BFS CAPSTONE PROJECT

Group Members:

1. Anmol Goel
2. Sarthak Bhatnagar

BFS CAPSTONE PROJECT

Problem Statement

- CredX is a leading credit card provider that gets thousands of credit card applications 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’.

- In this project, your task is to help CredX identify the right customers using predictive models. Using past data of the bank’s applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Data Dictionary - I

Demographic Data	
Variables	**Description**
Application ID	Unique ID of the customers
Age	Age of customer
Gender	Gender of customer
Marital Status	Marital status of customer (at the time of application)
No of dependents	No. of childrens of customers
Income	Income of customers
Education	Education of customers
Profession	Profession of customers
Type of residence	Type of residence of customers
No of months in current residence	No of months in current residence of customers
No of months in current company	No of months in current company of customers
Performance Tag	Status of customer performance (" 1 represents "Default")

Data Dictionary - II

Credit Bureau Data	
Variable	**Description**
Application ID	Customer application ID
No of times 90 DPD or worse in last 6 months	Number of times customer has not payed dues since 90days in last 6 months
No of times 60 DPD or worse in last 6 months	Number of times customer has not payed dues since 60 days last 6 months
No of times 30 DPD or worse in last 6 months	Number of times customer has not payed dues since 30 days days last 6 months
No of times 90 DPD or worse in last 12 months	Number of times customer has not payed dues since 90 days days last 12 months
No of times 60 DPD or worse in last 12 months	Number of times customer has not payed dues since 60 days days last 12 months
No of times 30 DPD or worse in last 12 months	Number of times customer has not payed dues since 30 days days last 12 months
Avgas CC Utilization in last 12 months	Average utilization of credit card by customer
No of trades opened in last 6 months	Number of times the customer has done the trades in last 6 months
No of trades opened in last 12 months	Number of times the customer has done the trades in last 12 months
No of PL trades opened in last 6 months	No of PL trades in last 6 month of customer
No of PL trades opened in last 12 months	No of PL trades in last 12 month of customer
loans)	Number of times the customers has inquired in last 6 months
loans)	Number of times the customers has inquired in last 12 months
Presence of open home loan	Is the customer has home loan (1 represents "Yes")
Outstanding Balance	Outstanding balance of customer
Total No of Trades	Number of times the customer has done total trades
Presence of open auto loan	Is the customer has auto loan (1 represents "Yes")
Performance Tag	Status of customer performance (" 1 represents "Default")

Approach

Approach to be written here, at last

- 1.) Inspecting the Demographic Data
- 2.) EDA DEMOGRAPHIC
  - 2.A) EDA DEMOGRAPHIC - CATEGORICAL VARIABLES
  - 2.B) EDA DEMOGRAPHIC - CONTINUOUS VARIABLES
- 3.) WoE and IV Analysis for Demographics Data
- 4.) Credit Bureau Data
- 5.) EDA CREIT Bureau Data
- 6.) WOE AND IV ANALYSIS OF Credit Bureau Data
- 7.) Merging the the data of demographics and CreditBureau
- 8.) Model Building
  - 8.1.) Models Building for Demographic Data (Non-transformed)
    - 8.1.A.) Logistic Regression Model
    - 8.1.B.) Decision Tree Model
    - 8.1.C.) Random Forest Model
  - 8.2.) Building Model with WOE Transformed Data (Demographic)
    - 8.2.A.) Logistic Regression Model
    - 8.2.B.) Decision Tree Model
    - 8.2.C.) Random Forest Model
  - 8.3.) Models Building for Combined(Demographic and Credit Bureau) (Non-transformed)
    - 8.3.A.) Logistic Regression Model
    - 8.3.B.) Decision Tree Model
    - 8.3.C.) Random Forst Model
  - 8.4.) Models Building for Combined(Demographic and Credit Bureau) WoE Data
    - 8.4.A.) Logistic Regression Model
    - 8.4.B.) Decision Tree Model
    - 8.4.C.) Random Forest Model
- 9.) Model Evaluation
- 10.) Application Scorecard

Approach

-
  -
    - 8.3.C.) Random Forst Model
-
  - 8.4.) Models Building for Combined(Demographic and Credit Bureau) WoE Data
-
  -
    - 8.4.A.) Logistic Regression Model
-
  -
    - 8.4.B.) Decision Tree Model
-
  -
    - 8.4.C.) Random Forest Model
- 9.) Model Evaluation
- 10.) Application Scorecard
- 11.) Benefits of our ML model
- 12.) Evaluating Financial Risk of our Model

BFS CAPSTONE PROJECT

Approach

- Understand Both data sets given for this project
- Clean and transform data according to the business logic. For example:
    - Rows containing Null value for Performance tag should be removed
    - Rows with duplicate ID should be removed
    - Negative Salary values replaced with median values across the column
- For IV analysis, imputing values needs to be ignored.
- Data needs to be divided into 2 parts; 1st having NULL rows and 2nd without Nulls to perform EDA and analyze important values
- Weight of Evidence (WoE) and Information Value (IV) analysis and prepare WoE transformed dataset.
    - Take Demographic data-set and perform WoE transformation and find significant variables based on IV.
    - Merge Demographic data-set with Credit Bureau dataset and perform WoE transformation and get most significant variables.

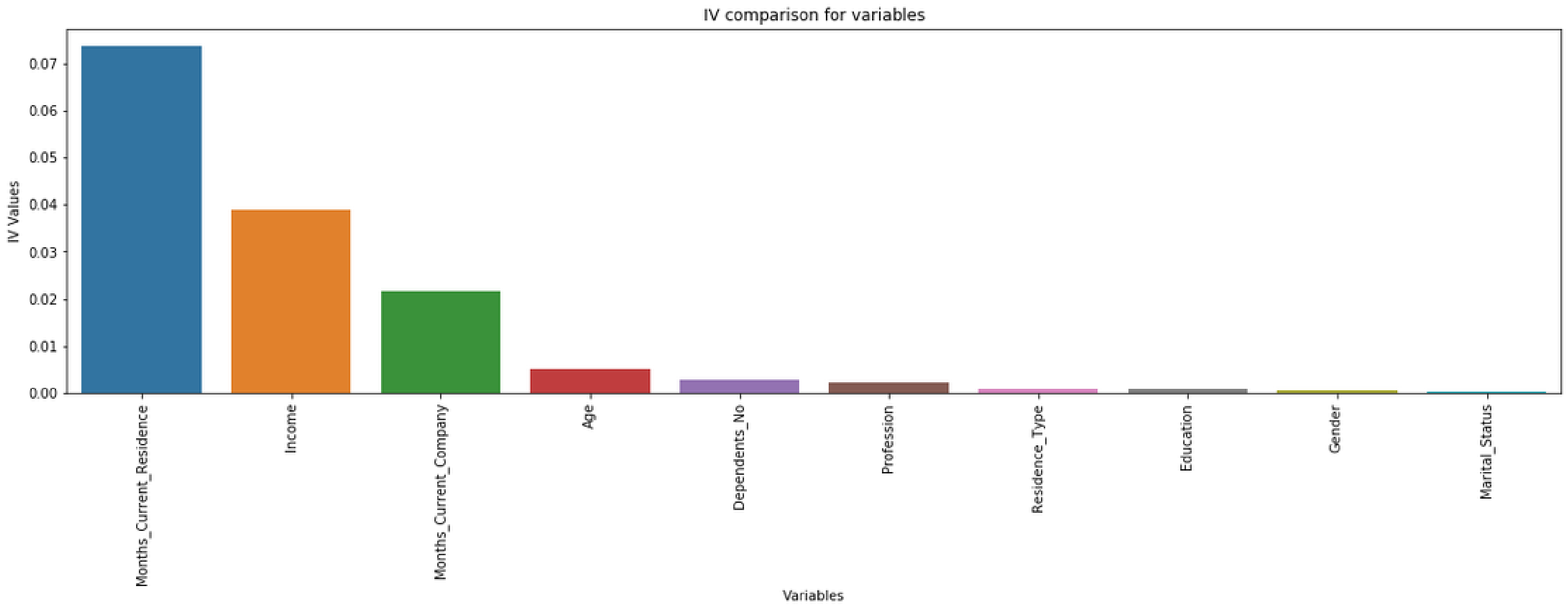
BFS CAPSTONE PROJECT

- Use both the original clean data-set and WoE transformed data set of demographics separately to prepare data models.
- For this bi-logit problem model preparation, begin with simple models like Logistic Regression model with RFE and step by step move on to relatively complex models like Logistic Regression with Regularization, Random Forest

BFS CAPSTONE PROJECT

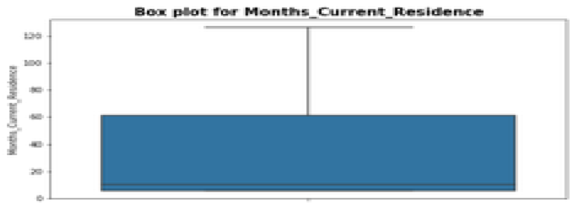
IV Values for demographic data in descending order of importance

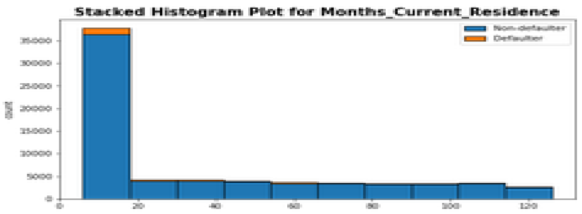
	Variable	IV
0	Months_Current_Residence	0.073689
0	Income	0.039013
0	Months_Current_Company	0.021577
0	Age	0.004955
0	Dependents_No	0.002823
0	Profession	0.002281
0	Residence_Type	0.000936
0	Education	0.000765
0	Gender	0.000562
0	Marital_Status	0.000143

