CAPSTONE PROJECT

Credit risk analysis

Objective

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, our task is to help CredX identify the right customers using predictive models. 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 your project.

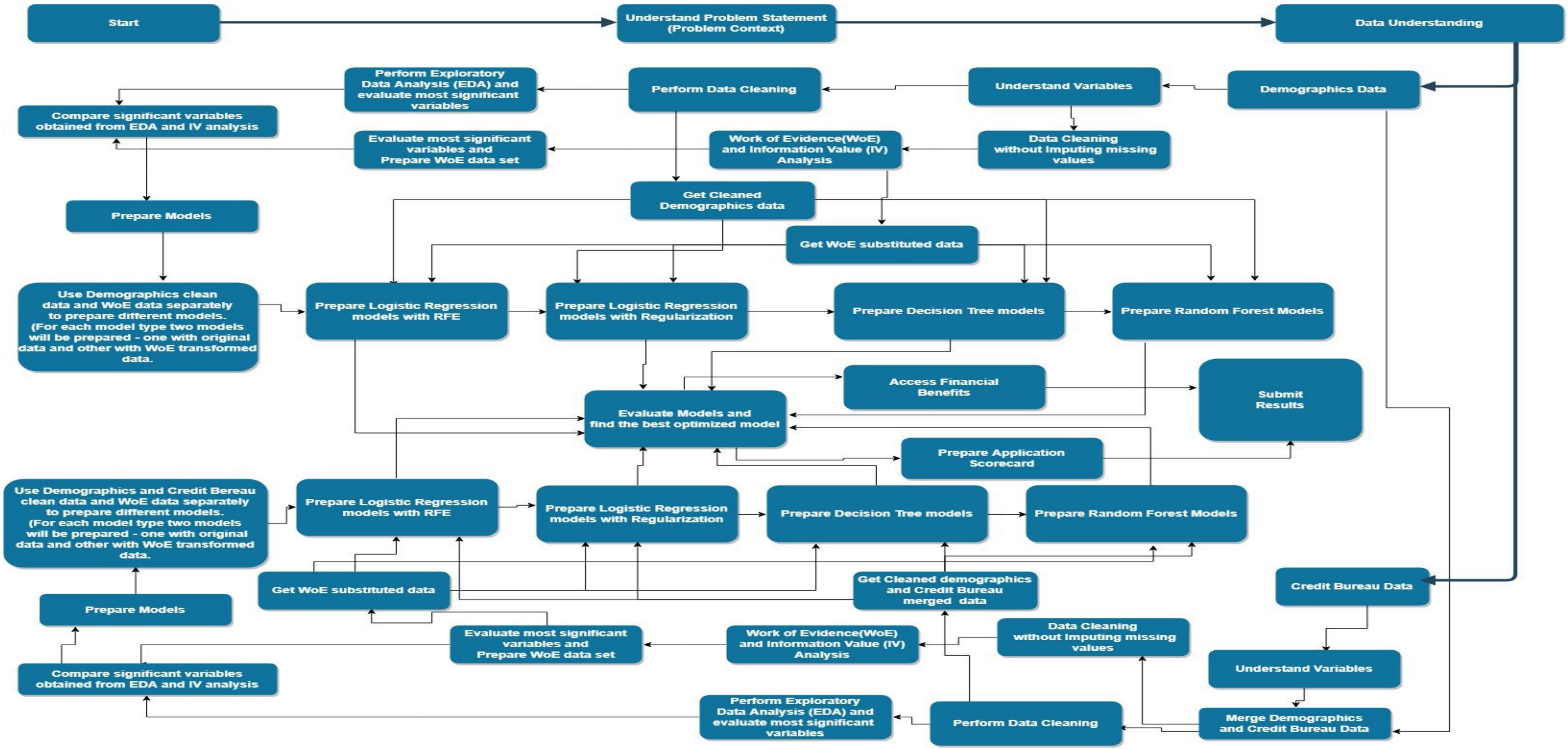
Steps to Problem Solving

- Understand the underlying problem and the domain for the problem.
- Understand the dataset provided (both Demographic and Credit Bureau) and inspect each attribute of both.
- Perform data cleaning on both the data sets. For input dataset of IV analysis, imputing values needs to be ignored.
- Choose each data-set individually and perform Exploratory Data Analysis to anticipate the significant variables.
- Work of Evidence (WoE) and Information Value (IV) analysis and preparing WoE transformed dataset.
  - ✓ Take Demographic data-set and perform WoE transformation and also obtain significant variables based on IV.
  - ✓ Merge Demographic data-set with Credit Bureau dataset and do WoE transformation and get significant variables.
- 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, Decision Tree, Random Forest etc. Following steps needs to be considered in each model building process :
  - ✓ Initially the dataset needs to be divided into Test and Train dataset.
  - ✓ There is a class imbalance in the dataset. This needs to be handled using balanced class during each model preparation.
  - ✓ Cross Validation needs to be done for each model.
  - ✓ Additional validation of data should be done on the dataset on the rejected applications (performance tag null) ignored for model building.
  - ✓ Hyperparameters for each type of models need to be optimized properly using GridSearch and model coming with optimized parameter should be chosen.

Steps to Problem Solving (continued …)

- The above step also needs to be carried out for merged data set of demographic and Credit Bureau.
- Evaluate all the models based on the following parameters :
  - ✓ Confusion matrix should be prepared for each model.
  - ✓ Sensitivity, specificity, accuracy curve for each model with different cut-offs.
  - ✓ AUC-ROC curve for the model using cut-off values for each model.
  - ✓ Precision and Recall curve for cut-off should be generated.
  - ✓ 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.
- The apt two stable and optimized models (with stable characteristics) - one for demographics and second for combined data needs to be chosen.
- On the basis of the chosen model and significant variables in the model, two application scorecard should be prepared for the two models.
- Access the financial benefits of the project by checking the underlying matrices that get optimized.
- Present all the results obtained in all the above steps to the management.

Problem Solving Methodology



Data Understanding

There are two data sets in this project: Demographic and Credit bureau data.

