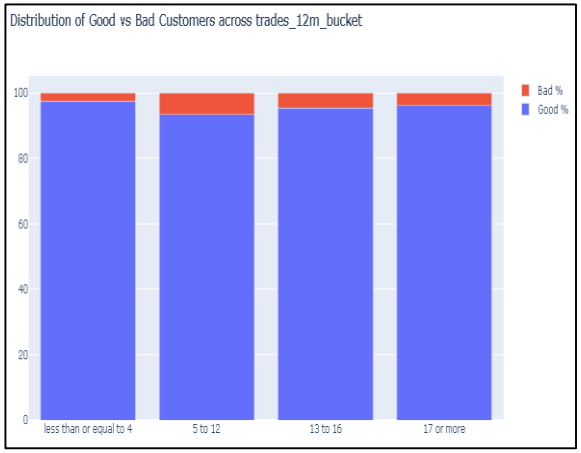
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

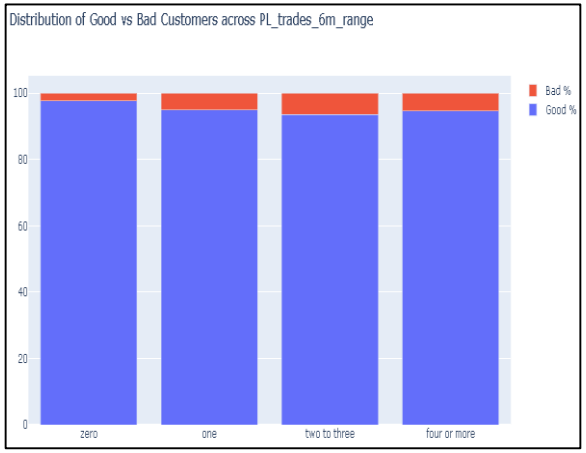


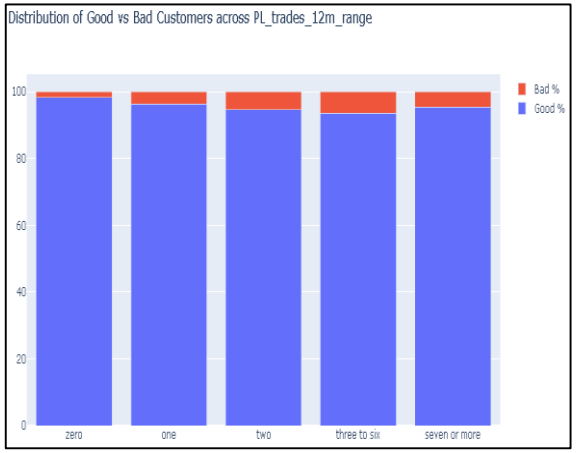




Graph showing percentage of good and bad customers in each of the categories of variable







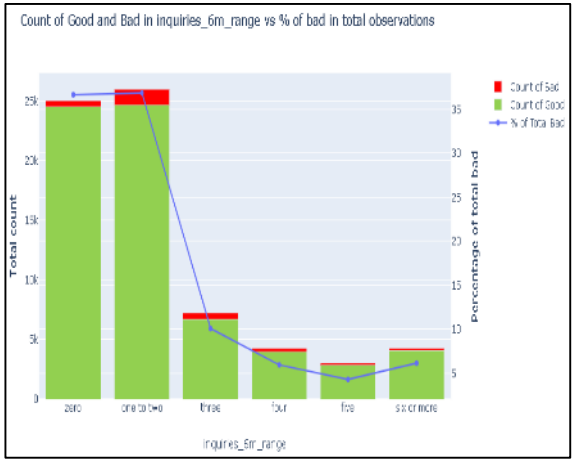
Exploratory Data Analysis

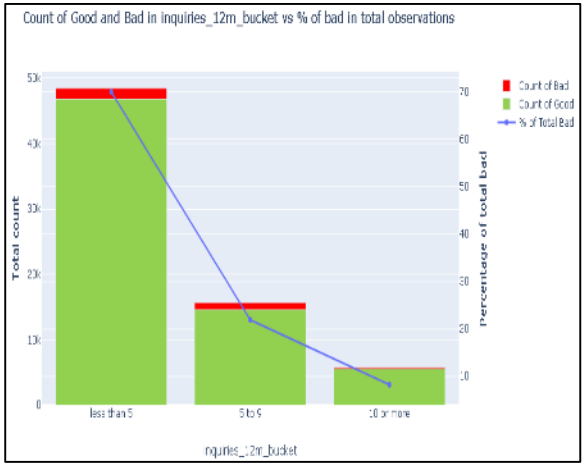
No of inquiries in last 6 months (excluding home & auto loans)

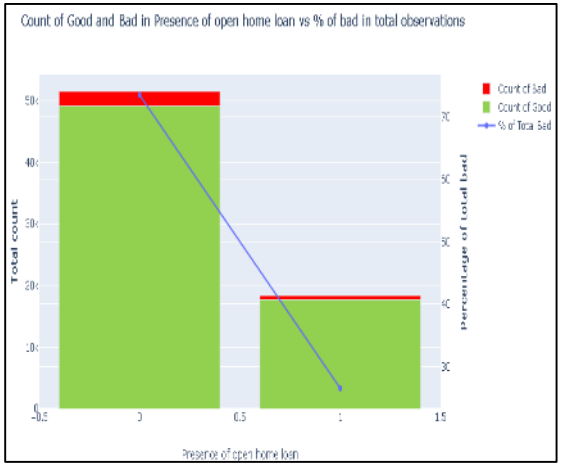
No of inquiries in last 12 months (excluding home & auto loans)

Presence of open home loan

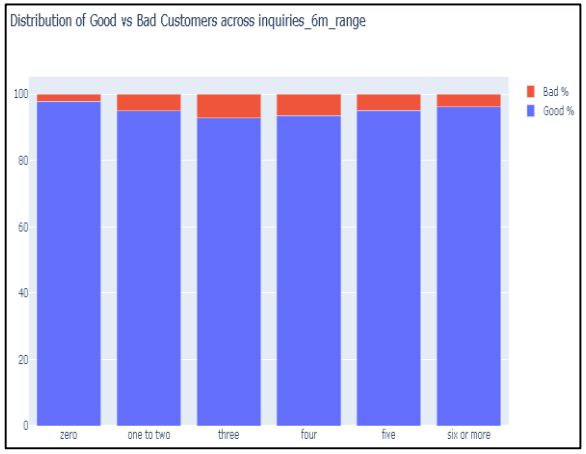
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

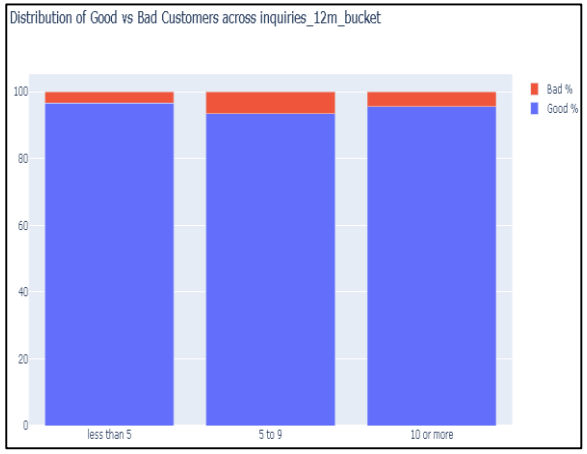


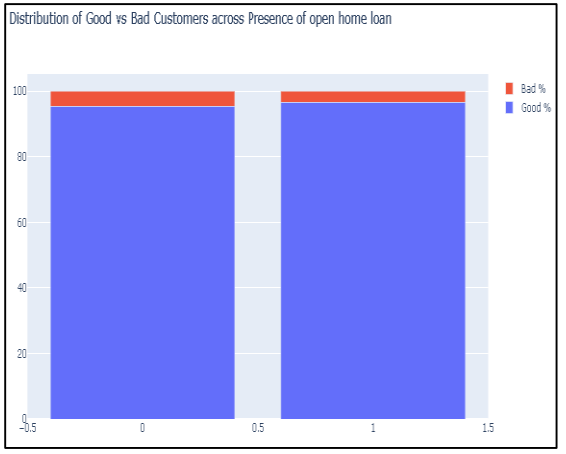




Graph showing percentage of good and bad customers in each of the categories of variable