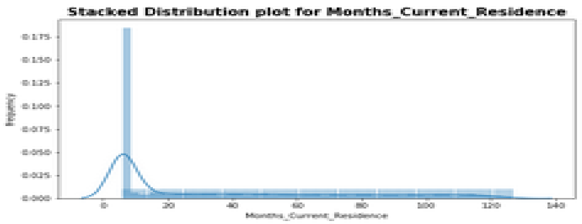


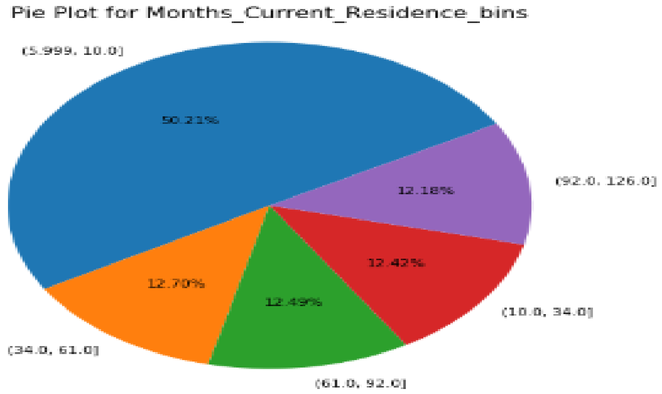
BFS CAPSTONE PROJECT

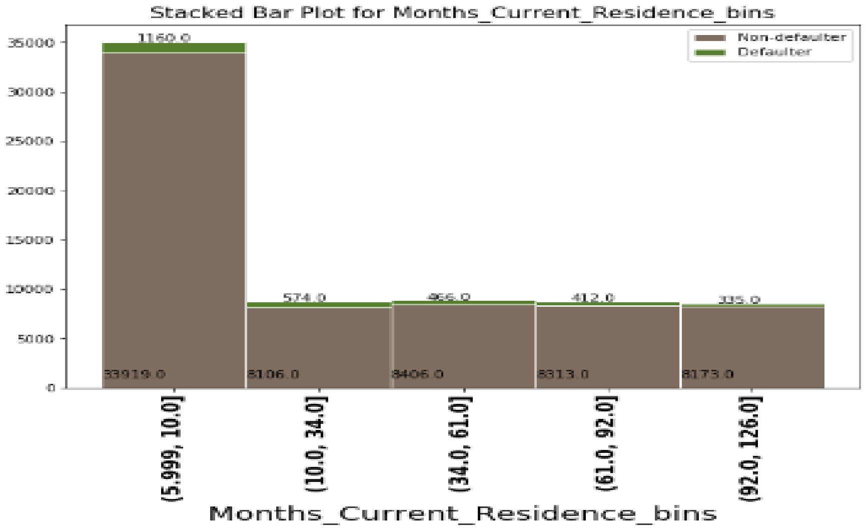
Understanding Months_Current_Residence as predictor variable

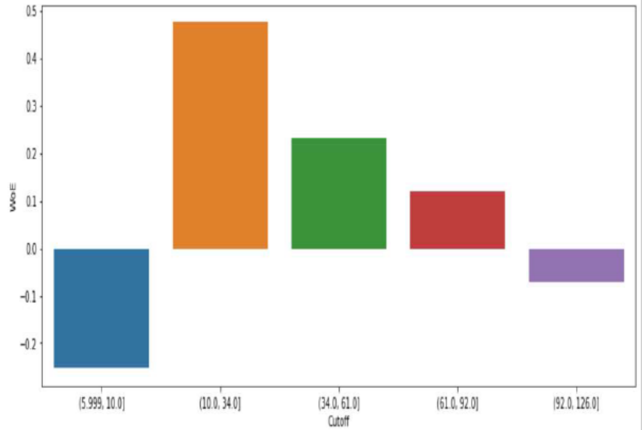


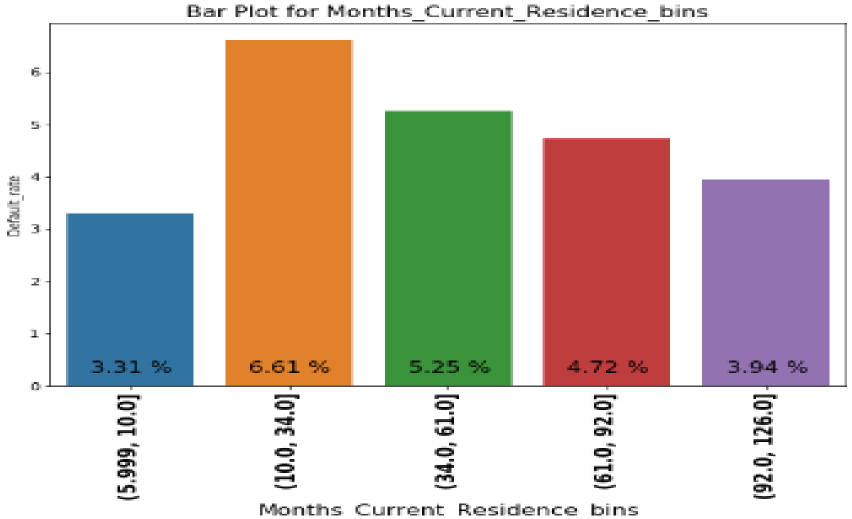








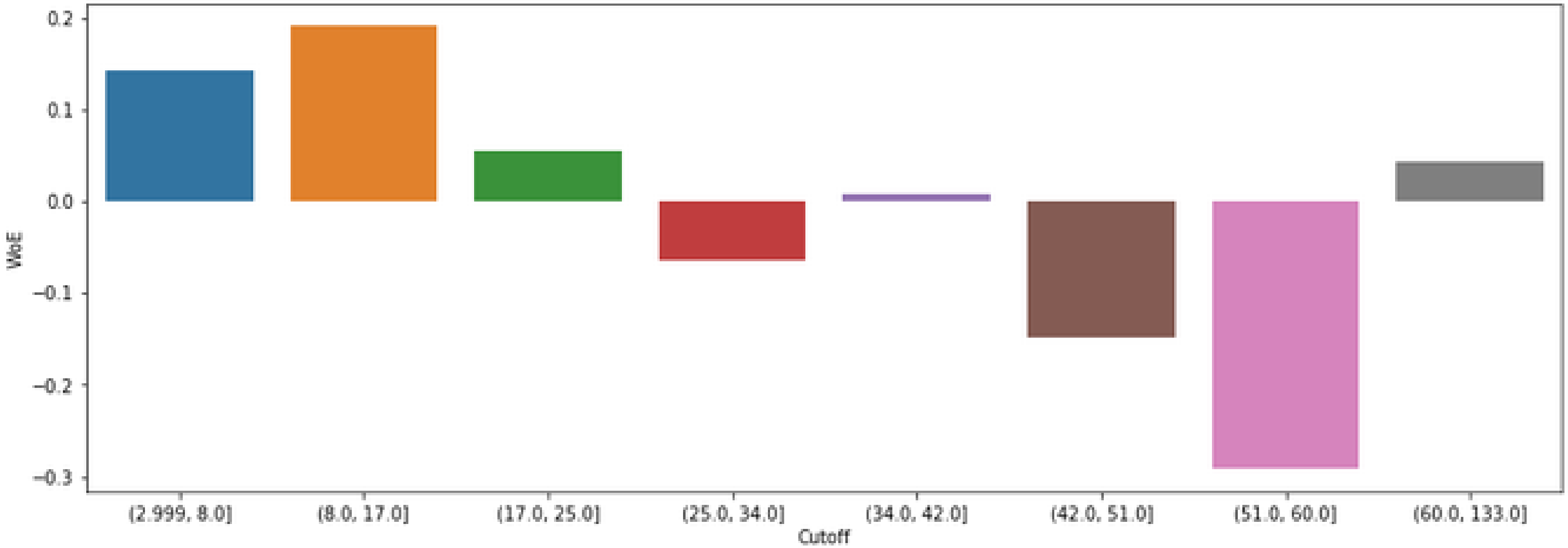




- WoE value decreases as the number of months in current residence increase
- We also see that the default rate increases then decreases across bins in the bar plot.

BFS CAPSTONE PROJECT

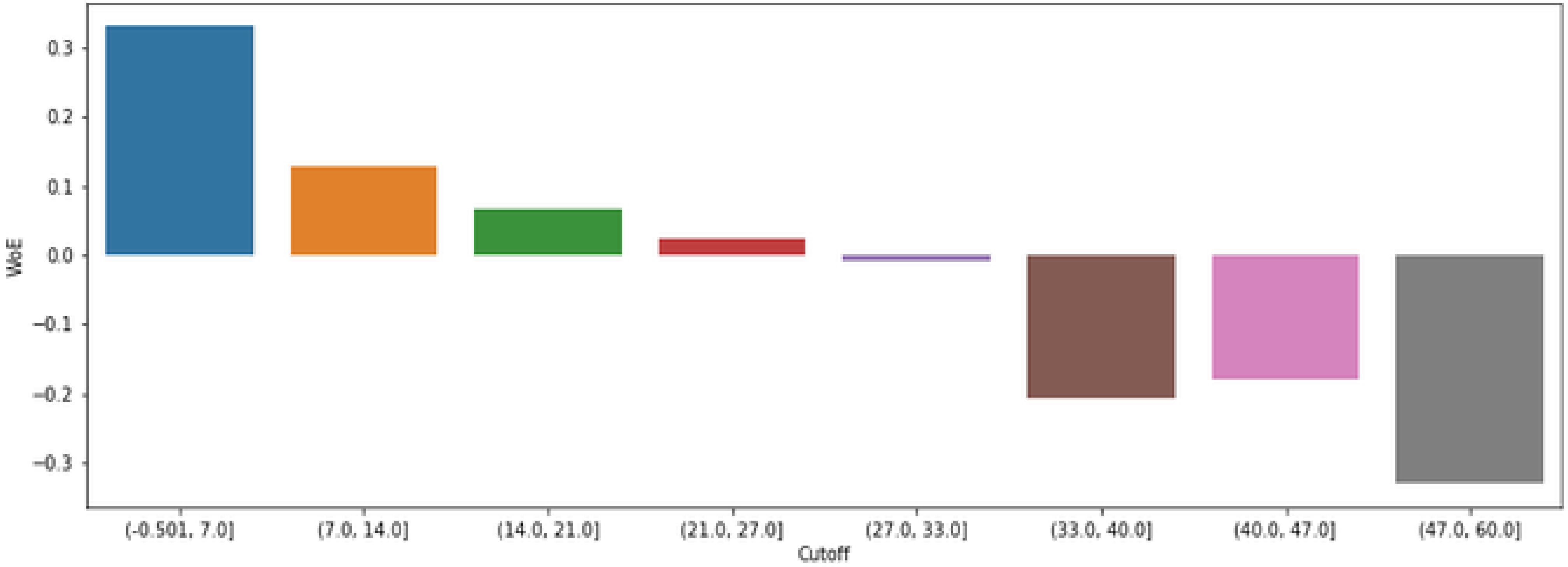
Understanding Months_Current_Company as predictor variable



- The trend of decrease in WoE with increase in Months_Current_Company is evident with some exceptions in the above plot.
- People who are relatively new in there company 2-18 months has higher WOE means that they govern the defaut rate more.