- **Demographic/application data:** This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- **Credit bureau data:** This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

Both files contain a performance tag, which indicates whether the applicant has gone 90 days past due (DPD) or worse in the past 12 months (i.e. defaulted) after getting a credit card.

In some cases, it is observed that all the variables in the credit bureau data are zero and credit card utilization is missing. These represent cases in which there is a no-hit in the credit bureau. The cases with missing credit card utilization are also observed. These are the cases in which the applicant does not have any other credit card.

Data Understanding and Handling Data Issues for Demographic Data

Variable Name	Description	Data Issues and their handling
Application Id	Unique Ids of the customer	3 non-unique Application Id values (6 records) were found. On manual check, it was found that the underlying customers with both records were different. All the 6 records, being less in number were removed considering them as junk data.
Age	Age of customer	• One customer had negative age (age was -3). • 19 customers had age as 0. • 45 customers had age less than 18 (assuming a person with age less than 18 is not eligible for applying the card). The age for all the above record was set as 18.
Gender	Gender of Customer	Gender was missing for 2 customers. These were imputed as ‘M’ based on maximum value counts (mode).
Marital Status	Marital status of customer (at the time of application)	Marital Status was missing for 6 customers. Missing values were imputed as ‘Married’ based on maximum value counts (mode).
Income	Income of customers	Income was negative for 81 customers. This negative count was imputed with median value.

Data Understanding and Handling Data Issues for Demographic Data (continued …)

Variable Name	Description	Data Issues and their handling
No. of Dependents	No. of children of customers	No. of children were missing for 3 customers. All these 3 customers had Marital status ‘Married’. So, the values were imputed with value 3 based on maximum occurring value.
Education	Education of customers	Education was missing for 119 customers. Missing values were imputed with value ‘UNKNOWN’.
Profession	Profession of customers	Profession was missing for 14 customers. Missing values were imputed with 'SAL’ based on maximum occurring values.
Type of Residence	Type of residence of customers	Residence type was missing for 8 customers. Missing values were imputed with 'Rented’ based on maximum occurring values.
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")	Performance Tag was missing for 1425 customers. All these customers were removed from the data set as Credit Card was never issued to these customers.

Data Understanding and Handling Data Issues for Credit Bureau Data

Variable Name	Description	Data Issues and their handling
Application Id	Customer application ID	3 non-unique Application Id values (6 records) were found. On manual check, it was found that the underlying customers with both records were different. All the 6 records, being less in number were removed considering them as junk data.
No of times 90 DPD or worse in last 6 months	Number of times customer has not payed dues since 90 days 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 in 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 in 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 in 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 in last 12 months	

Data Understanding and Handling Data Issues for Credit Bureau Data (continued …)

Variable Name	Description	Data Issues and their handling
No of times 30 DPD or worse in last 12 months	Number of times customer has not payed dues since 30 days in last 12 months	
Avgas CC Utilization in last 12 months	Average utilization of credit card by customer	This was missing for 1058 customers. This missing values were imputed by median value except for the values where Outstanding Balance was missing (in that scenario it was imputed with 0).
No of trades opened in last 6 months	Number of times the customer has done the trades in last 6 months	This was missing for 1 customer. Missing value was imputed with the maximum occurring value.
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	
No of Inquiries in last 6 months (excluding home & auto loans)	Number of times the customers has inquired in last 6 months	

Data Understanding and Manipulation for Credit Bureau Data (continued …)

Variable Name	Description	Data Issues and their handling
No of Inquiries in last 12 months (excluding home & auto loans)	Number of times the customers has inquired in last 12 months	
Presence of open home loan	If the customer has home loan (1 represents "Yes")	This was missing for 272 customers. This was assumed that these haven't taken home loan so imputed with 0.
Outstanding Balance	Outstanding balance of customer	This was missing for 272 customers. This was assumed that these haven't taken/applied any credit card yet so imputed with 0.
Total No of Trades	Number of times the customer has done total trades	
Presence of open auto loan	If the customer has auto loan (1 represents "Yes")	
Performance Tag	Status of customer performance (" 1 represents "Default")	Performance Tag was missing for 1425 customers. All these customers were removed from the data set as Credit Card was never issued to them.

EDA and WoE / IV Analysis for Demographic Data

```
iv_demographics.sort_values(by='IV', ascending = False)
```

	Variable	IV
0	Months_Current_Residence	0.073546
0	Income	0.039054
0	Months_Current_Company	0.021682
0	Age	0.004896
0	Dependents_No	0.002818
0	Profession	0.002221
0	Residence_Type	0.000942
0	Education	0.000782
0	Gender	0.000568
0	Marital_Status	0.000147

Following predictor variables were obtained as part of EDA and WoE Analysis for Demographic data

(top five based on IV values).

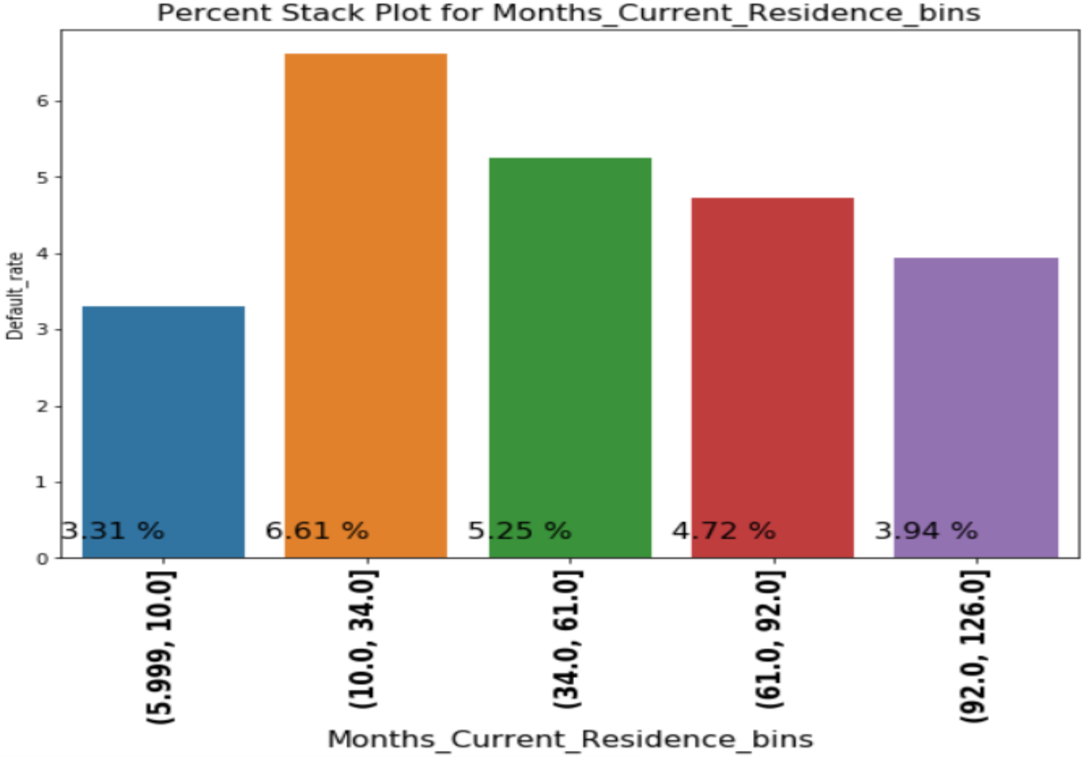
- Months_Current_Residence
- Income
- Months_Current_Company
- Age
- Dependents_No