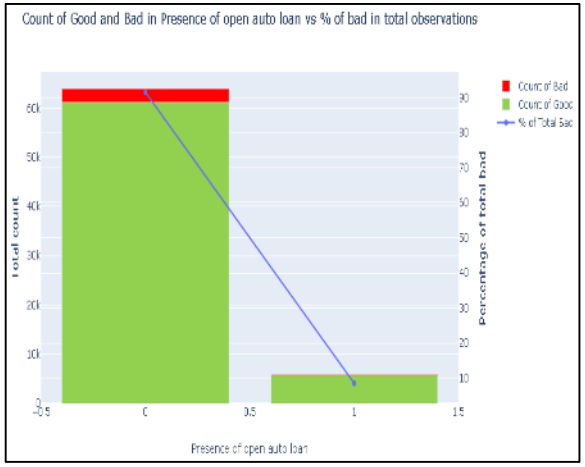
Exploratory Data Analysis

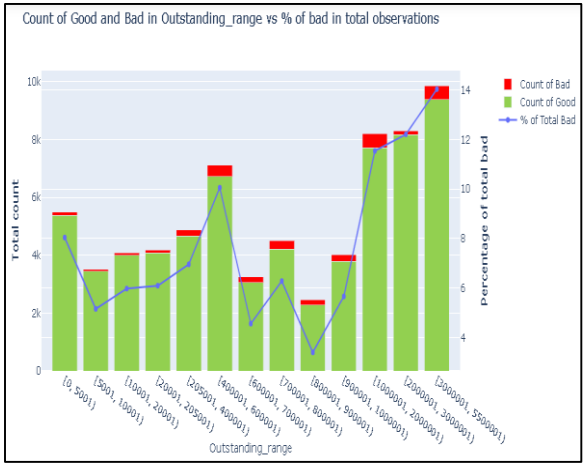
Presence of open auto loan

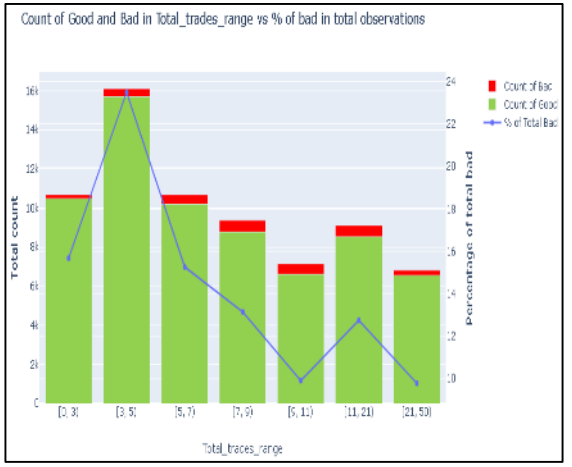
Outstanding Balance

Total number of trades

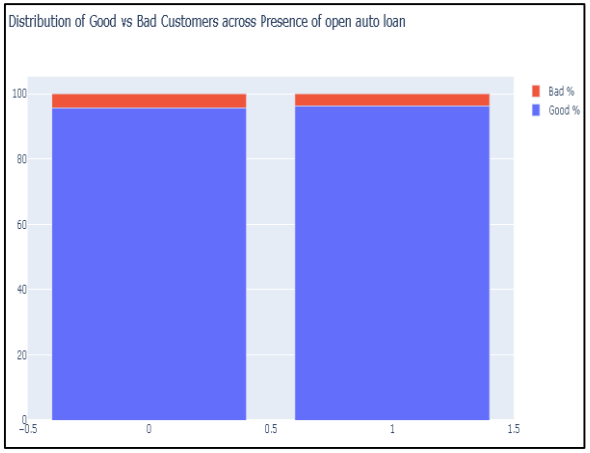
Graph showing count of good and bad across categories of a variable along with percentage of bad in total observations on secondary axis

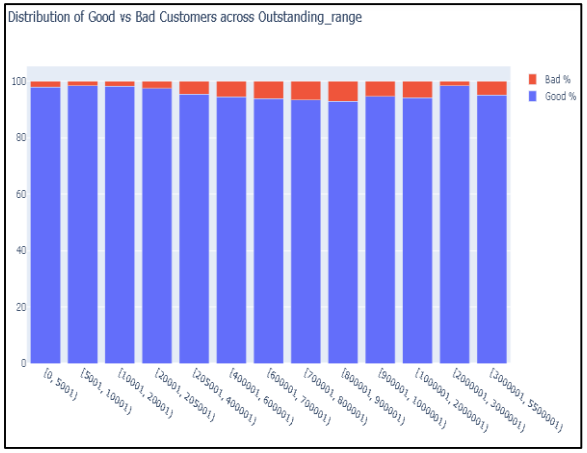


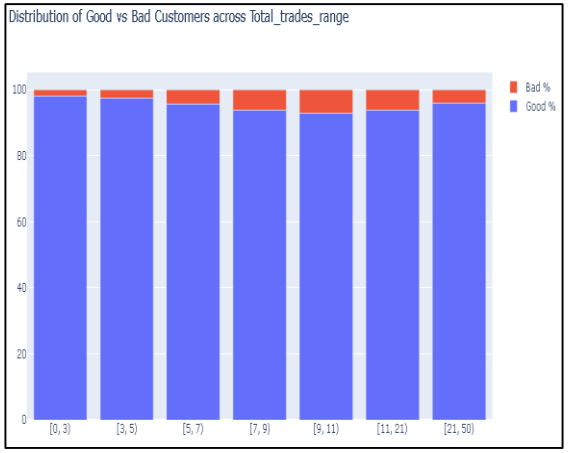




Graph showing percentage of good and bad customers in each of the categories of variable







IV Table

Credit Beaureu Data IV

	Variable	IV
2	Avgas CC Utilization in last 12 months_woe	0.307859
7	No of PL trades opened in last 12 months_woe	0.297757
12	No of trades opened in last 12 months_woe	0.297070
3	No of Inquiries in last 12 months (excluding h...	0.294731
5	Total No of Trades_woe	0.245290
13	No of times 30 DPD or worse in last 6 months_woe	0.243130
11	Outstanding Balance_woe	0.242113
4	No of PL trades opened in last 6 months_woe	0.220931
15	No of times 30 DPD or worse in last 12 months_woe	0.216964
0	No of times 90 DPD or worse in last 12 months_woe	0.213784
14	No of times 60 DPD or worse in last 6 months_woe	0.209280
10	No of Inquiries in last 6 months (excluding ho...	0.205999
9	No of trades opened in last 6 months_woe	0.189444
6	No of times 60 DPD or worse in last 12 months_woe	0.187297
8	No of times 90 DPD or worse in last 6 months_woe	0.160044
1	Presence of open home loan_woe	0.000000
	en auto loan_woe	0.000000

nal_Capstone Project BFSI-Avineet.ipynb

Demographics IV

	Variable	IV
6	No of months in current residence_woe	0.094518
9	Income_woe	0.045757
7	No of months in current company_woe	0.032684
8	Age_woe	0.006919
2	No of dependents_woe	0.002653
0	Profession_woe	0.002172
5	Gender_woe	0.000334
1	Education_woe	0.000283
3	Type of residence_woe	0.000108
4	Marital Status (at the time of application)_woe	0.000100

Conclusions from EDA

Note

From the earlier graphs it can be concluded that, there are plenty of variables wherein certain categories reflect high percentage of total bad customers. However, the same also needs to be analysed based on the no of applications received for each of those categories. Accordingly conclusion of EDA is in line with variables selected using IV

1. Demographic variables are not very good predictors of defaulting. Only below 3 variables seems significant.

1.Income
2.No of months in current residence
3.No of months in current company

2.Credit bureau dataset has many variables which seems like good predictors of defaulters.