BFS CAPSTONE PROJECT

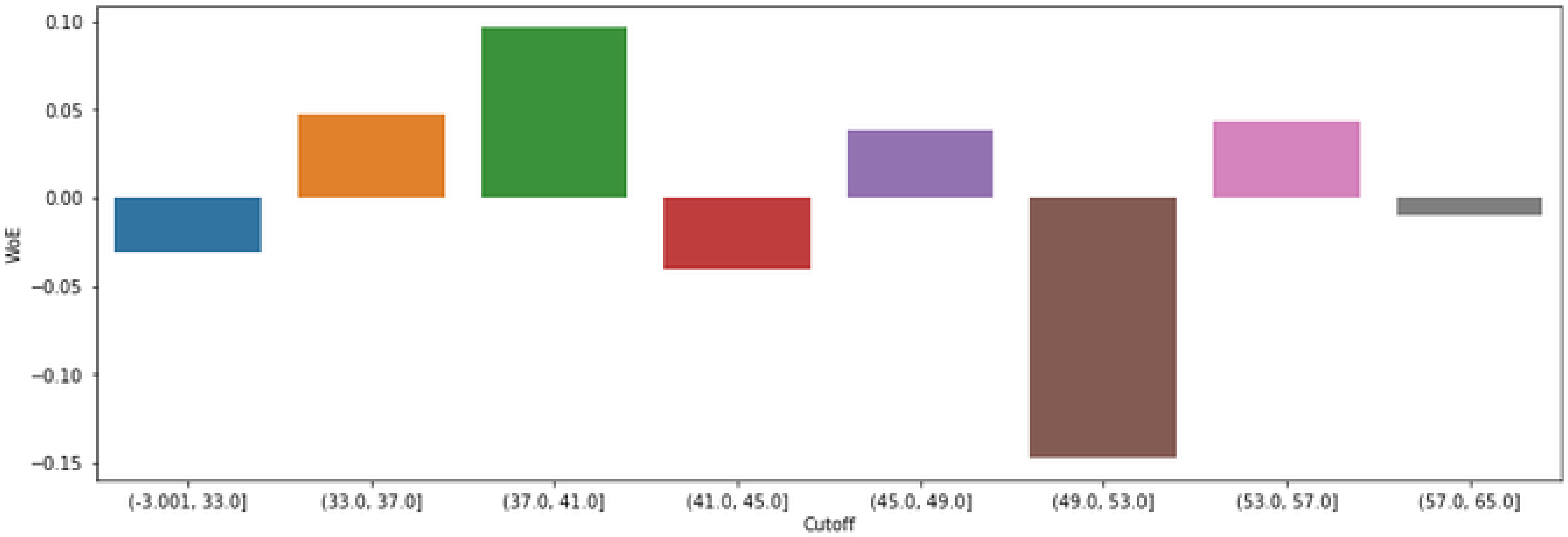
Understanding income as predictor variable



- The trend of decrease in WoE values with increase in income bins is very much evident from the above plot. Means that people with lower income has more weight of evidence and governs the default rate more.

BFS CAPSTONE PROJECT

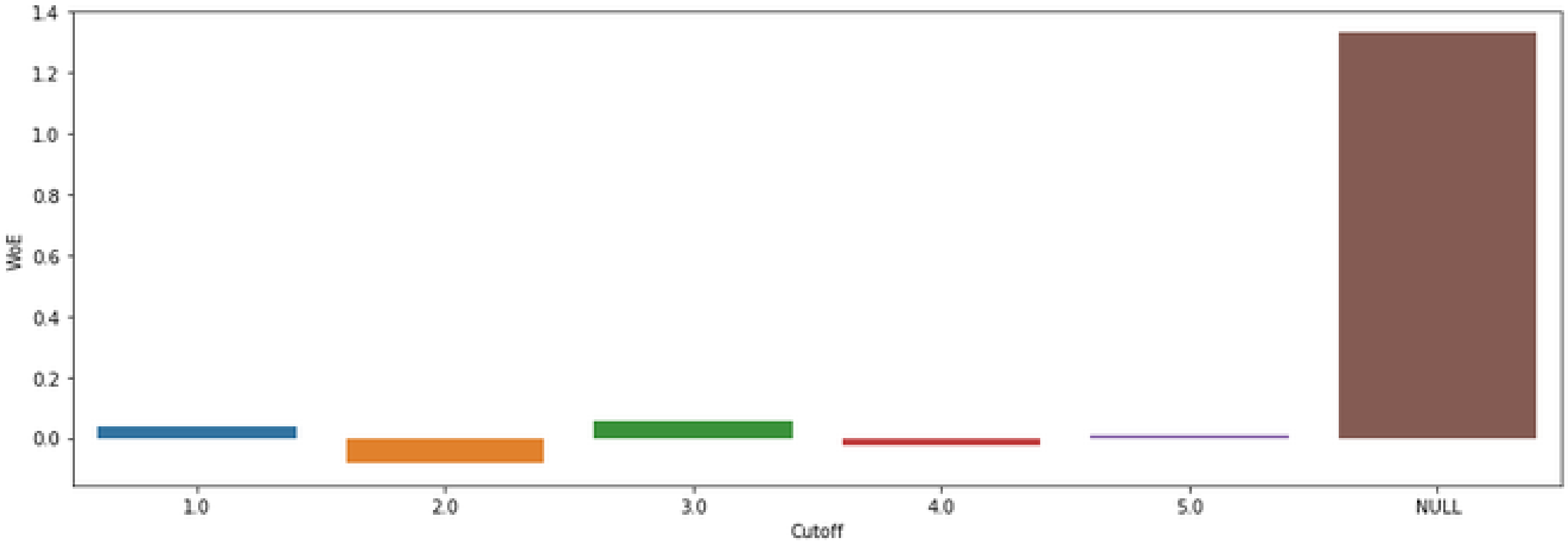
Understanding Age as predictor variable



- There is no clear trend of WOE in Age column

BFS CAPSTONE PROJECT

Understanding Months_Current_Residence as predictor variable

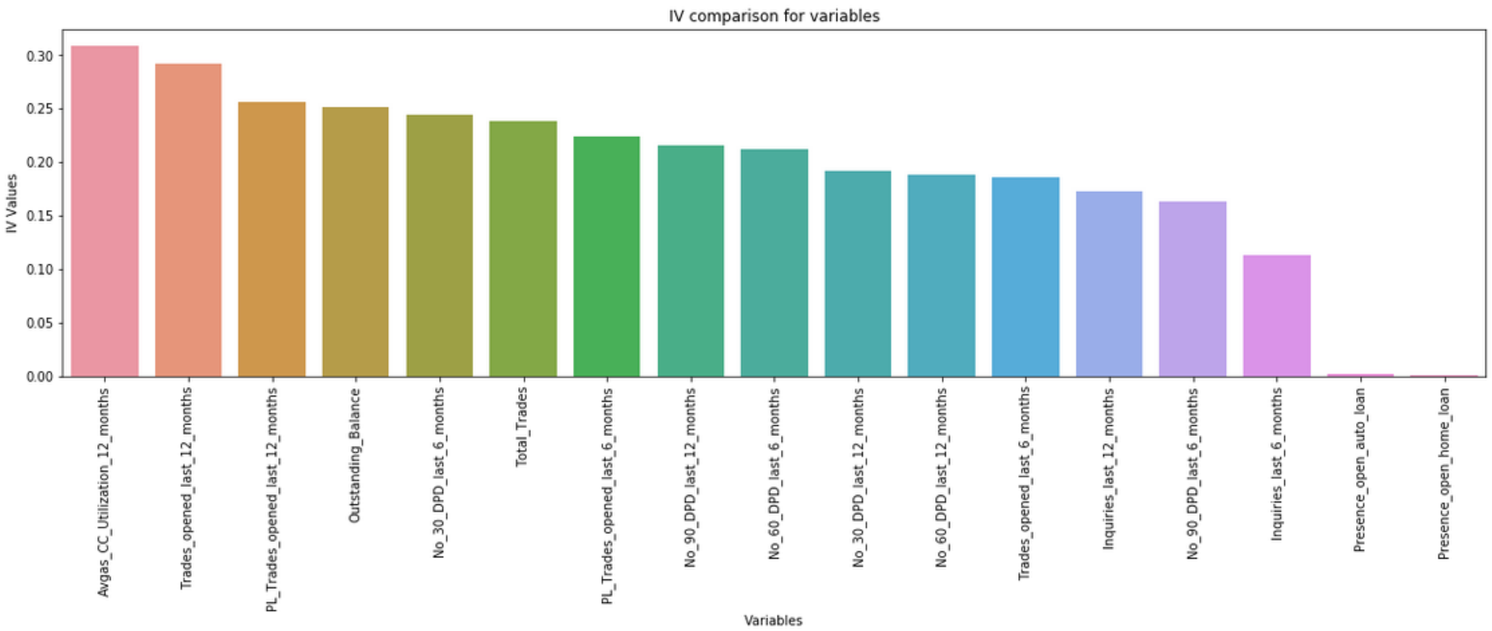


- Being the fifth important IV variable, the trend is not much observed with rtespect to variable Dependents_No.