The overall information value of the data set was **0.1467**.

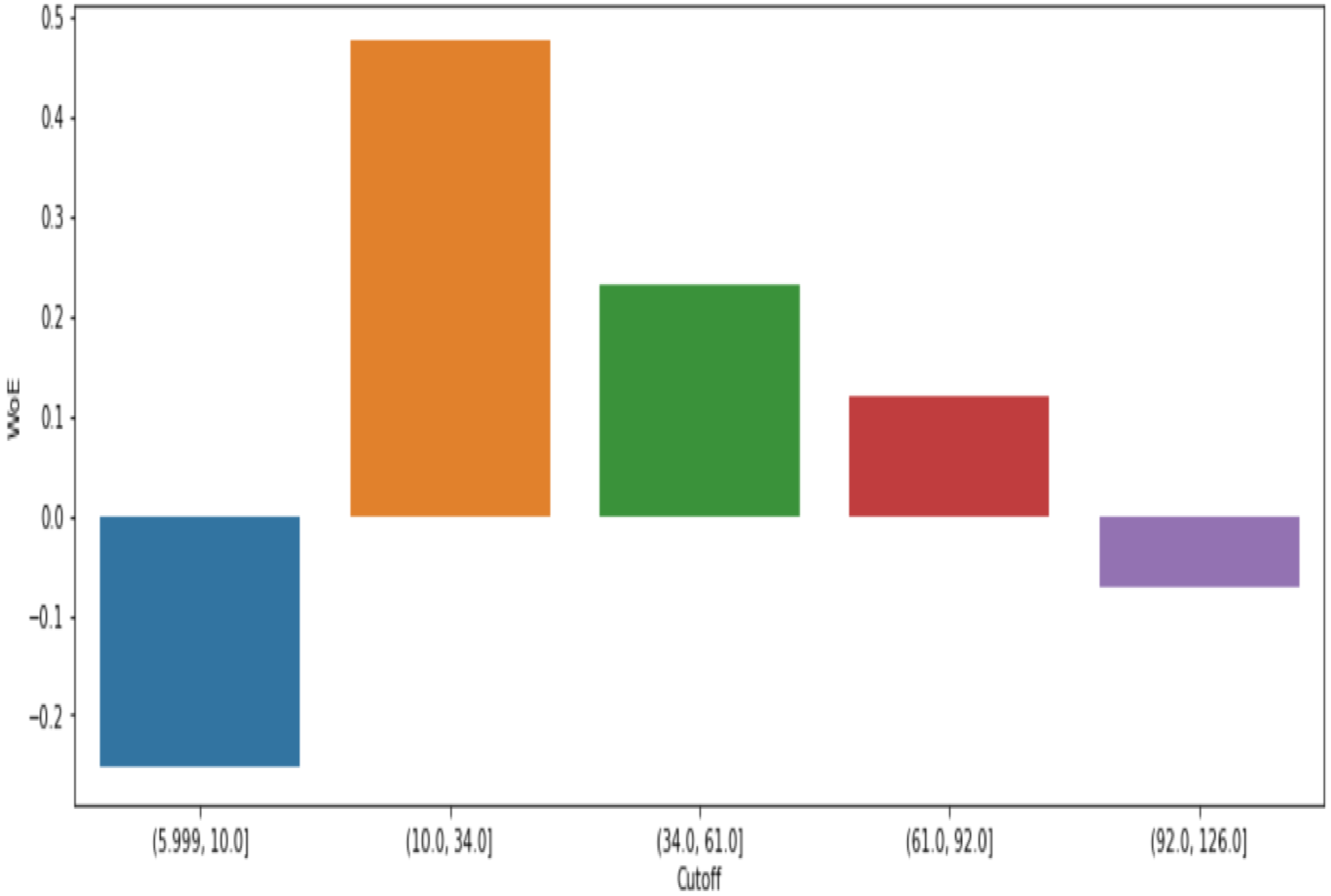
Understanding Months_Current_Residence as predictor demographic variable

The WoE values across rising in bins show monotonic decrease in WoE as months of current residence increase across bins (except the lowest value bin).

Similar trend of monotonic decrease is also observed in the bar plot for the bins created for Number of months in current residence.