1.No of times 90 DPD or worse in last 6 months
2.No of times 60 DPD or worse in last 6 months
3.No of times 30 DPD or worse in last 6 months
4.No of times 90 DPD or worse in last 12 months
5.No of times 60 DPD or worse in last 12 months
6.No of times 30 DPD or worse in last 12 months
7.Avgas CC Utilization in last 12 months
8.No of trades opened in last 6 months
9.No of trades opened in last 12 months

3. There is no correlation between numeric variables of demographic dataset.

4. Few numeric variables of Credit bureau dataset show strong positive correlation with other variables.

1.The 6 variables – No of times 90/60/30 DPD or worse in last 6/12 months are highly correlated among themselves.
2.No of enquiries in last 6 months/12 months excluding home, auto loan variables are highly correlated.
3.No. of trade opened in 6/12 months, total number of trades, no of PL trades in 6/12 months are correlated.

MODEL BUILDING APPROACH

- 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 is split into Train and Test in 70:30 ratio for model building. All models are trained on training datasets and regularization was done by tuning of hyper parameters with cross validation on validation datasets. All the models are tested on test datasets that were kept separate from training and validation dataset.
- DATA SAMPLING: The given data is highly imbalanced. We have sampled data using Parameter Class="Balanced "for balancing the training data sets. The cutoff value for the probability of default was chosen such that model evaluation metrics like accuracy ,sensitivity and specificity were almost equal to each other.
- Logistic Regression was built by iteratively removing using these two algorithms
    1. Finding best Features using RFECV
    2. Building model using selected features by RFECV.

LOGISTIC REGRESSION MODEL ON DEMOGRAPHIC DATA

- Important predictors :Five features have been selected using RFECV.

1.Age_woe

2.Income_woe

3.No of months in current company_woe

4.No of months in current residence_woe

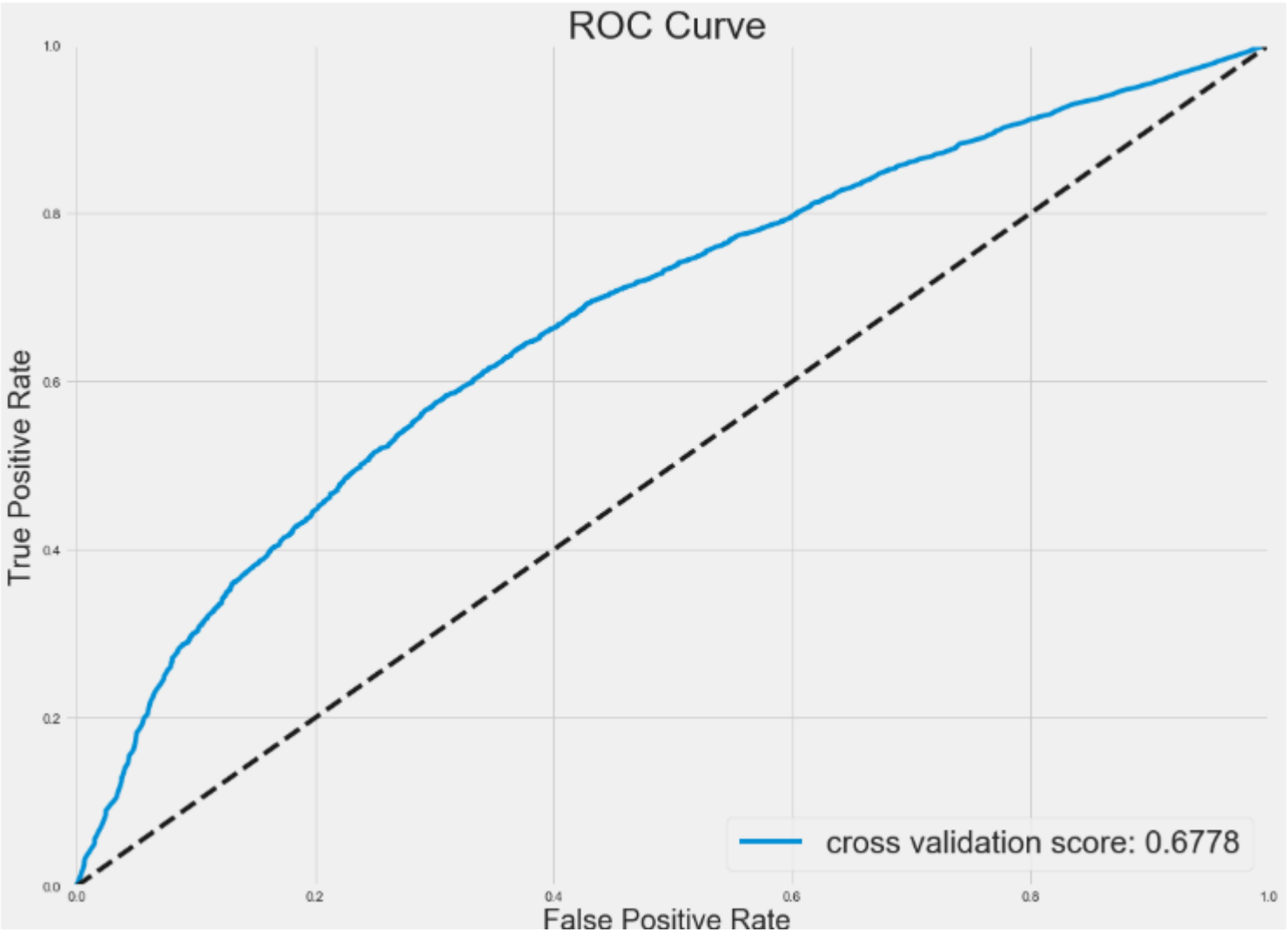
5.Profession_woe

Thus a logistic regression model based only on demographic data seems to have low performance.

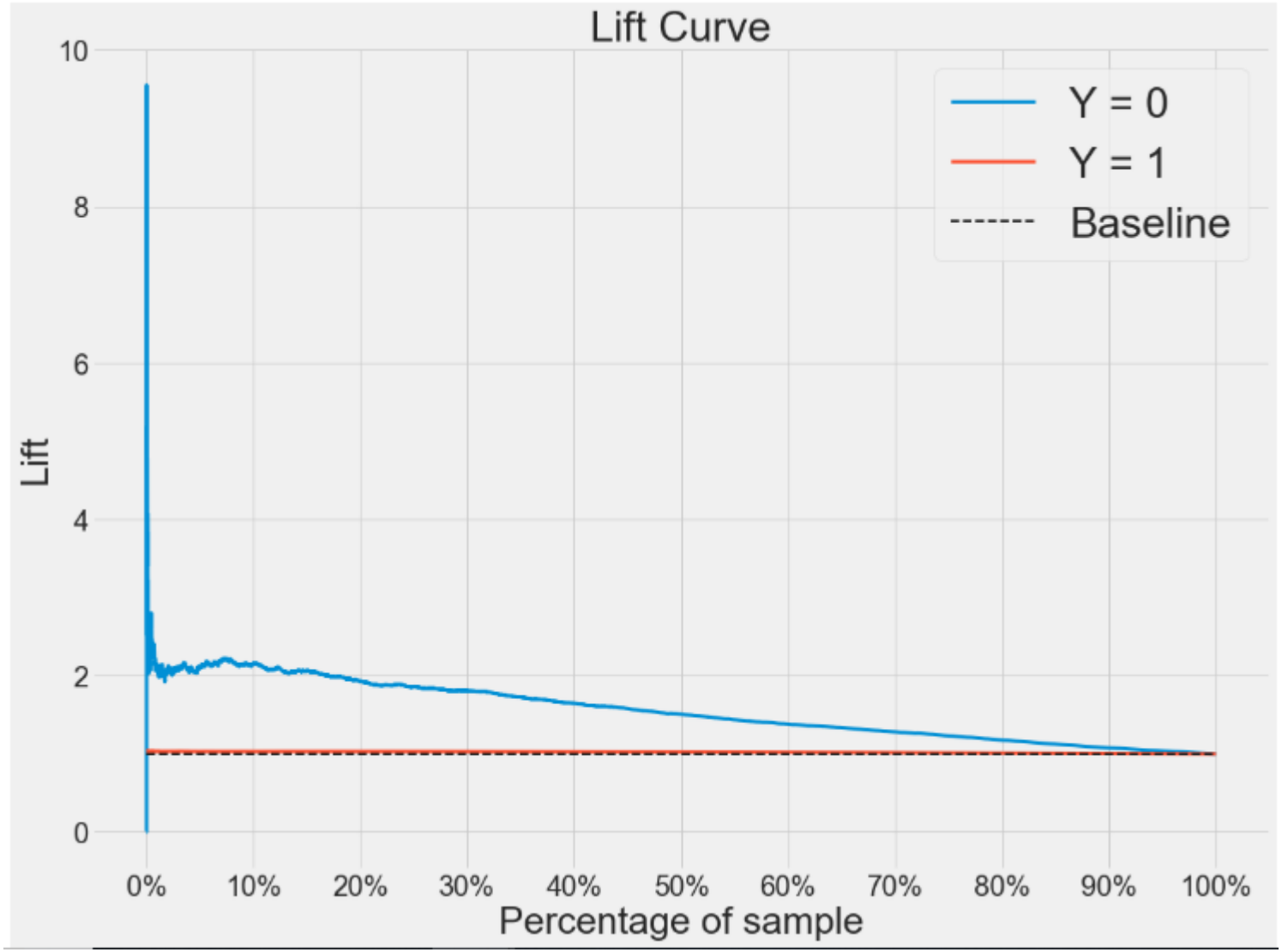
Statistics	Values
Cut_off	0.5
Accuracy	0.60
Sensitivity	0.60
Specifivity	0.56

Thus a logistic regression model based only on demographic data seems to have low performance.