BFS CAPSTONE PROJECT

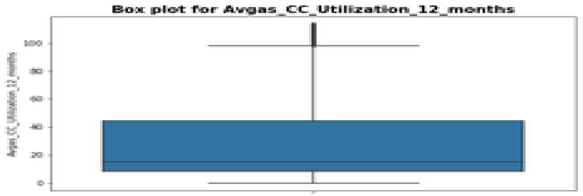
IV Values for Credit Bureau data in descending order of importance

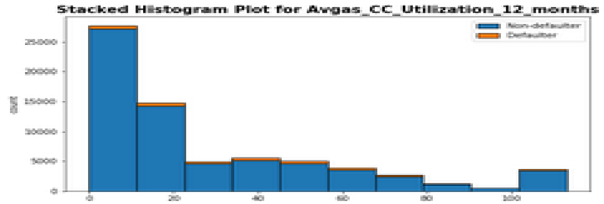
	Variable	IV
0	Avgas_CC_Utilization_12_months	0.308880
0	Trades_opened_last_12_months	0.291790
0	PL_Trades_opened_last_12_months	0.256190
0	Outstanding_Balance	0.251292
0	No_30_DPD_last_6_months	0.244460
0	Total_Trades	0.238446
0	PL_Trades_opened_last_6_months	0.224342
0	No_90_DPD_last_12_months	0.216015
0	No_60_DPD_last_6_months	0.211539
0	No_30_DPD_last_12_months	0.191285
0	No_60_DPD_last_12_months	0.188539
0	Trades_opened_last_6_months	0.186282
0	Inquiries_last_12_months	0.172768
0	No_90_DPD_last_6_months	0.162983
0	Inquiries_last_6_months	0.112865
0	Presence_open_auto_loan	0.001665
0	Presence_open_home_loan	0.000463

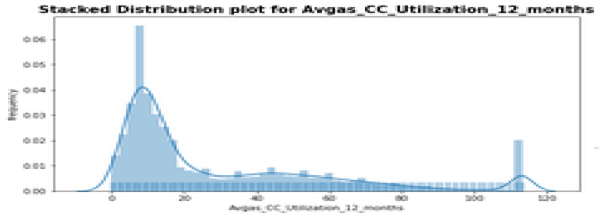


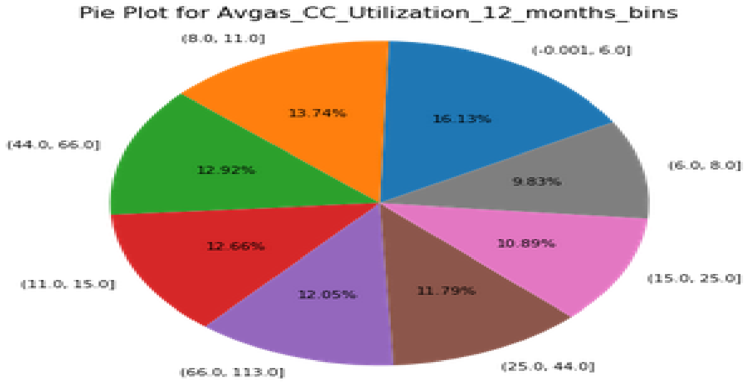
BFS CAPSTONE PROJECT

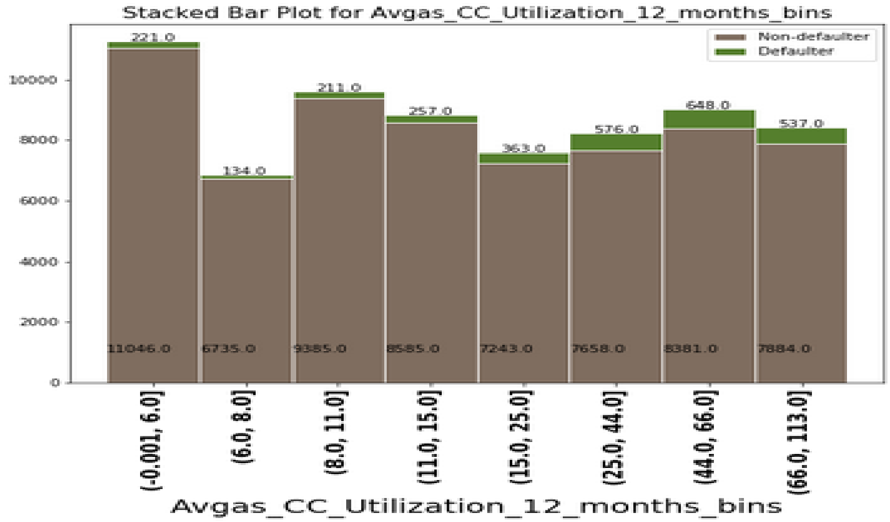
Understanding Avgas_CC_Utilization_12_months as predictor variable

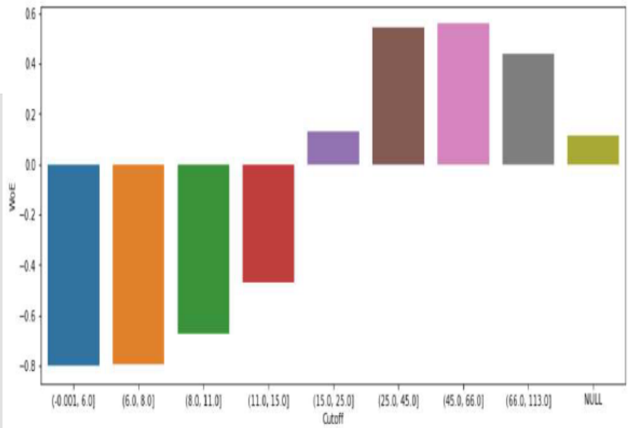


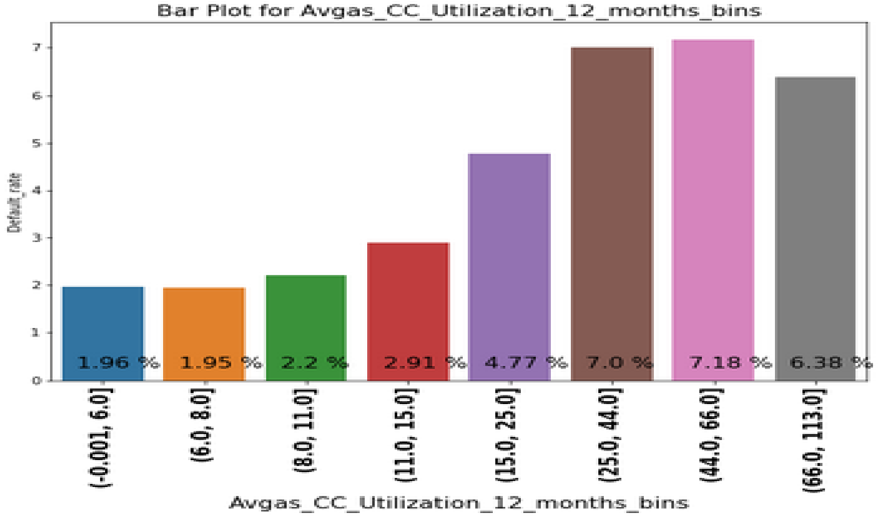








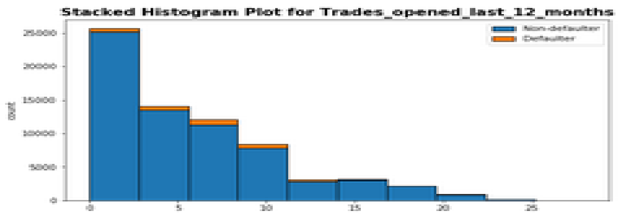


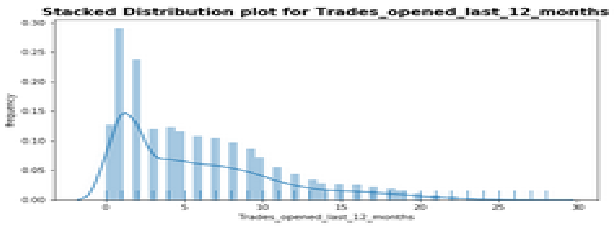


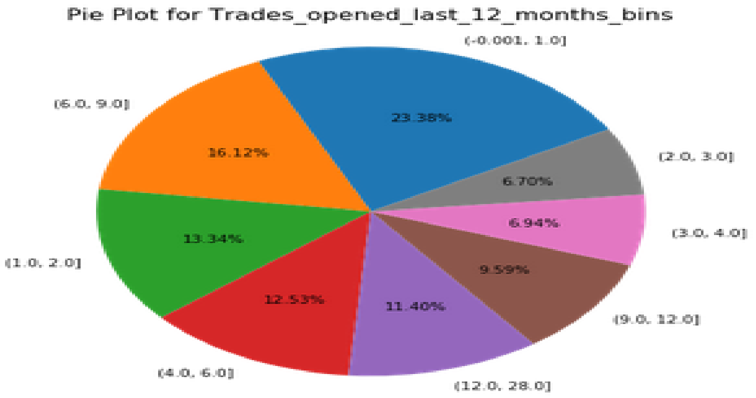
- WoE value increases as the **Avgas_CC_Utilization_12_months** increase.
- Also it is very clear with default rate vs the avg CC utilization bar plot that with more usage the chances of defaulting increases.

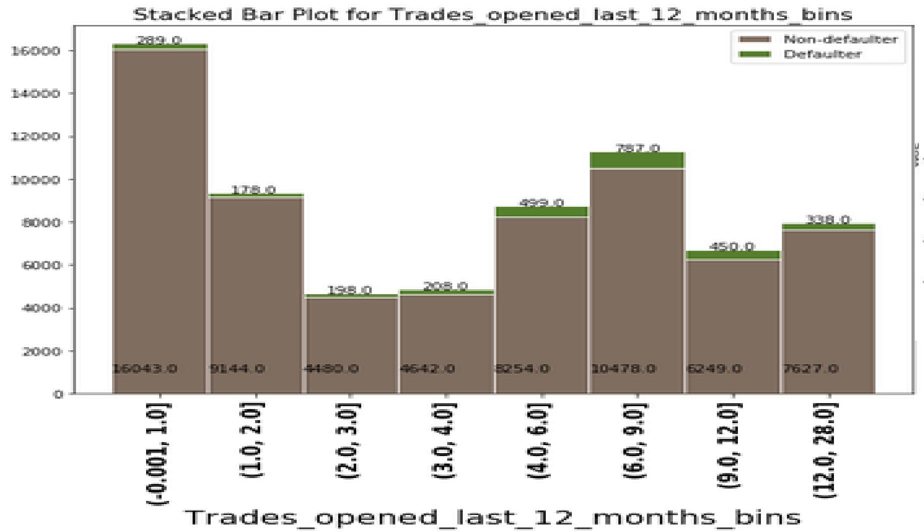
BFS CAPSTONE PROJECT

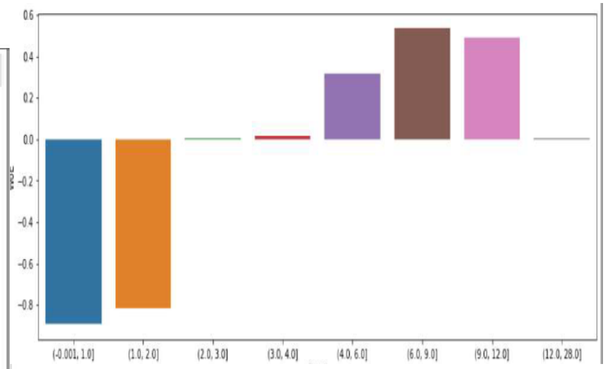
Understanding Trades_opened_last_12_months as predictor variable