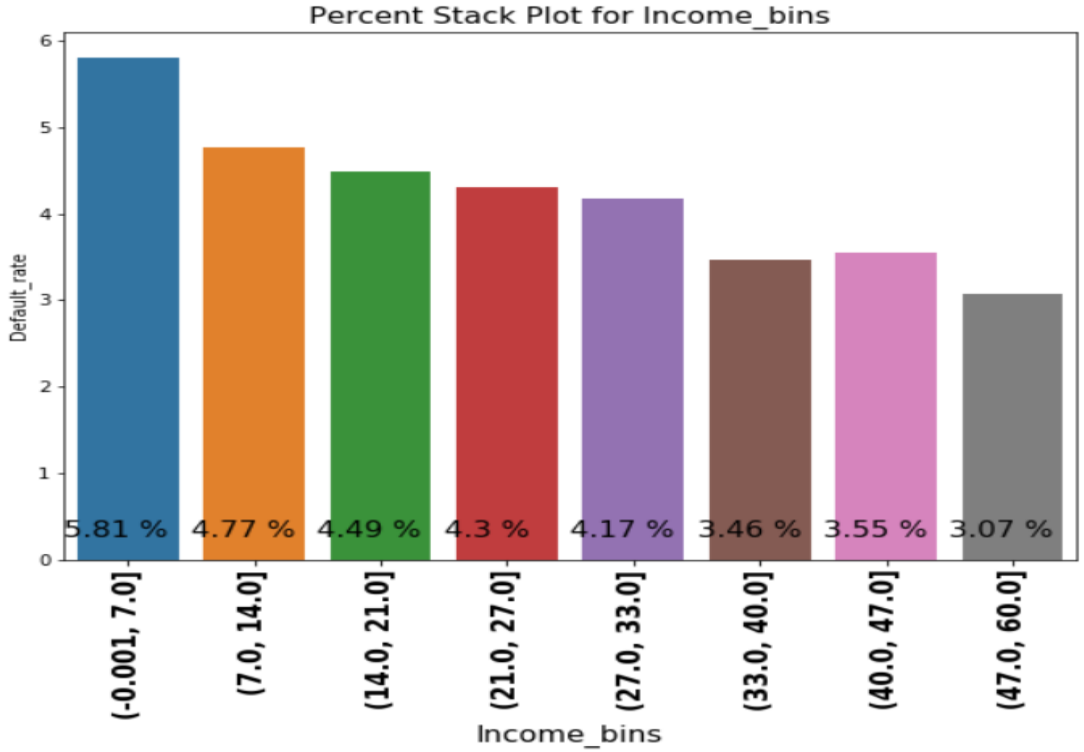

```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_demographics[woe_demographics.Variable=='Months_Current_Residence'])
plt.show()
```



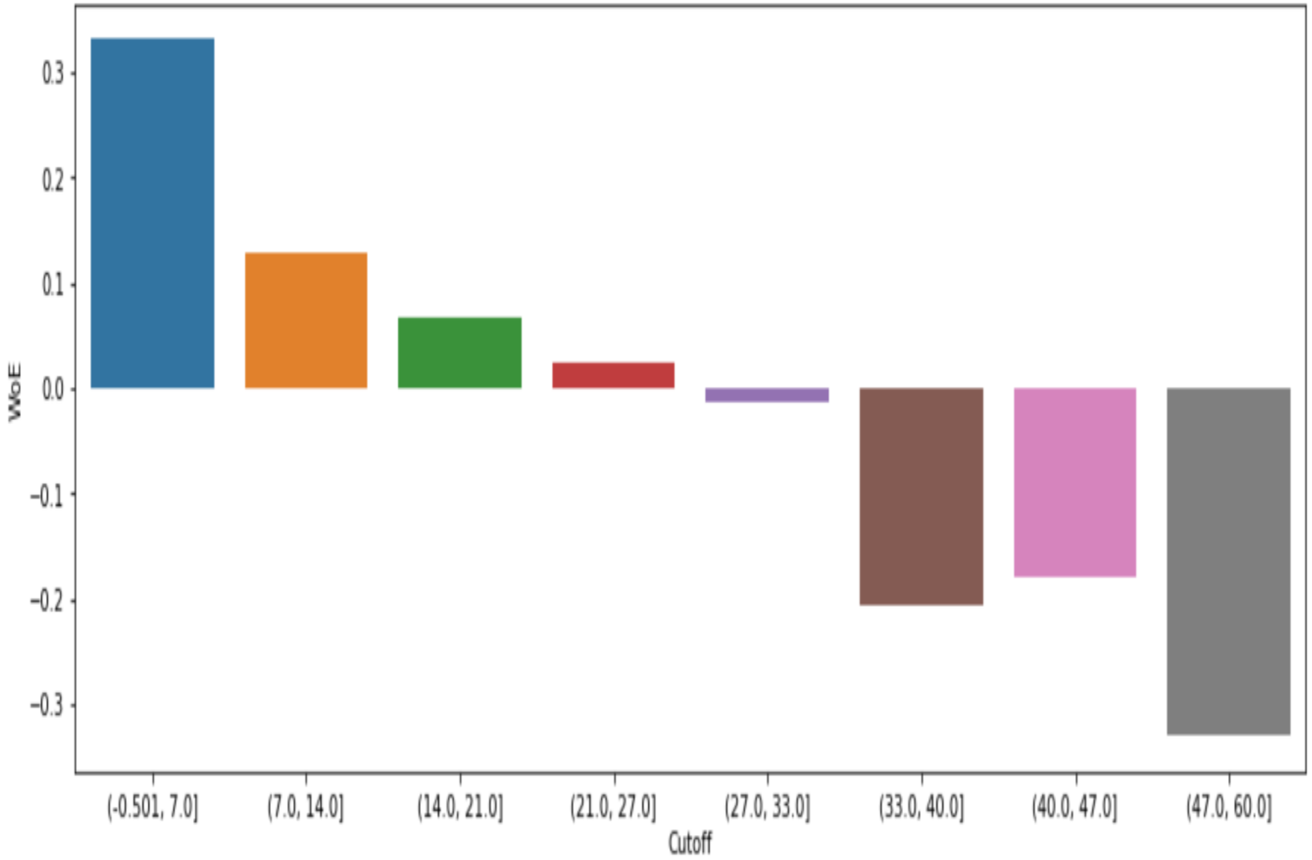
Understanding Income as predictor demographic variable

The WoE and bins plot chart shows monotonic decrease in default rate as income increase across bins.