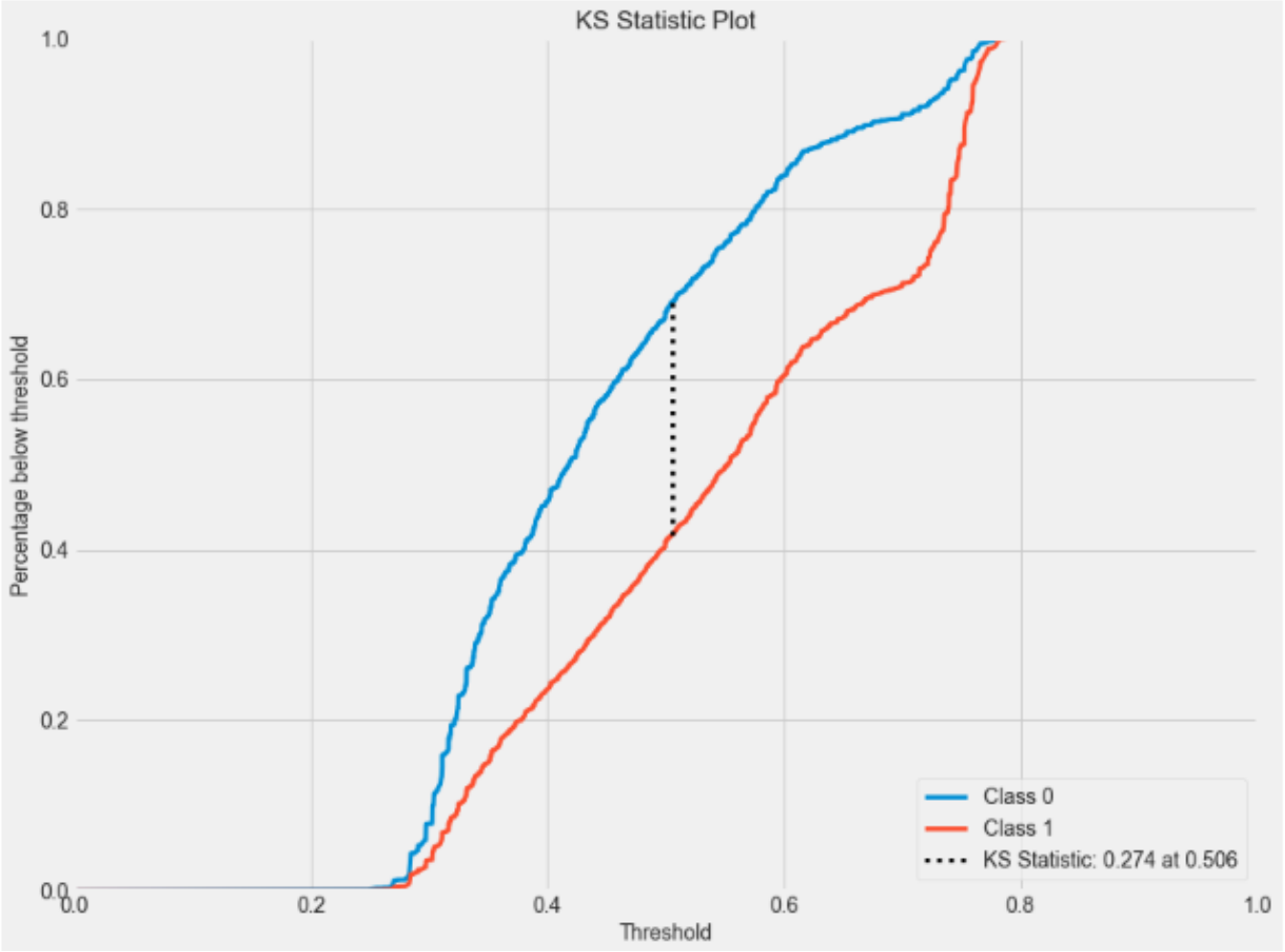
ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
cross validation score: 0.6778

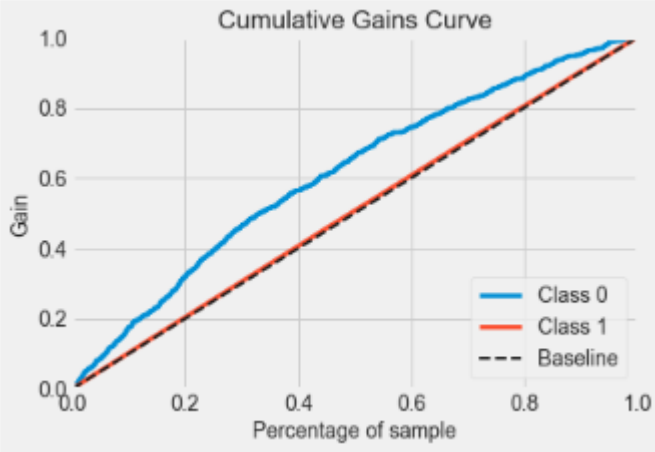
Lift Curve

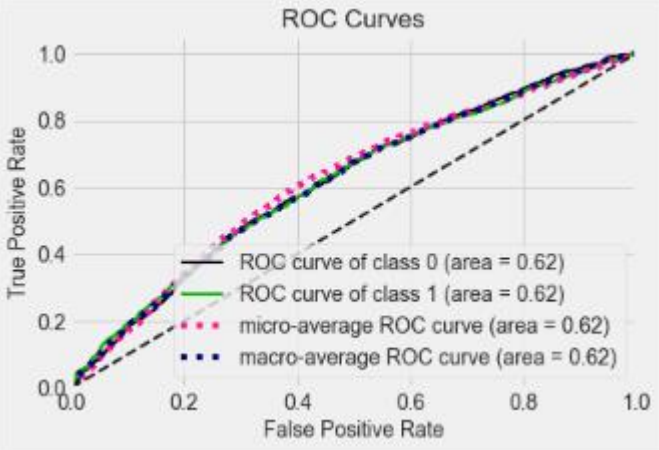

Lift Curve
Y = 0
Y = 1
Baseline
Lift
Percentage of sample
0
2
4
6
8
10
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