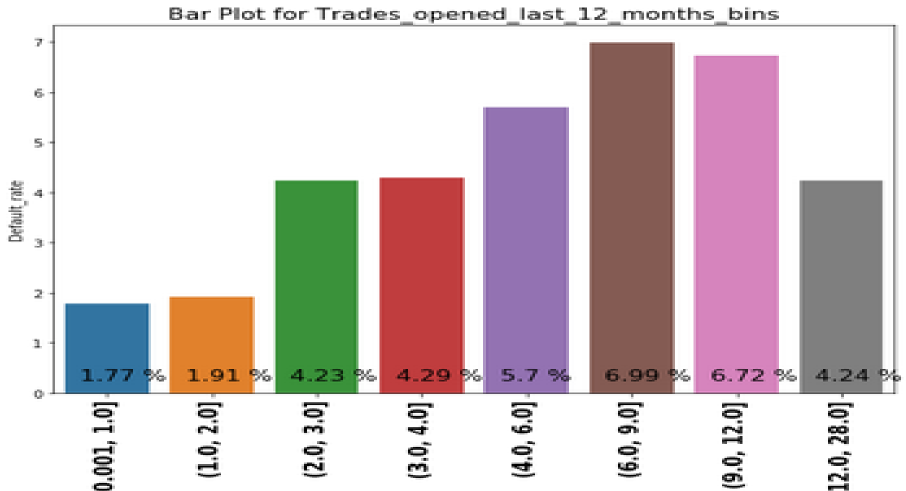








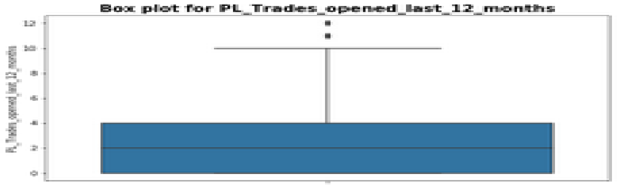


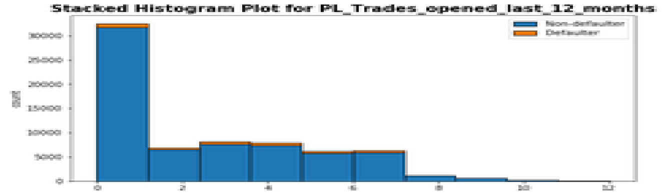


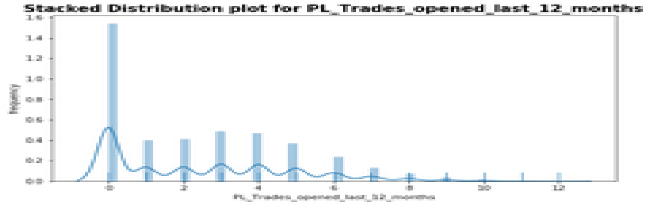
- WoE value increases as the **Trades_opened_last_12_months** increase.
- Also similar trend is observer across bins in Bar Plot

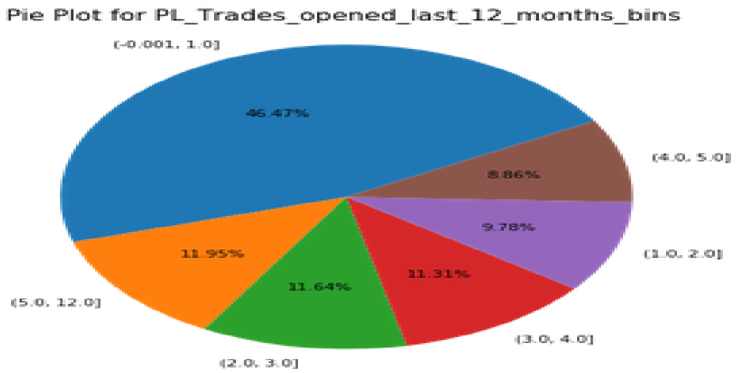
BFS CAPSTONE PROJECT

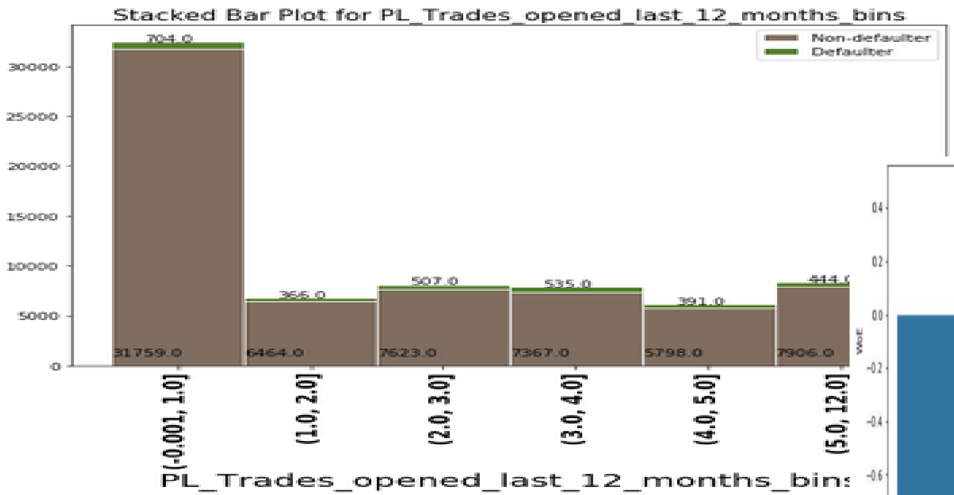
Understanding PL_Trades_opened_last_12_months as predictor variable