Similar trend is observed across Income bins plotted along the bar plot which shows decrease in percentage of defaulters across income bins.




```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_demographics[woe_demographics.Variable=='Income'])
plt.show()
```

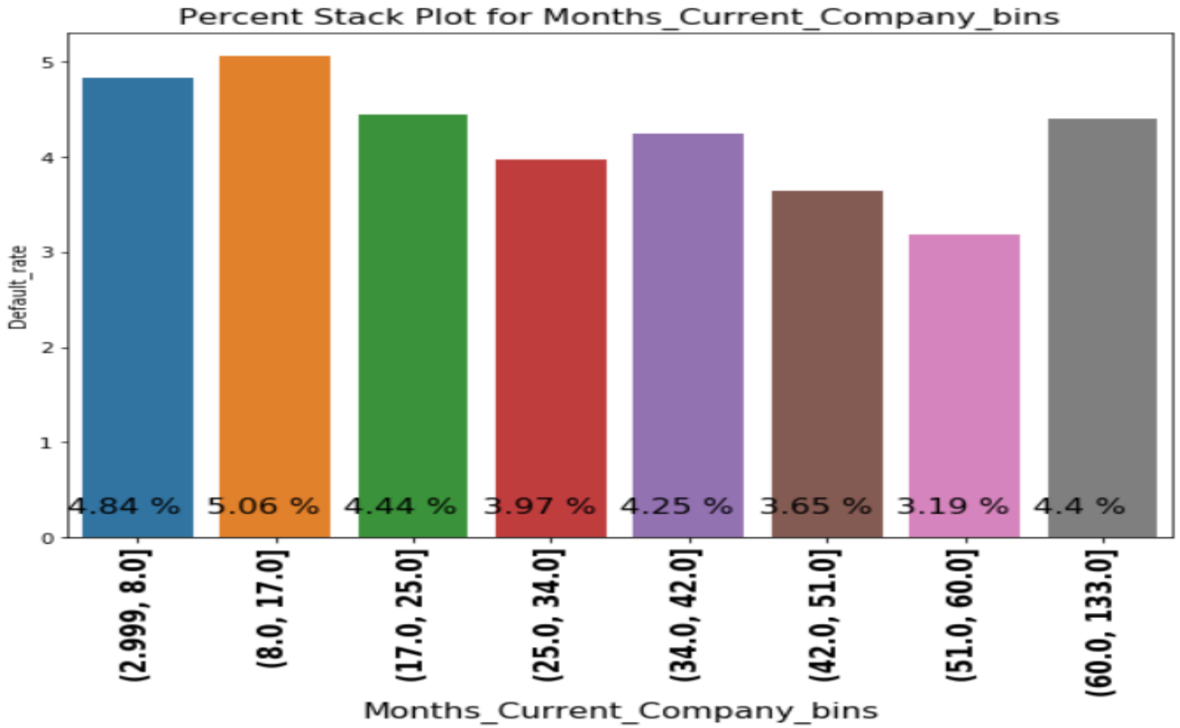


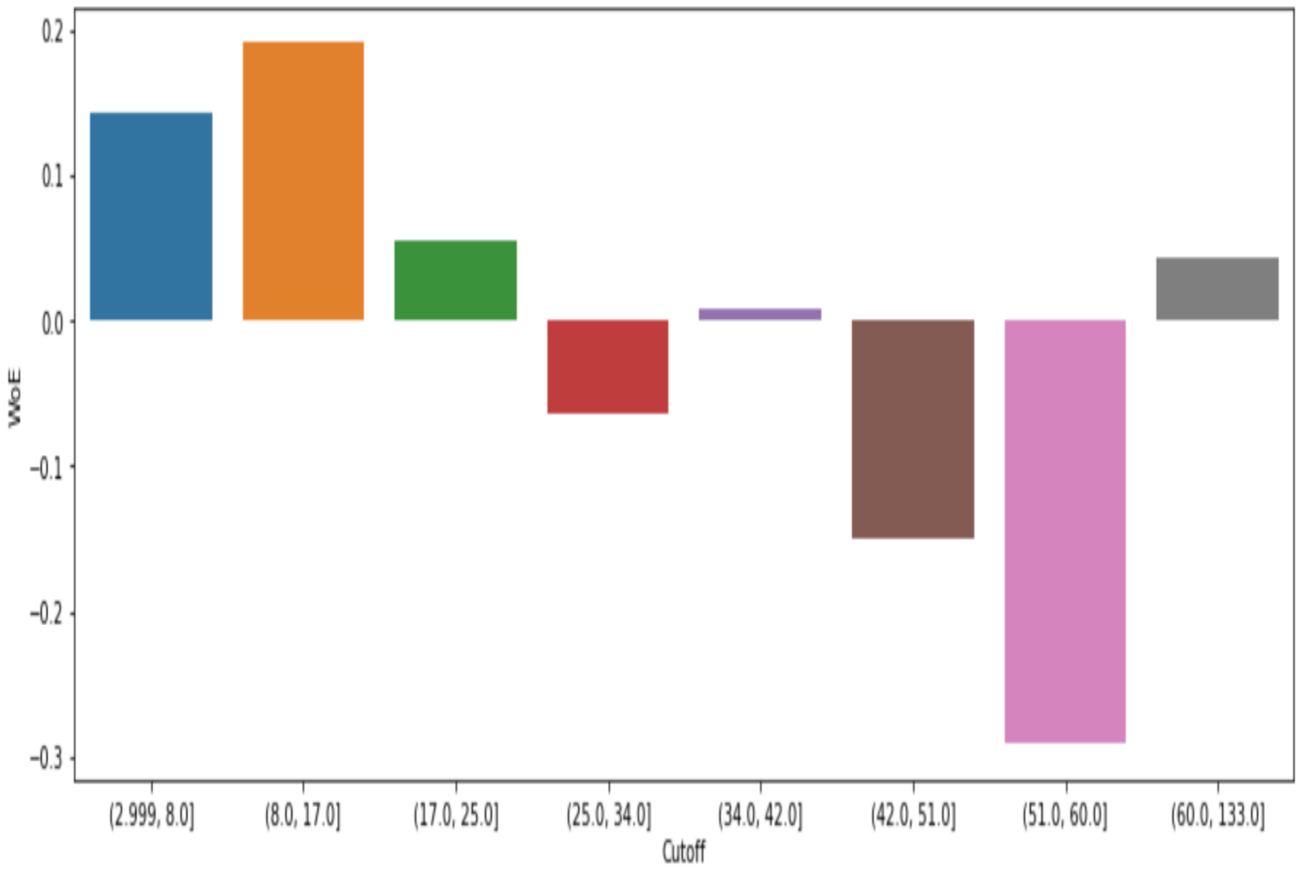
Understanding Months_Current_Company as predictor demographic variable

The WoE and bins plot chart shows monotonic decrease in default rate as Months_Current_Company increase across bins (with some exceptions).