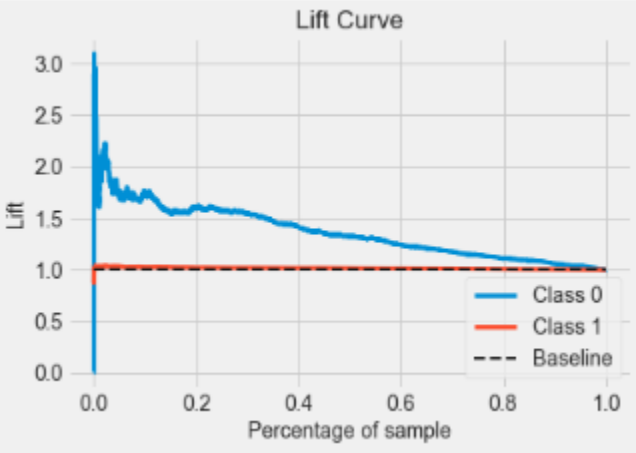
KS Statistics



Cummulative ,Lift and KS on Test DataSet







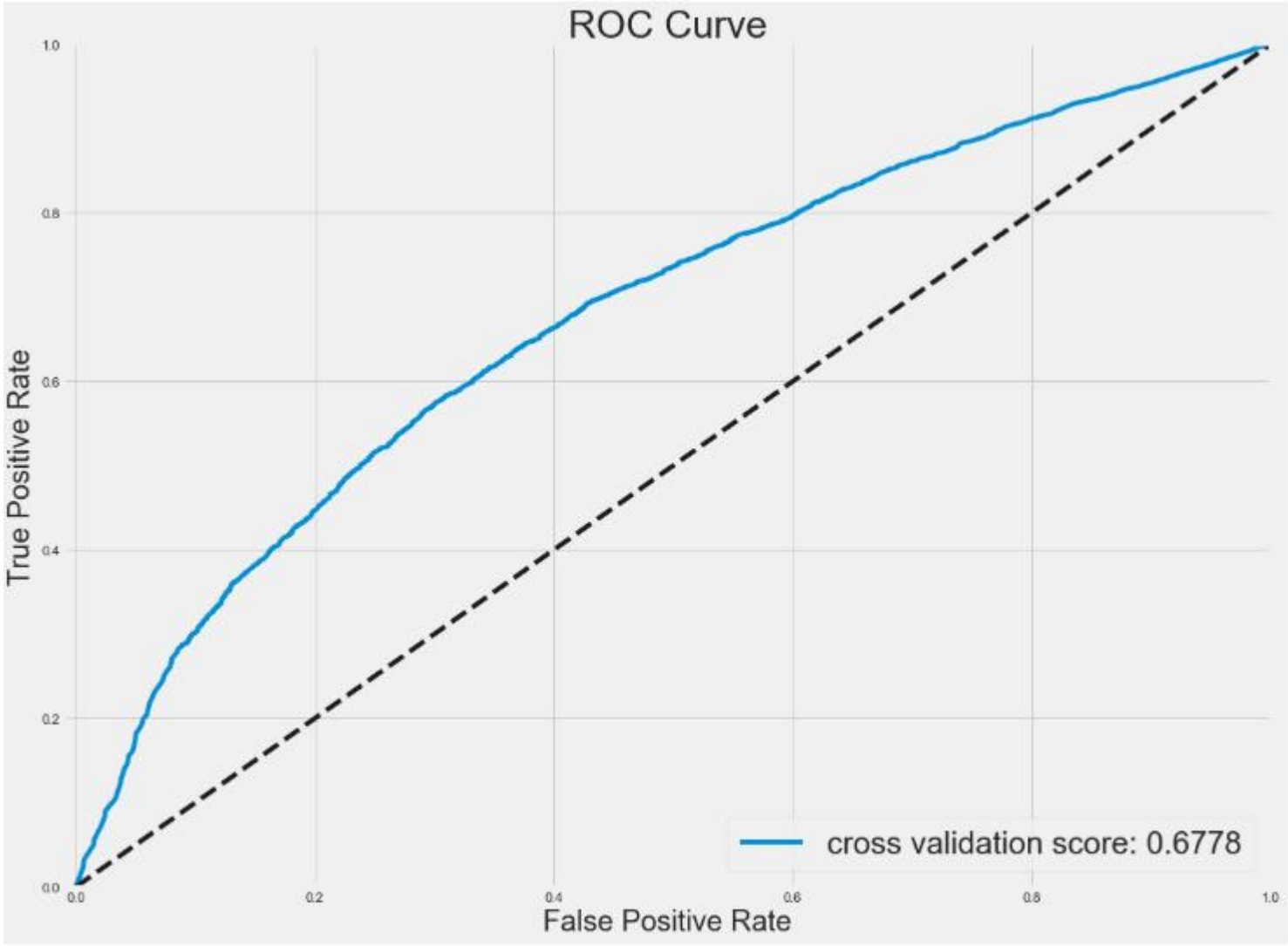
LOGISTIC REGRESSION MODEL ON MERGED CREDIT BUREAU AND DEMOGRAPHIC DATASET WITHOUT REJECTED 1425 RECORDS

Predictors in logistic regression model trained on a part of merged credit bureau and demographic dataset (merged on the application id column) without rejected 1425 records which does not have performance tags are as follows :-

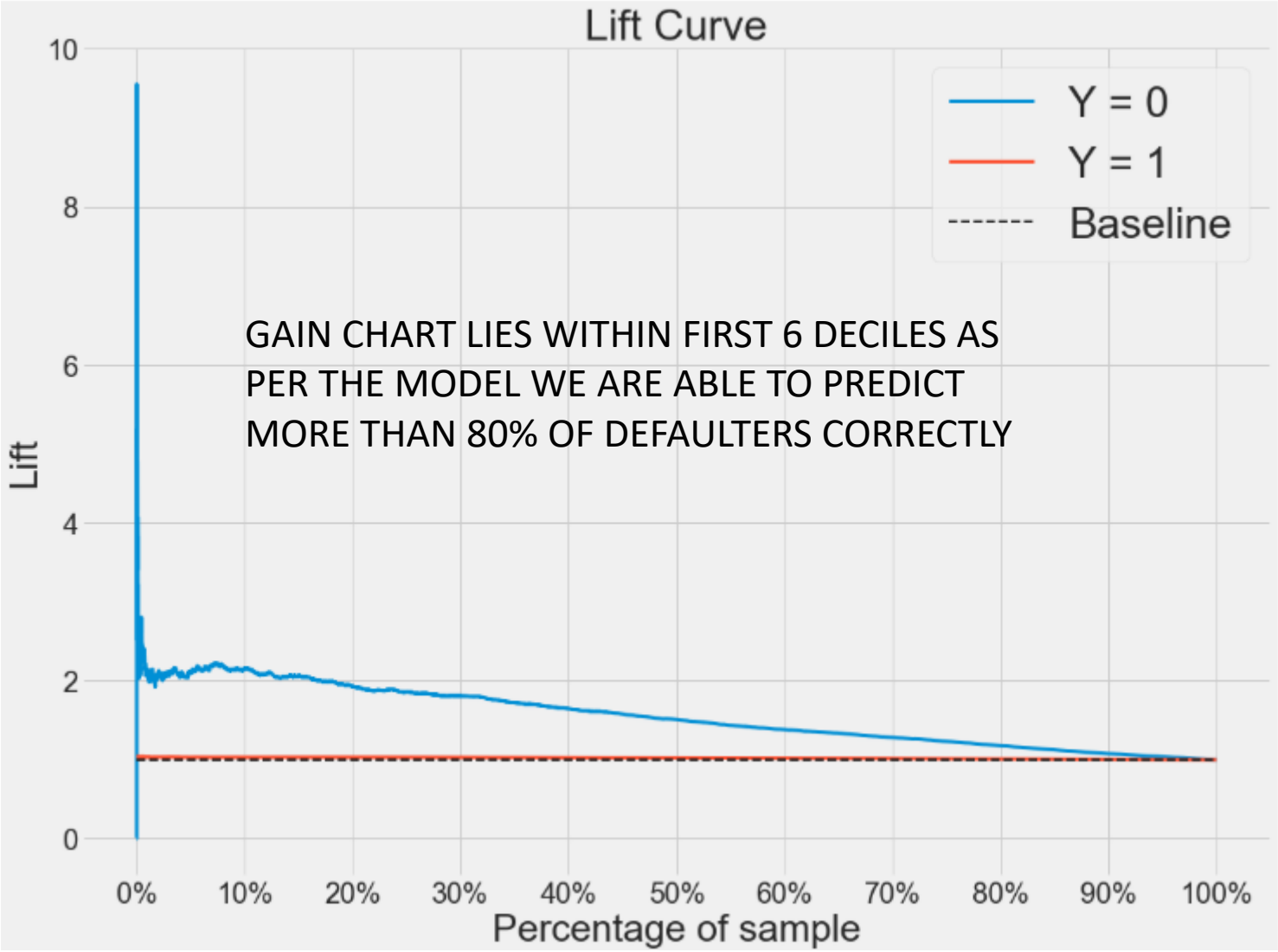
1.No of times 30 DPD or worse in last 6 months_woe
2.No of Inquiries in last 12 months (excluding home & auto loans)_woe
3.No of months in current company_woe
4.No of times 60 DPD or worse in last 12 months_woe
5.No of trades opened in last 12 months_woe
6.Avgas CC Utilization in last 12 months_woe
7.No of times 30 DPD or worse in last 12 months_woe