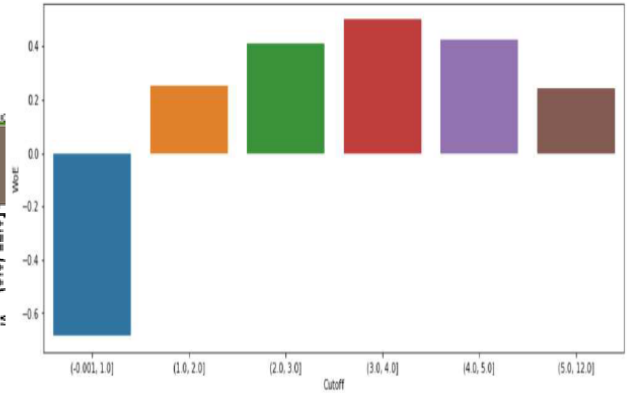


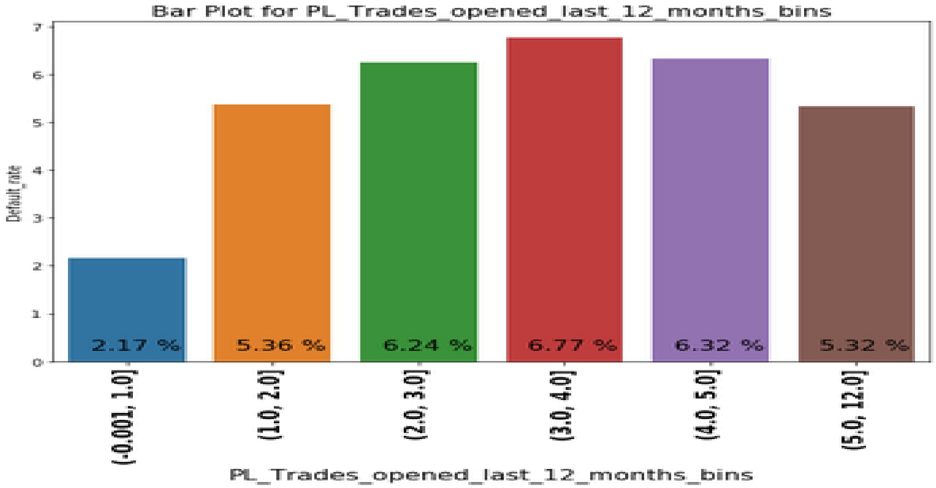






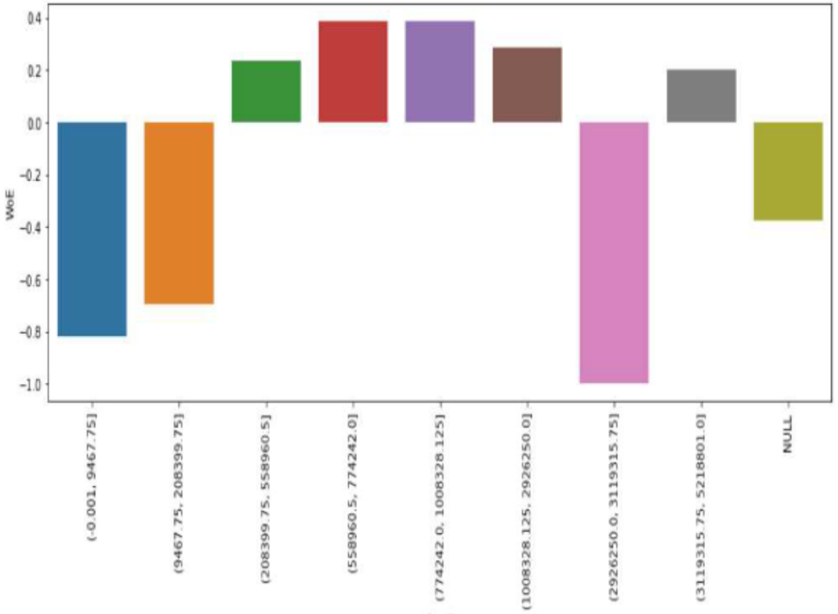


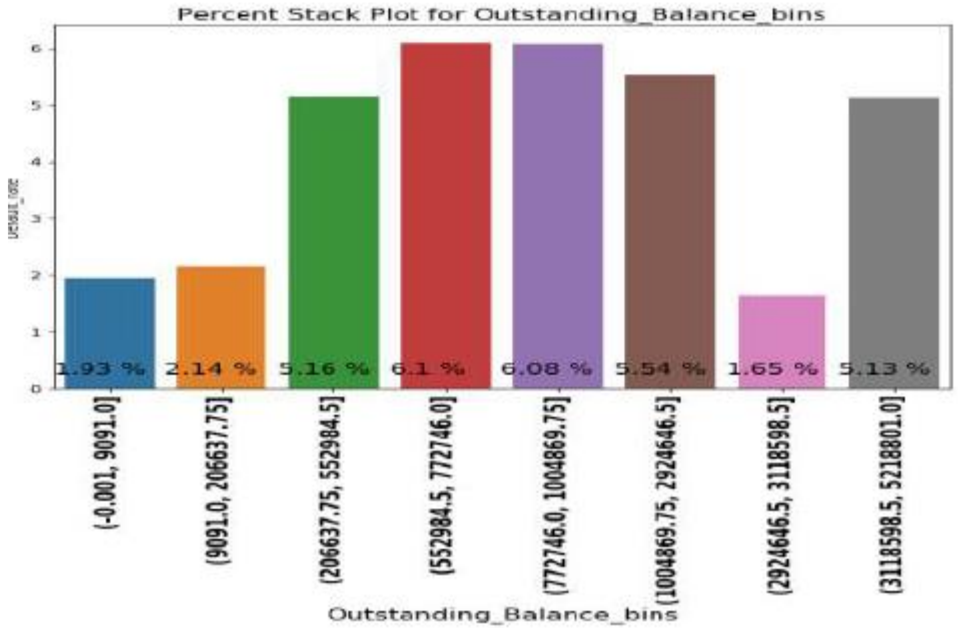




- WoE value increases as the **PL_Trades_opened_last_12_months** decrease
- Similar trend is observed in the bins across the bar plot too

Understanding Outstanding_Balance as predictor variable

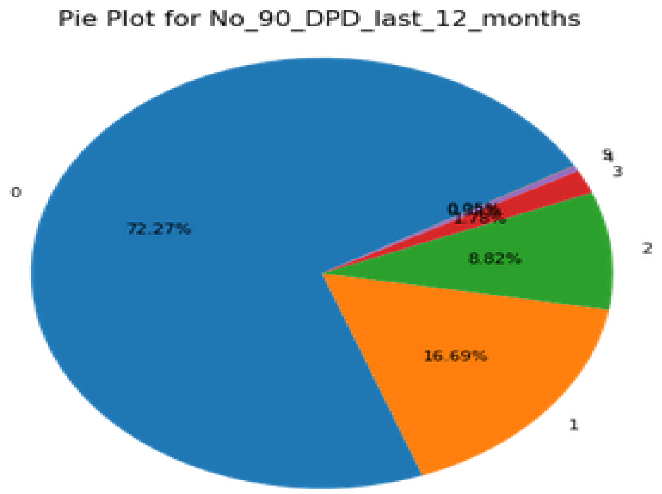


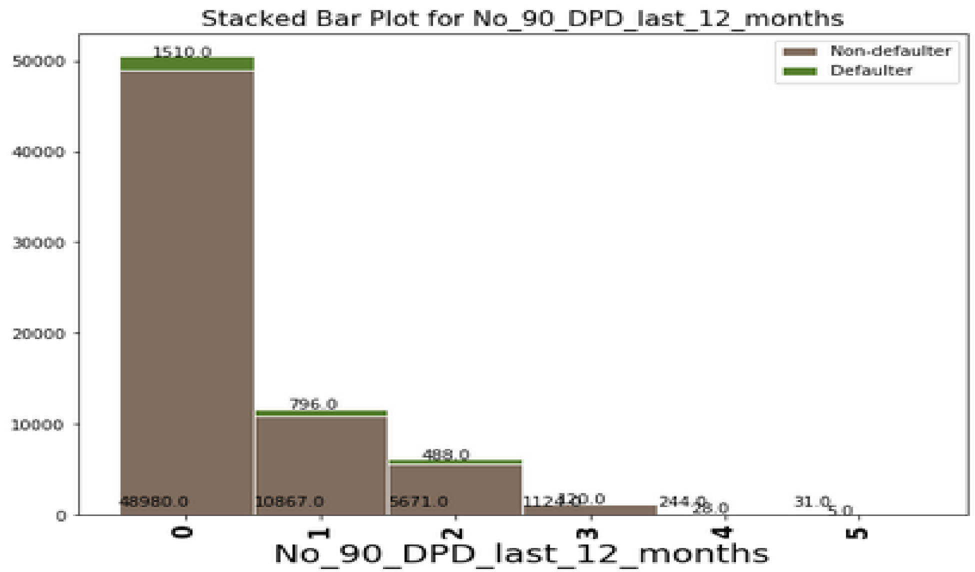


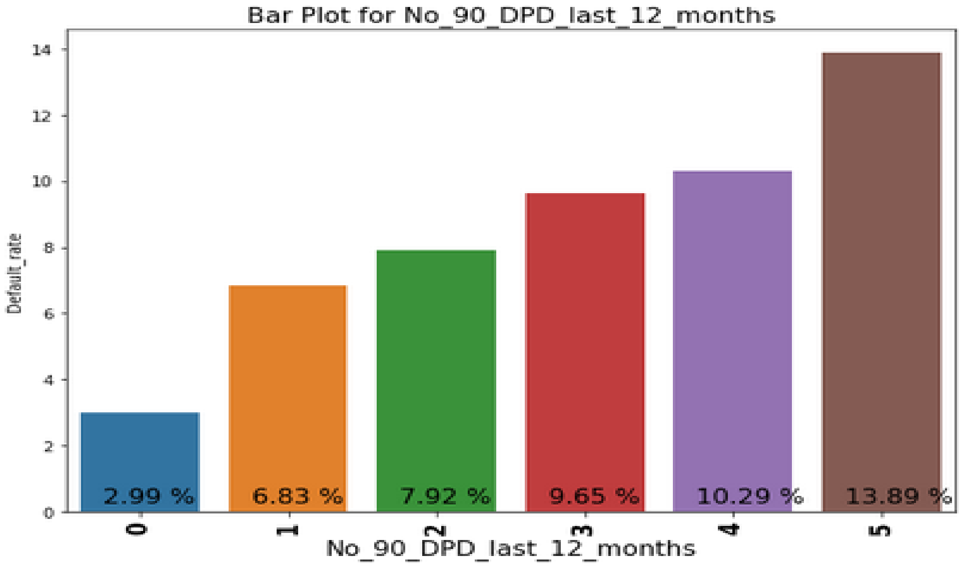
- We can clearly see with the increase in Outstanding Balance, the default rate % and WOE increases.
- Exception :- we see this trend across bins, but bin (2924646-3118598] has the exception where the default rate falls to 1.65 % strangely.

BFS CAPSTONE PROJECT

Understanding No_30_DPD_last_6_months as predictor variable







- 72.27 % of people have defaulted 0 times in paying dues in 90 days past due in 12 months.The default rate is 2.99%
- We see the default rate percent increases with the increasing value in No_90_DPD_last_12_months. The trend is stedy

Model Building

Data Sets chosen for model

Demographics data set

Demographics WoE transformed data set

Combined (Demographics and Credit Bureau) data set

Combined (Demographics and Credit Bureau) WoE transformed data set

3 models for each data set

Logistic regression with RFE

Decision tree

Random forest

Total 12 models

Model Building Results for Demographics Dataset

UpGrad

Model	Accuracy (Test data)	Precision (Test data)	Recall (Test data)	Precision (Rejected app. data)	Recall (Rejected app. data)
Logistic Regression	78%	5.46%	26%	100%	58%
Decision Tree	56.35%	5.18%	54.13%	100%	51%
Random Forest	62.20%	6.17%	56.06%	100%	30%

Hyperparameters chosen to tune the model:

- Logistic Regression
  - AUC: 0.57
  - Cut off point: 0.05
- Decision tree
  - max_depth : 5
  - min_samples_leaf : 100
  - min_samples_split : 50
  - Criterion : gini
- Random Forest
  - max_depth : 4
  - min_samples_leaf : 350
  - min_samples_split : 400
  - n_estimators : 1000
  - max_features : 10

Model Building Results for Demographics-WoE transformed Dataset

Model	Accuracy (Test data)	Precision (Test data)	Recall (Test data)	Precision (Rejected app. data)	Recall (Rejected app. data)
Logistic Regression	73%	6.32%	39.42%	100%	58.31%
Decision Tree	62.48%	5.80%	52.46%	100%	57.68%
Random Forest	63.41%	5.96%	52.57%	100%	73.47%

Hyperparameters chosen to tune the model:

- Logistic Regression
  - AUC: 0.60
  - Cut off point: 0.05

- Decision tree
  - max_depth : 5
  - min_samples_leaf : 200
  - min_samples_split : 50
  - Criterion : gini

- Random Forest
  - max_depth : 4
  - min_samples_leaf : 300
  - min_samples_split : 450
  - n_estimators : 1000
  - max_features : 2

Model Building Results for Combined (Demographics and Credit Bureau) data set

Model	Accuracy (Test data)	Precision (Test data)	Recall (Test data)	Precision (Rejected app. data)	Recall (Rejected app. data)
Logistic Regression	68%	7%	51%	100%	93.96%
Decision Tree	53.90%	6.52%	74.63%	100%	95.29%
Random Forest	57%	6.79%	72.36%	100%	99.92%

Hyperparameters chosen to tune the model:

- Logistic Regression
  - AUC: 0.67
  - Cut off point: 0.05

- Decision tree
  - max_depth : 5
  - min_samples_leaf : 50
  - min_samples_split : 200
  - Criterion : gini

- Random Forest
  - max_depth : 4
  - min_samples_leaf : 350
  - min_samples_split : 400
  - n_estimators : 900
  - max_features : 15

Model Building Results for Combined (Demographics and Credit Bureau)-WoE dataset

Model	Accuracy (Test data)	Precision (Test data)	Recall (Test data)	Precision (Rejected app. data)	Recall (Rejected app. data)
Logistic Regression	65%	7.15%	61.03%	100%	97.89%
Decision Tree	53.87%	6.49%	73.34%	100%	99.78%
Random Forest	57.28%	6.75%	70.54%	100%	99.85%

Hyperparameters chosen to tune the model:

- Logistic Regression
  - AUC: 0.67
  - Cut off point: 0.05

- Decision tree
  - max_depth : 5
  - min_samples_leaf : 200
  - min_samples_split : 50
  - Criterion : entropy

- Random Forest
  - max_depth : 4
  - min_samples_leaf : 350
  - min_samples_split : 400
  - n_estimators : 900
  - max_features : 10

Model Evaluation Techniques

UpGrad

Basis of Evaluation To Get Optimal Model for each type:

- The objective of the model is to optimize **Sensitivity / Recall** .
- Confusion matrix prepared for each model.
- Sensitivity, specificity, accuracy curve for Logistic Regression models.
- AUC-ROC curve for the Logistic Regression models using cut-off values for each model.
- Plots showing optimized values for Regularization hyperparameter.
- Use of GridSearchCV and plotting its results for all models.
- Gini-Index needs to be evaluated for Tree based models like decision tree and random forest.
- Within each model type evaluation using GridSerach based on **recall** values should be done to get models with optimized hyperparameters.
- For evaluation among models, the dataset for rejected applications (with performance tag missing), which were assumed as potentially defaulters should be considered for evaluations. Ideally, the output for all these applications should be defaulters.