Similar trend is observed across Months_Current_Company bins plotted along the bar plot which shows decrease in percentage of defaulters across bins.

```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_demographics[woe_demographics.Variable=='Months_Current_Company'])
plt.show()
```

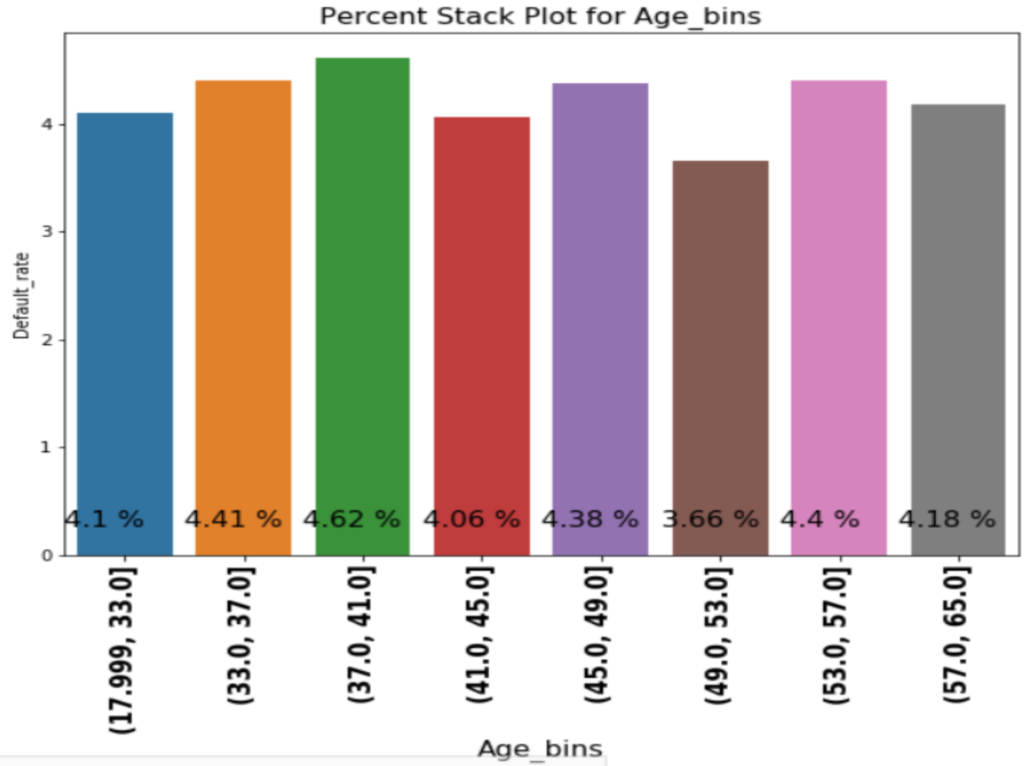




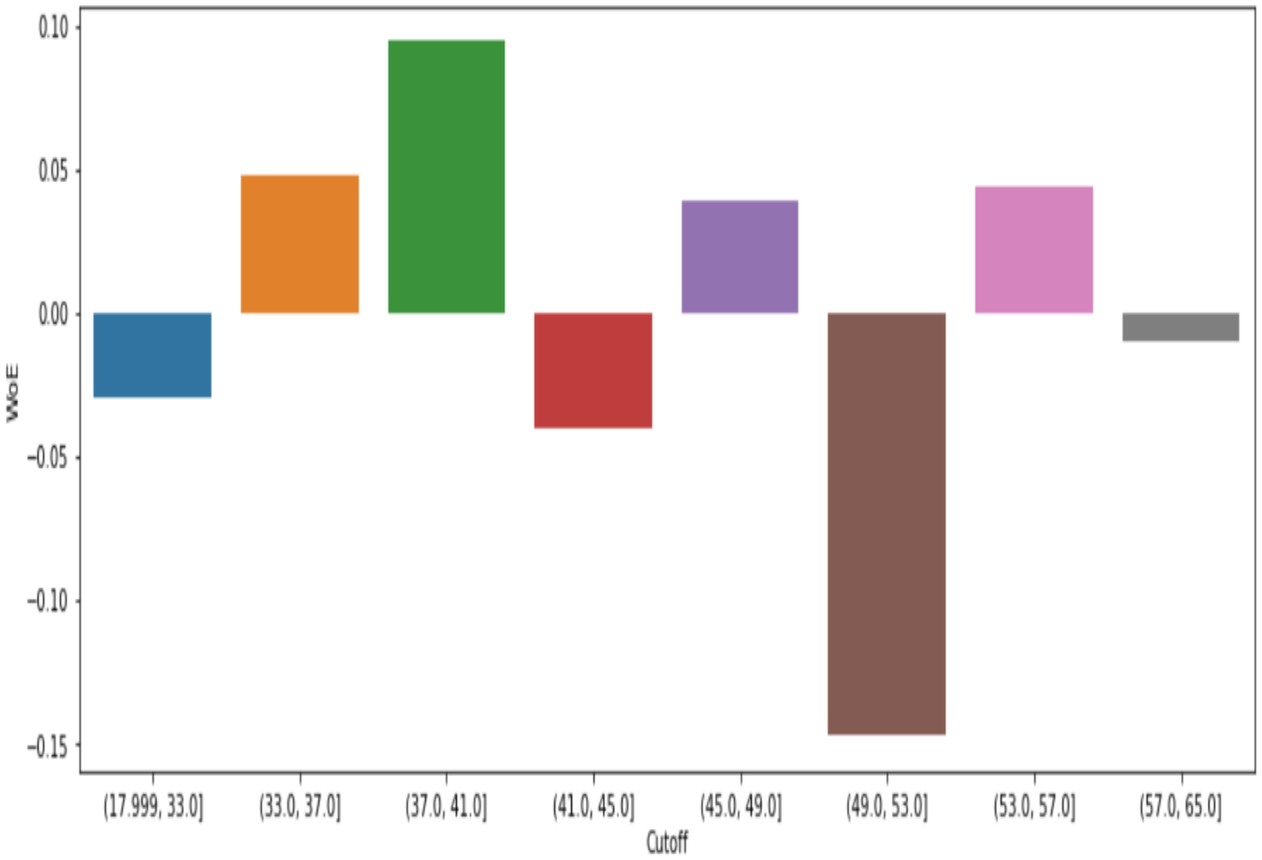
Understanding Age as predictor demographic variable