Statistics	Values
Cutoff	0.51
Accuracy	60%
Specificity	66%
Sensitivity	58%

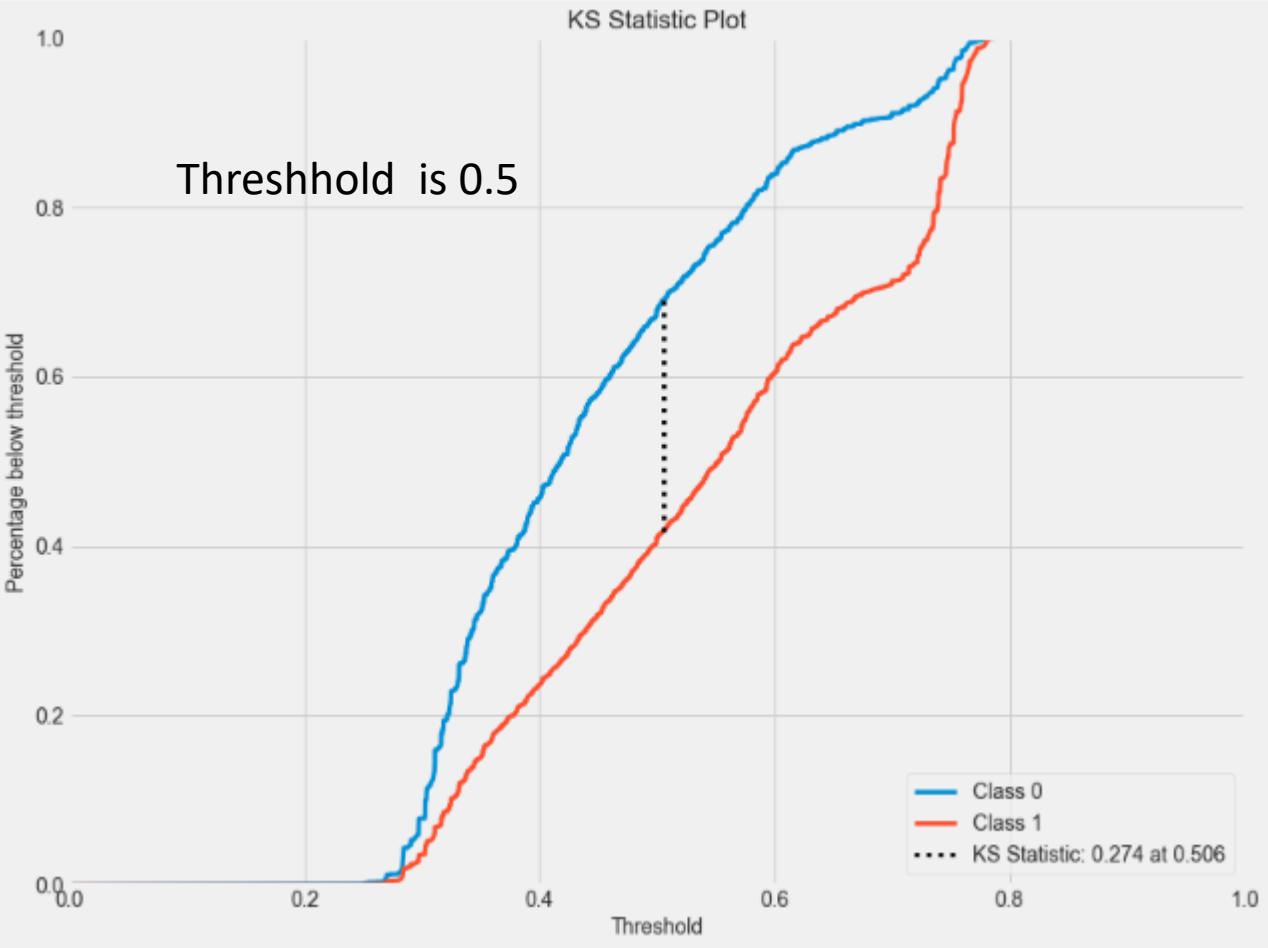
ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
cross validation score: 0.6778

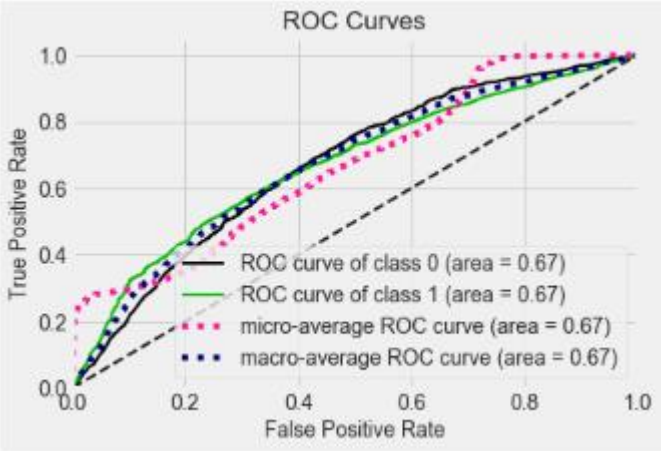
LIFT Curve

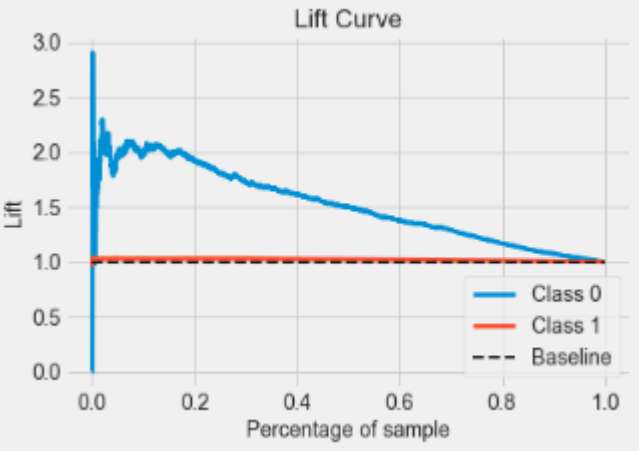

Lift Curve
Y = 0
Y = 1
Baseline
GAIN CHART LIES WITHIN FIRST 6 DECILES AS PER THE MODEL WE ARE ABLE TO PREDICT MORE THAN 80% OF DEFAULTERS CORRECTLY
Lift
Percentage of sample
0
2
4
6
8
10
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