Model Evaluation

- **Final Model chosen :** Random Forest model on WoE transformed Combined data set.

- **Reasons for Choosing this Model :**
- Have high test recall
- The models were able to reject almost all the manually rejected applications.(As high as 92%)
- Random Forest is an ensemble model which means the diversity is intact, as it incorporates various diverse models adding almost all available information.
- The model is very stable (especially the WOE one). The use of WoE values and multiple decision trees provide this stability. The WoE values are bound to show less variance making the model stable.
- The model did not overfit the data
- The model is expected **not** to overfit on any data.

Application Scorecard

- **Model chosen :** Logistic regression with Lasso regularization on WOE transformed Combined data set

- **Scorecard Evaluation variables and formulae :**

  target_score = 400

  target_odds = 10

  pts_double_odds = 20

  factor = pts_double_odds / log10(2)

  offset = target_score - factor × log10(target_odds)

  scorecard['logit’] = ∑ (β×WoE) + α

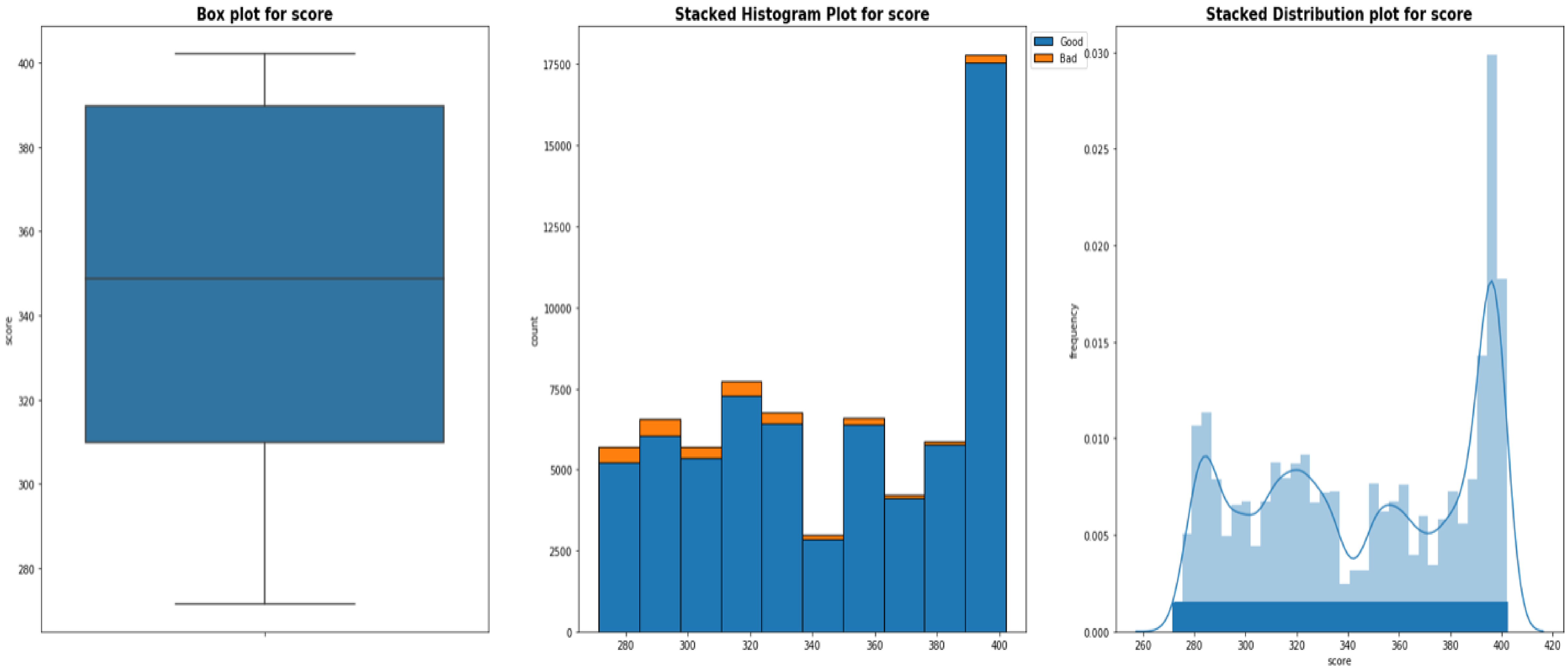
  (where β — logistic regression coefficient and α — logistic regression intercept)