The WoE and bins plot chart shows variations across Age group bins.

Similar kind of variations are also observed in bar plot for Age bins created.




```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_demographics[woe_demographics.Variable=='Age'])
plt.show()
```

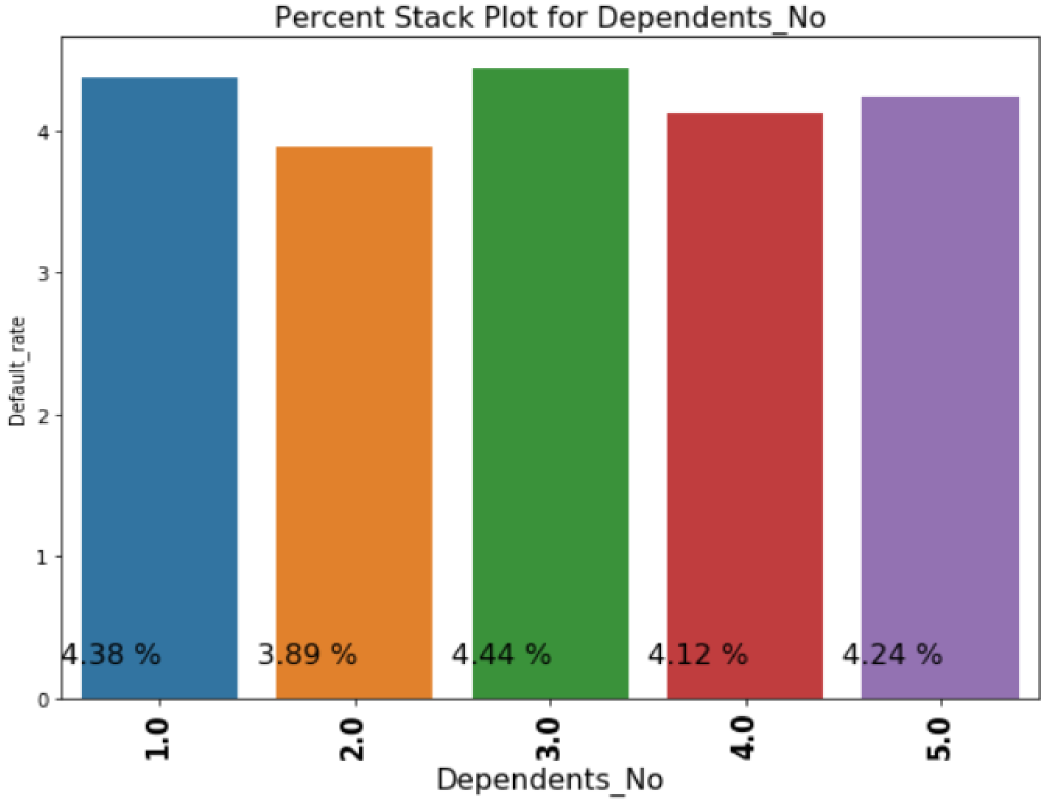


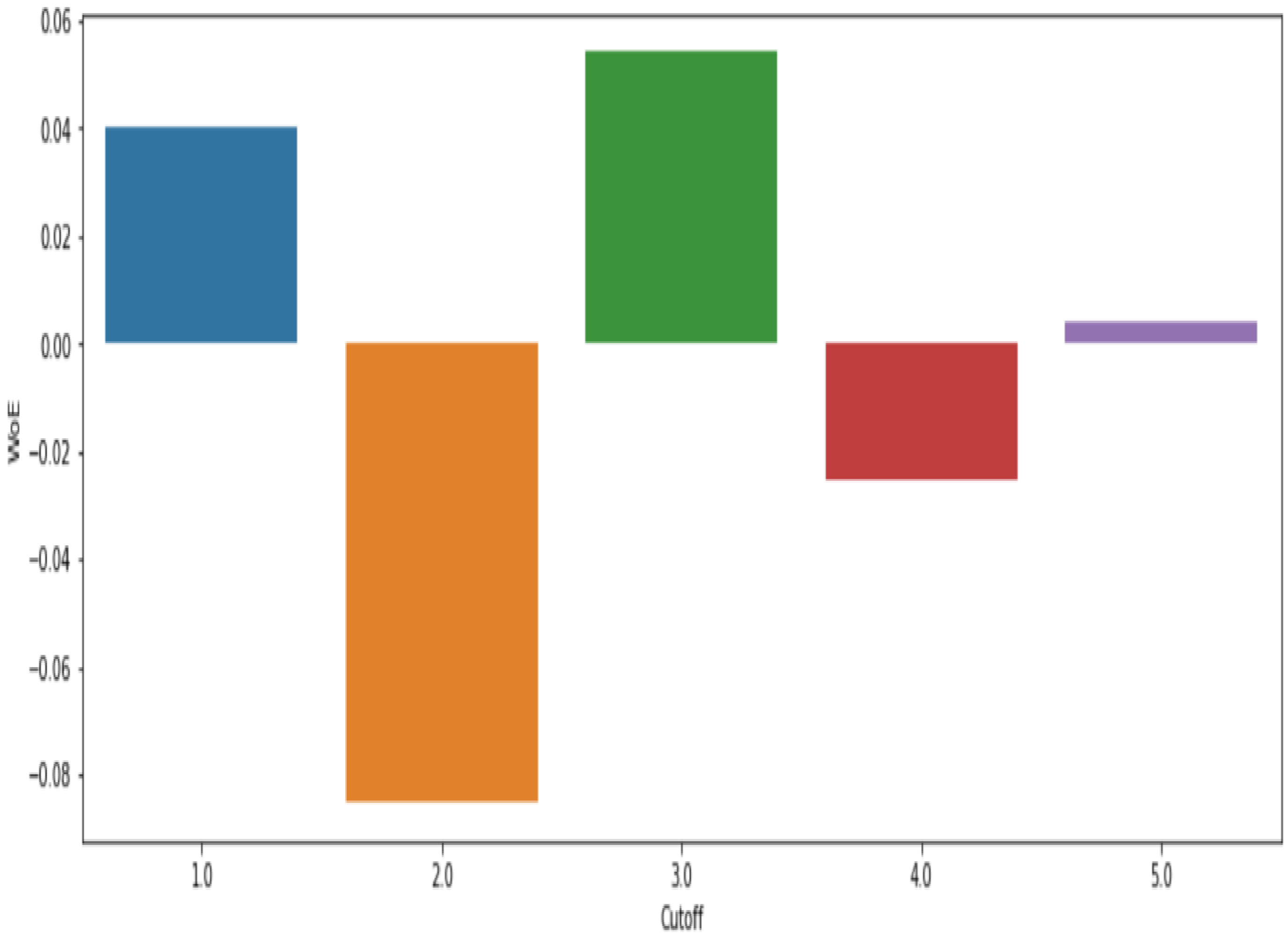
Understanding Dependent No. as predictor demographic variable