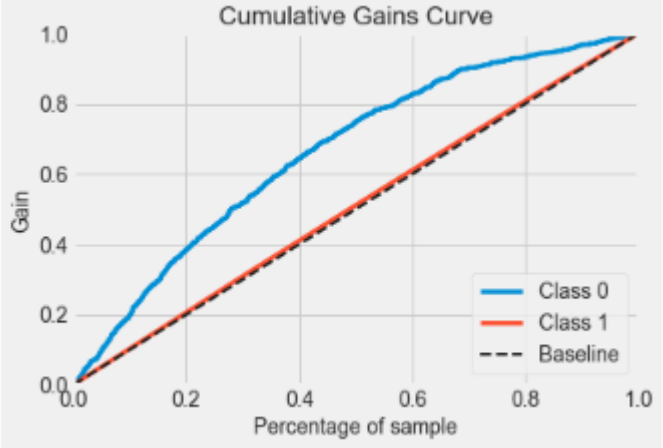
KS Statistics



ROC Curve,Lift Curve and KS Statistics on Test Data

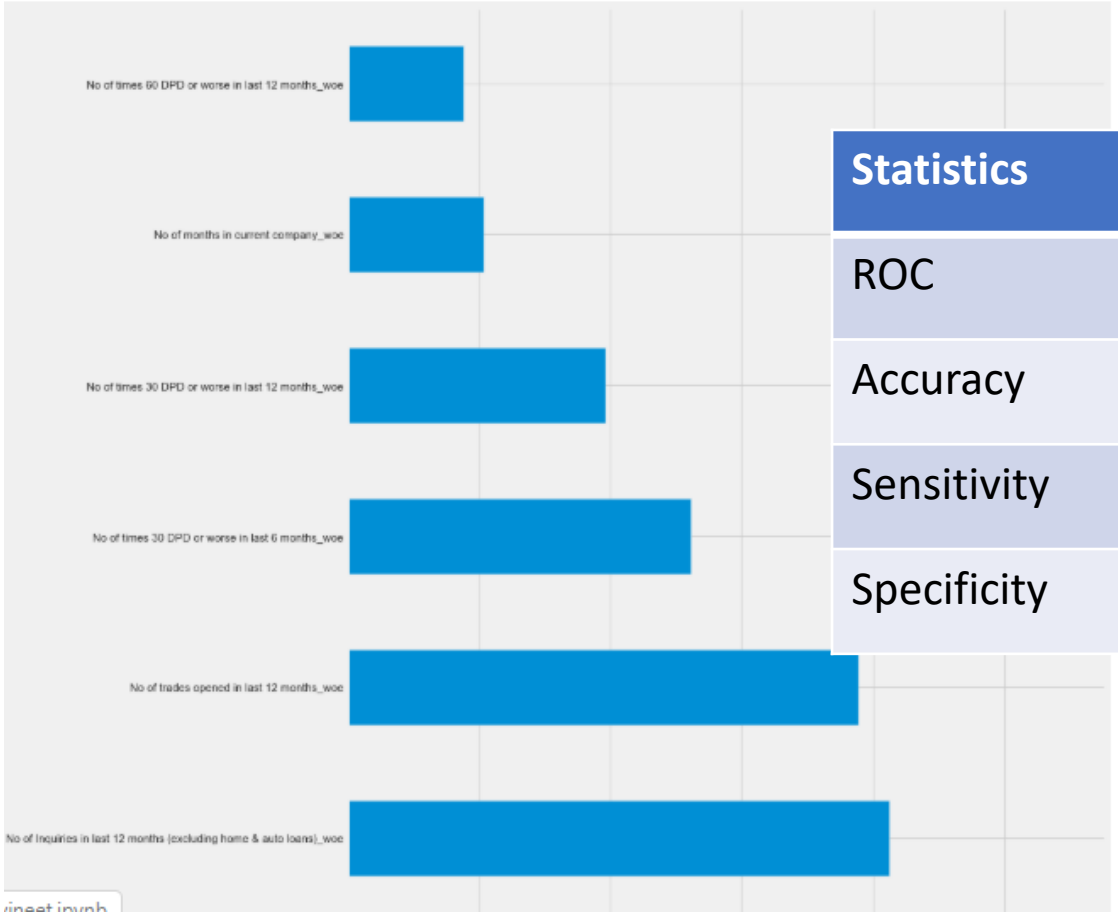






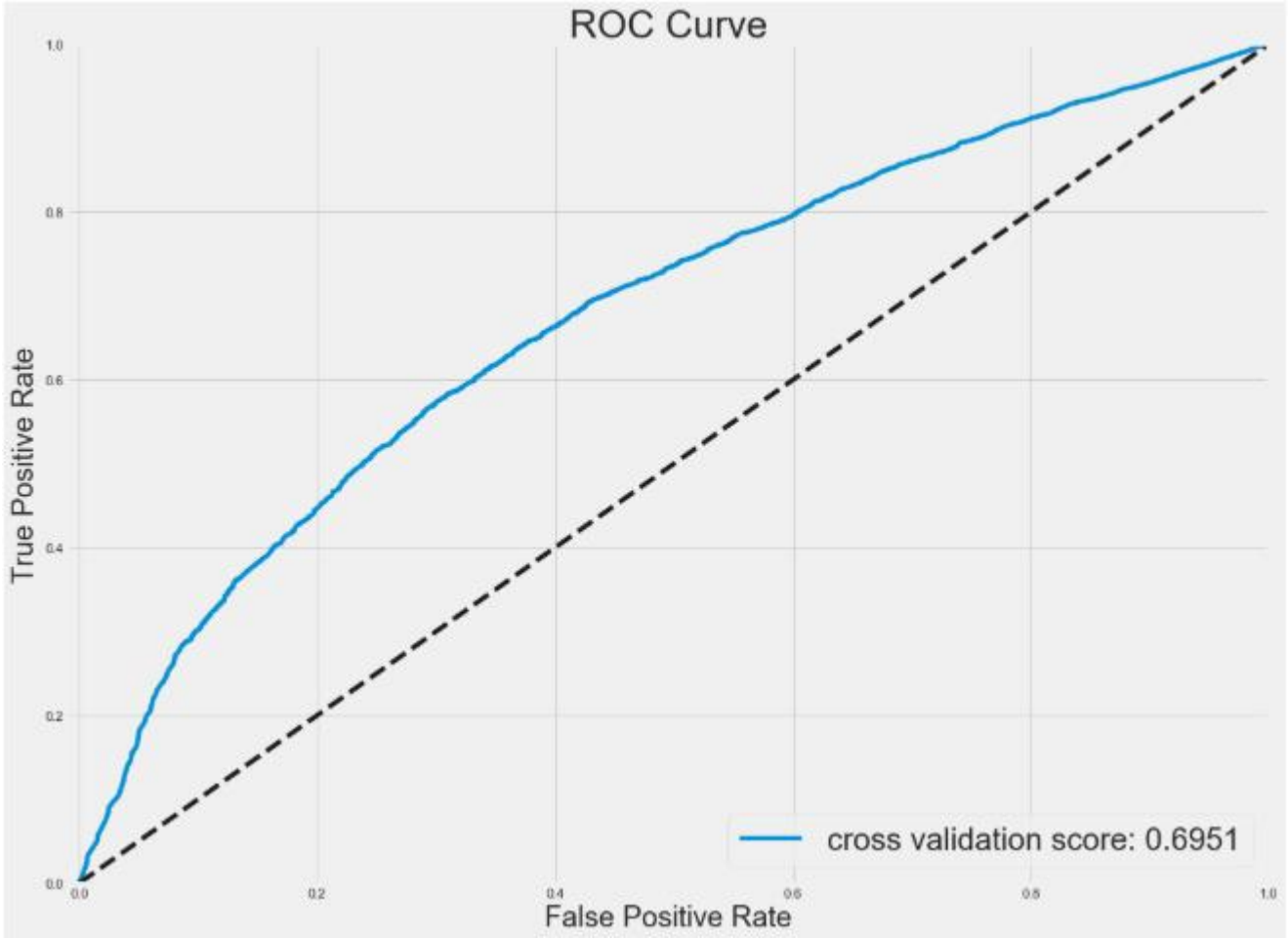
Random Forest Based on Merged DataSet

- Important Features obtained by RF is:-

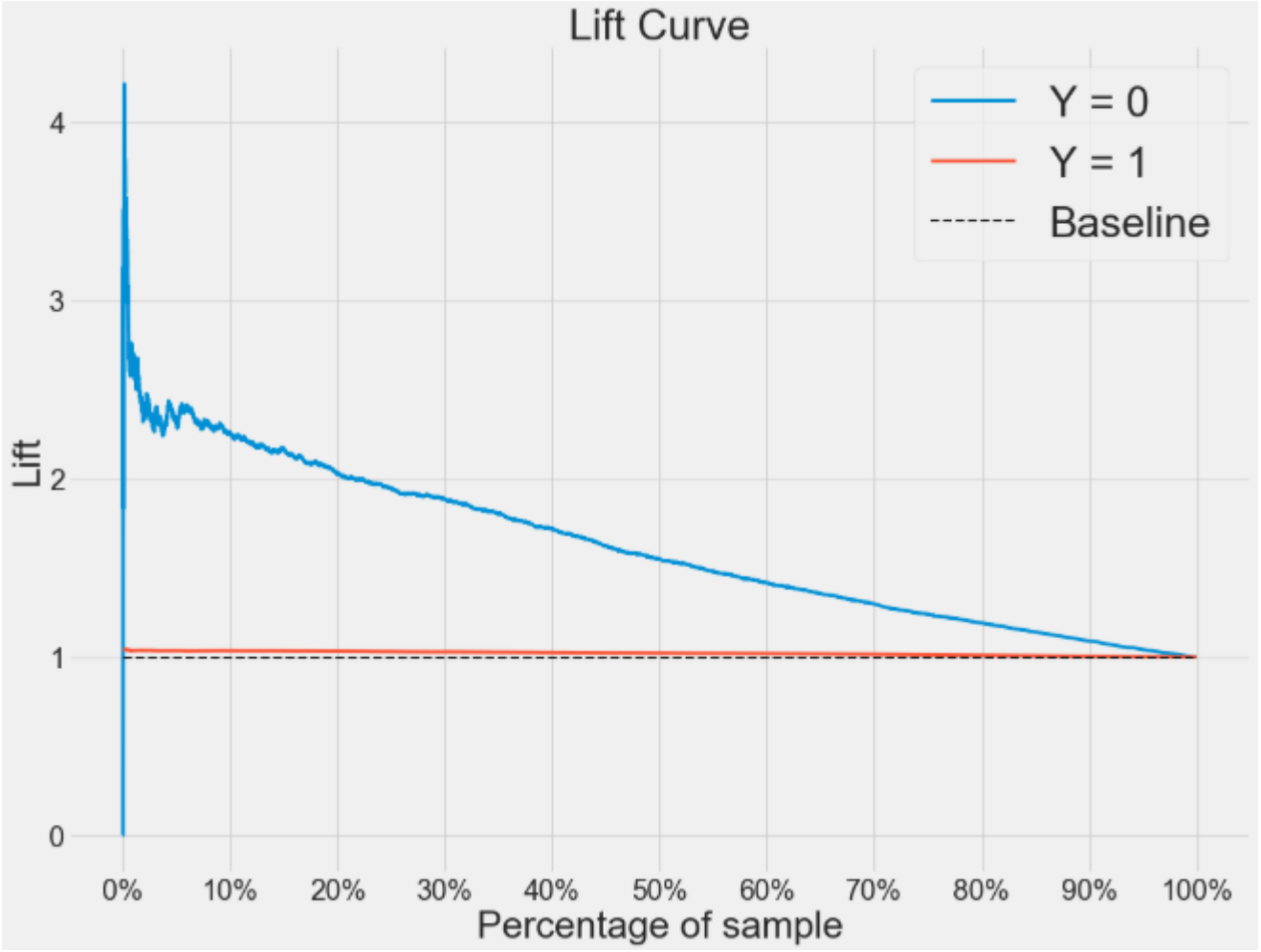


Statistics	Value
ROC	0.5
Accuracy	58%
Sensitivity	57%
Specificity	68%

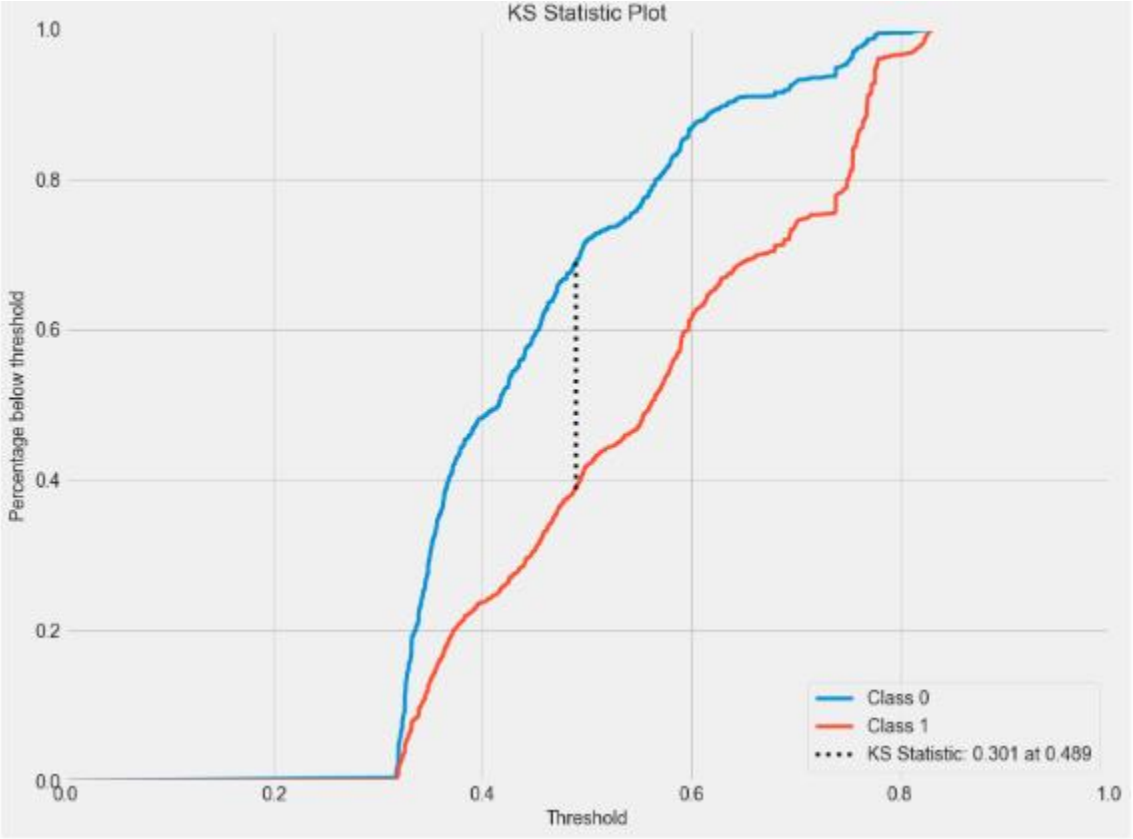
ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
cross validation score: 0.6951

LIFT Chart


Lift Curve
Y = 0
Y = 1
Baseline
Lift
0
1
2
3
4
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Percentage of sample

KS Statistics



Application Scorecard

- Final application scorecard was made using the Logistic regression model on the entire dataset which does not contain predictions for missing values in “Performance Tag” in 1425 records.
- The logistic regression model was chosen since its evaluation metrics were comparable to other models as well it’s an easily interpretable simple model.
- 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. ln(odd(good)) was calculated

5. Used the following formula for computing application score card: 400 + slope * (ln(odd(good)) - ln(10)) where slope is 20/(ln(20)-ln(10)) Where, slope=20/(log(20)-log(10)

Summary of application_score_card values

Scores range from 325 to 389 for applicants with median score being 380

. • Higher scores indicate less risk for defaulting

- Cutoff selected for probability of default for logistic regression model was 0.5
- CUTOFF_SCORE= 400 + (slope * (log((1-0.5)/0.5) - log(10)))
- CUTOFF SCORE is equal to 338.18
- No.of applicants with Default is 901
- No.of applicants with non default model is 20021