  Finally, **scorecard['score'] = offset - factor × scorecard['logit']**

Application Scorecard Variation Plots

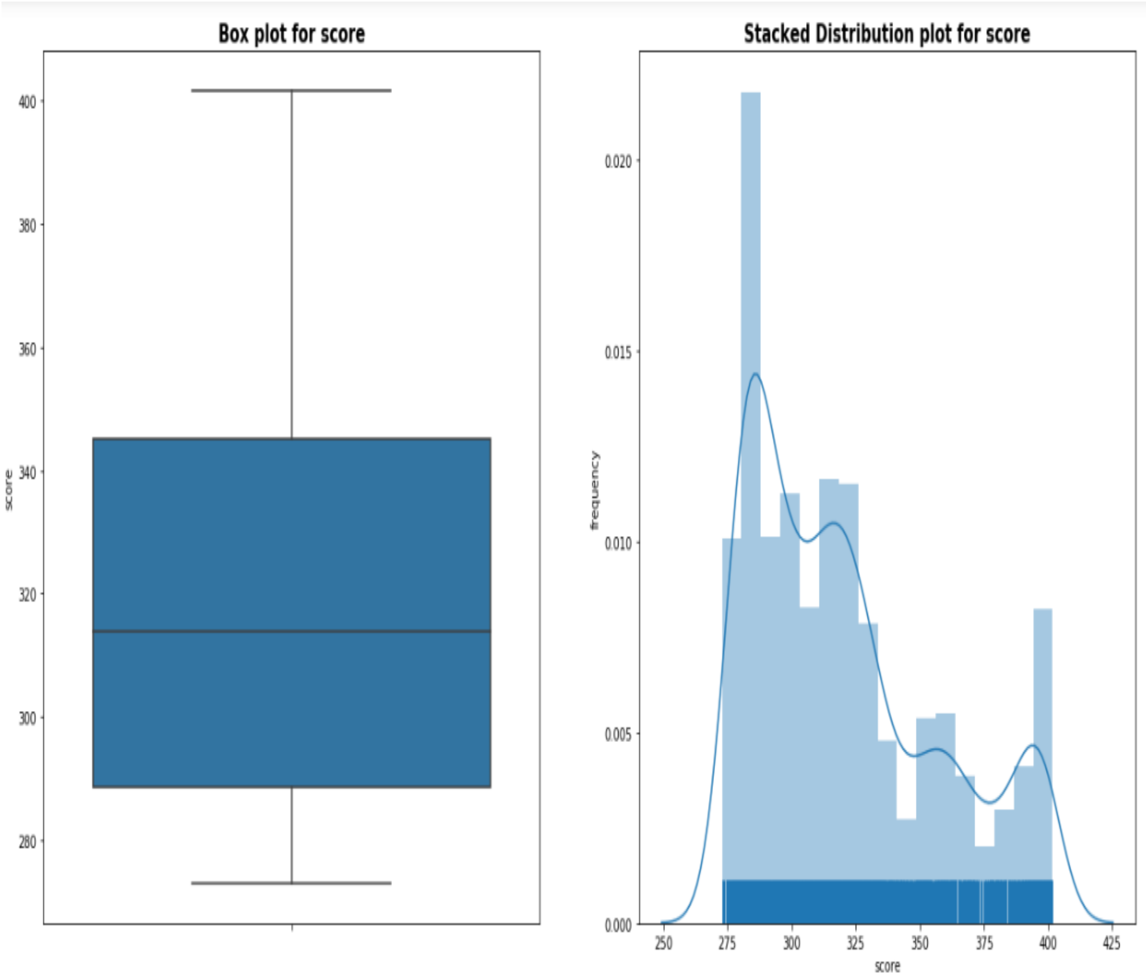
UpGrad

Overall Scorecard Variation Plots

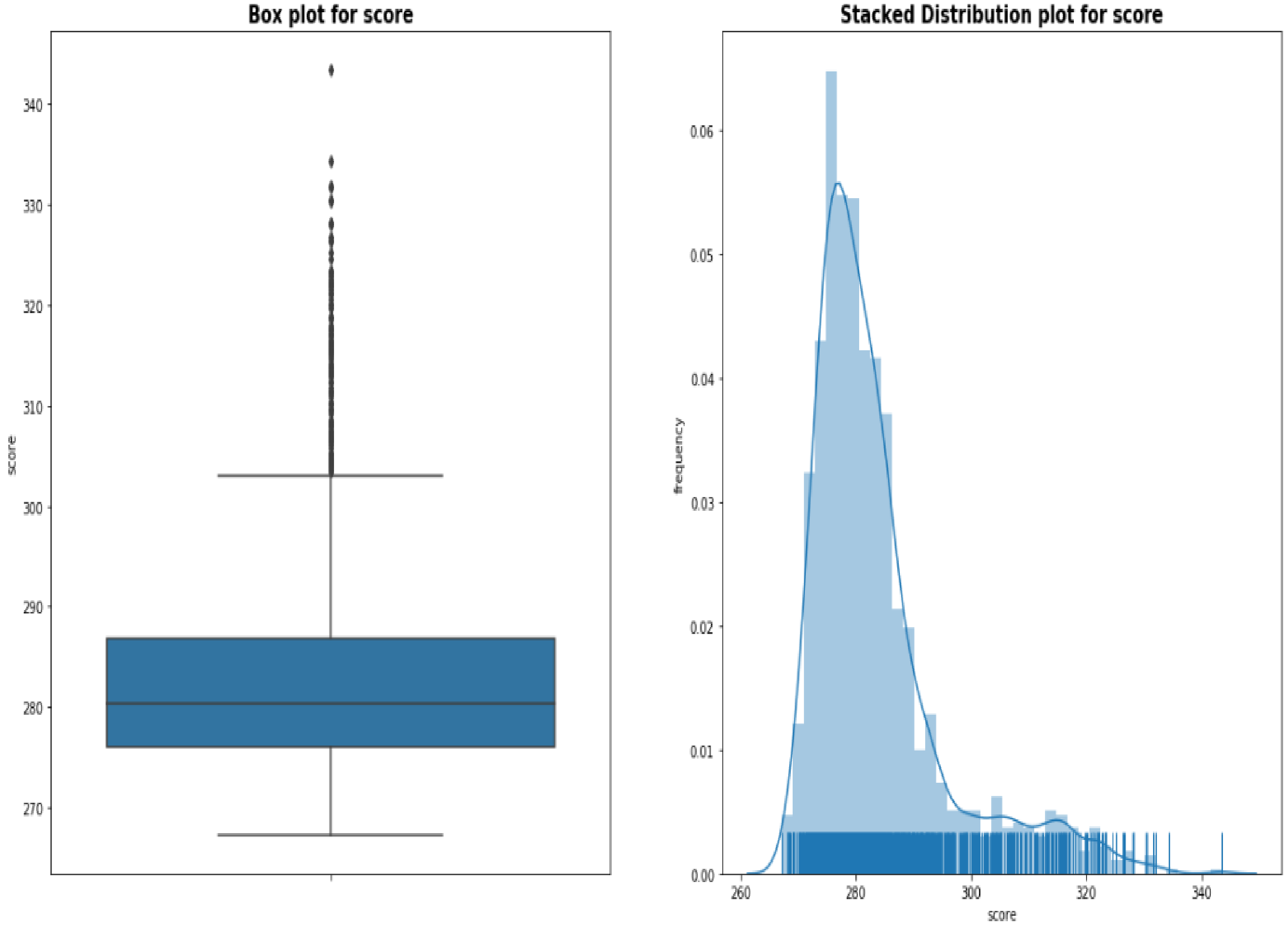


Application Scorecard Variation Plots

Defaulters Scorecard Plots



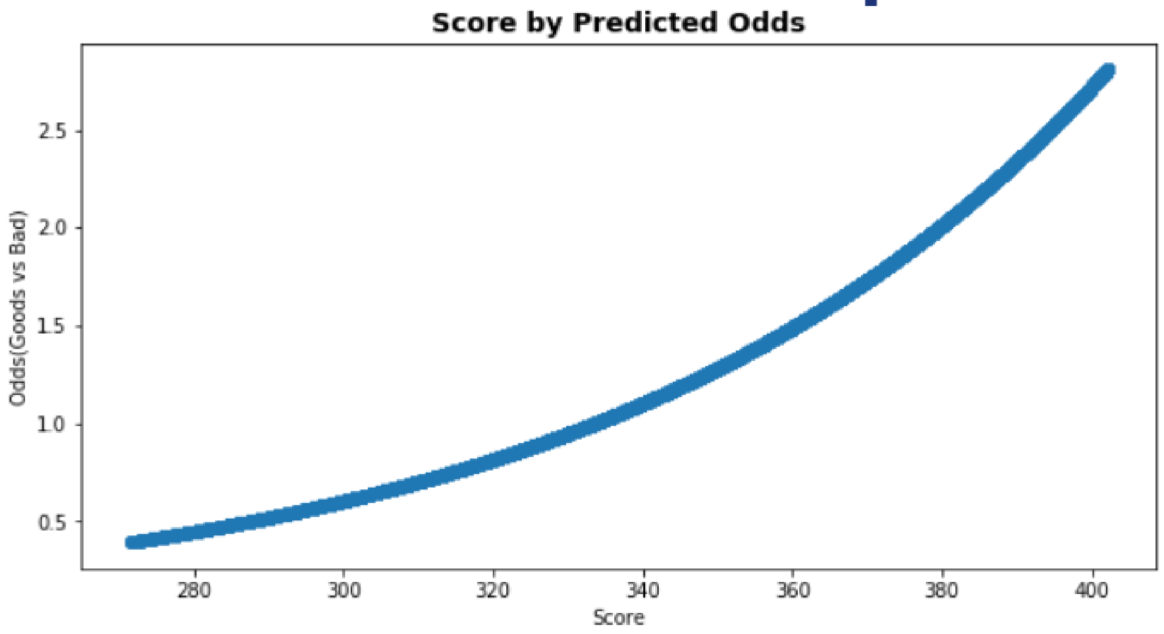
Rejected Population Scorecard Plots

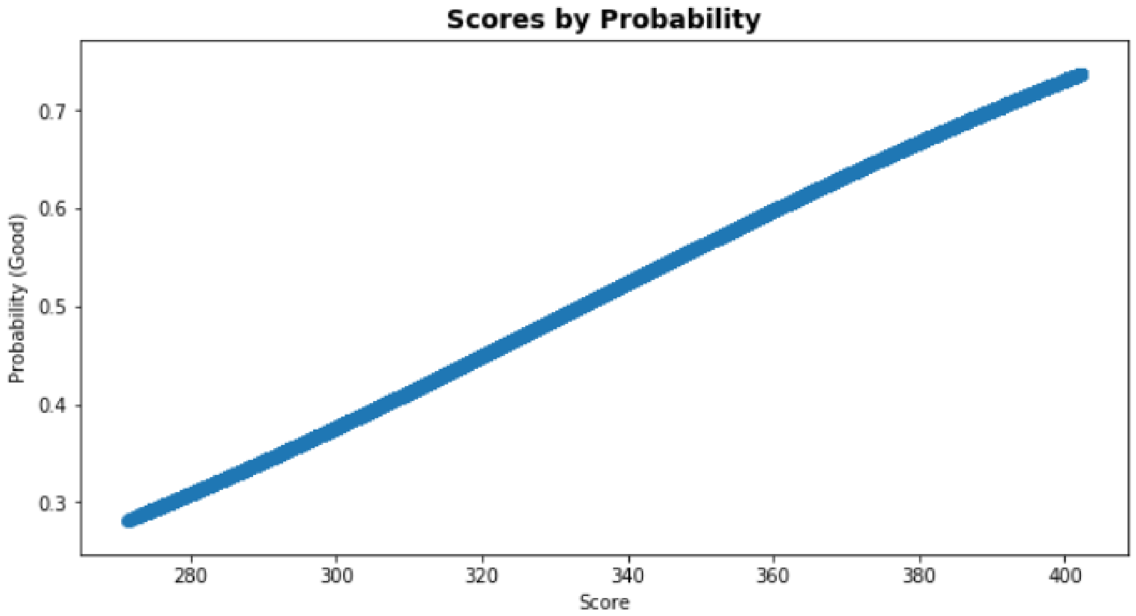


Application Scorecard

- **Cut-Off : 330**

- **Reason for Choosing this Cut-Off**
  - Recommended Strategy "Acquire the right customers" (a bit conservative owing to previous losses)
  - Caters almost all the rejected population
  - Prevents two-third of the default cases.
  - Impacts one-third of the approved cases.
  - A less cut-off would dilute the purpose of the model.
  - A higher cut-off will impact the business of the bank.
  - Discussions and recommendations by CredX Operations and Strategy team may change this cut-off.





Benefits of ML model

- Our objective is to minimize "Net Credit Loss" from Profit & Loss perspective.
- With ML model we get good discriminatory power over pre-identifying risky costumers.
- Reduces the cost spent on Underwriters which rejects the application by reviewing manually.
- Reduces time for processing of application requests as Underwriters are not involved and the process is automated.
- Prevents manual error made by Underwriters.
- Any kind of bias can be easily removed which creeps in due to sex, race or religion.
- Scorecard and cut-off provides clear instructions as how to proceed with application. Decision making is also fast.

Financial Risk in Current Operations

- Total number of applications = 71295
- Credit Card given to applicants = 69870
- Customers that made Credit Loss = 2948
- Assumptions on unit Applications :
  - Acquisition Cost + Credit Report Cost –100 INR
  - Calling Cost – 10 calls avg × 10p = 1 INR
  - Operations {Agents + Infra + Others) = 1000 INR approx.
  - Credit Default = Rs 48,000 average
- For every customer defaulted we are at a risk of loosing 50,000 INR (48k + 2k) on an average (assuming 50K is average credit line used by defaulter)
- Total Credit Loss for all customers = 50000 × 2948

  = 150 Million INR (approx.)

Financial Benefit of the ML Model

- The model giving a recall of 70% which means it is preventing 70% of losses.

Potential Loss prevented with using model = 0.7 × 150 Million INR

= 105 Million INR

Loss after prevention = 150 – 105 Million INR

= 45 Million INR

The Credit loss after applying the model has slashed to 30% compared to the Original Credit loss, which did not include using any model.

- Saving the amount paid to Underwriters for Credit application approval is added advantage.

- **Important Note :** There will be a tradeoff between the increase in approval rate and credit loss – increase of one will lead to increase of other. With this model the approval rate is bound to be less business to the bank and so will be the profits of the bank. However, profits are very small in margins (5-7%) as compare to the Principal amount in Credit Line. Hence percentage would be very small overall the Credit Line amount.

Thank You