Financial Benefits of the Model

- Profit calculations – with model Vs without model
- We have considered an average profit of Rs.5000 from each non defaulters and an average loss of Rs.1,00,000 when each accepted applicant defaults
- Net Profit without model = Rs 3.9665 crores
- Profit using model will be total profit due to each true positive and each true negative minus loss from each false positive and each false negative prediction
- Profit with model = Rs15.6865 crores
- Net financial gain with using our model = Rs. 11.72 crores
- Percentage financial gain = 295.47%

Revenue loss and Potential Credit loss saved

- Revenue Loss : Occurs when good customers are identified as bad and credit card application is rejected.
  - No of candidates rejected by the model who didn't default – 20980.
  - Total No of candidates who didn't default – 66853
  - % of good candidates rejected by our model – 31.38%
  - About 31.38% of the non defaulting customers are rejected which resulted in revenue loss.

- Credit Loss Saved : The candidates who have been selected by the bank and have defaulted are responsible for the credit loss to the bank.
  - % of candidates approved and then defaulted when model was not used = 4.2%
  - % of candidates approved and then defaulted when model was used = 1311/69799 = 1.8%
  - Credit loss saved => 4.2 – 1.8 = 2.4%

Conclusion

- Logistic regression model is chosen as the final Model with 70% of Accuracy.
- Optimal score cut-off value of 338.18 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 in rejecting the candidate who may default in future.
- There is Net Financial gain of 295.47% after using the model.
- In 372 case ,model have predicted actual default correctly which I s actual benefit of project.