The Dependent No. is the fifth significant variable as per IV analysis and there is a big fall in the Information value as comparison to fourth significant variable.

Although this variable is significant as per IV value, the bar plot is unable to convey much to be classified as trend.

```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_demographics1[woe_demographics1.Variable=='Dependents_No'])
plt.show()
```





EDA and WoE / IV Analysis for Credit Bureau Data

Following predictor variables were obtained as part of EDA and WoE Analysis for Credit Bureau data

(top five based on IV values).

- Avgas_CC_Utilization_12_months
- Trades_opened_last_12_months
- PL_Trades_opened_last_12_months
- Outstanding_Balance
- Inquiries_last_12_months

The overall information value of the data set was **3.287**.

```
iv_creditbureau.sort_values(by='IV', ascending = False)
```

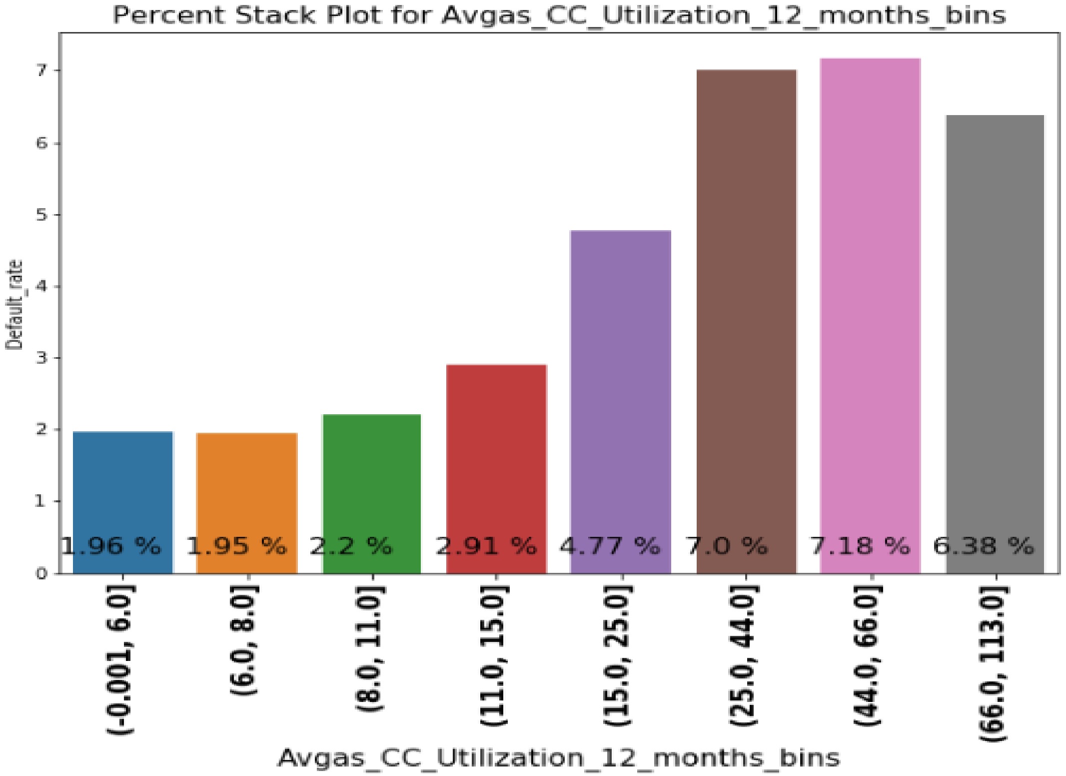
	Variable	IV
0	Avgas_CC_Utilization_12_months	0.308660
0	Trades_opened_last_12_months	0.291635
0	PL_Trades_opened_last_12_months	0.255968
0	Outstanding_Balance	0.253333
0	No_30_DPD_last_6_months	0.244250
0	Total_Trades	0.238455
0	PL_Trades_opened_last_6_months	0.224219
0	No_30_DPD_last_12_months	0.218609
0	No_90_DPD_last_12_months	0.215653
0	No_60_DPD_last_6_months	0.211274
0	No_60_DPD_last_12_months	0.188230
0	Trades_opened_last_6_months	0.186148
0	Inquiries_last_12_months	0.172713
0	No_90_DPD_last_6_months	0.162659
0	Inquiries_last_6_months	0.113141
0	Presence_open_auto_loan	0.001655
0	Presence_open_home_loan	0.000462

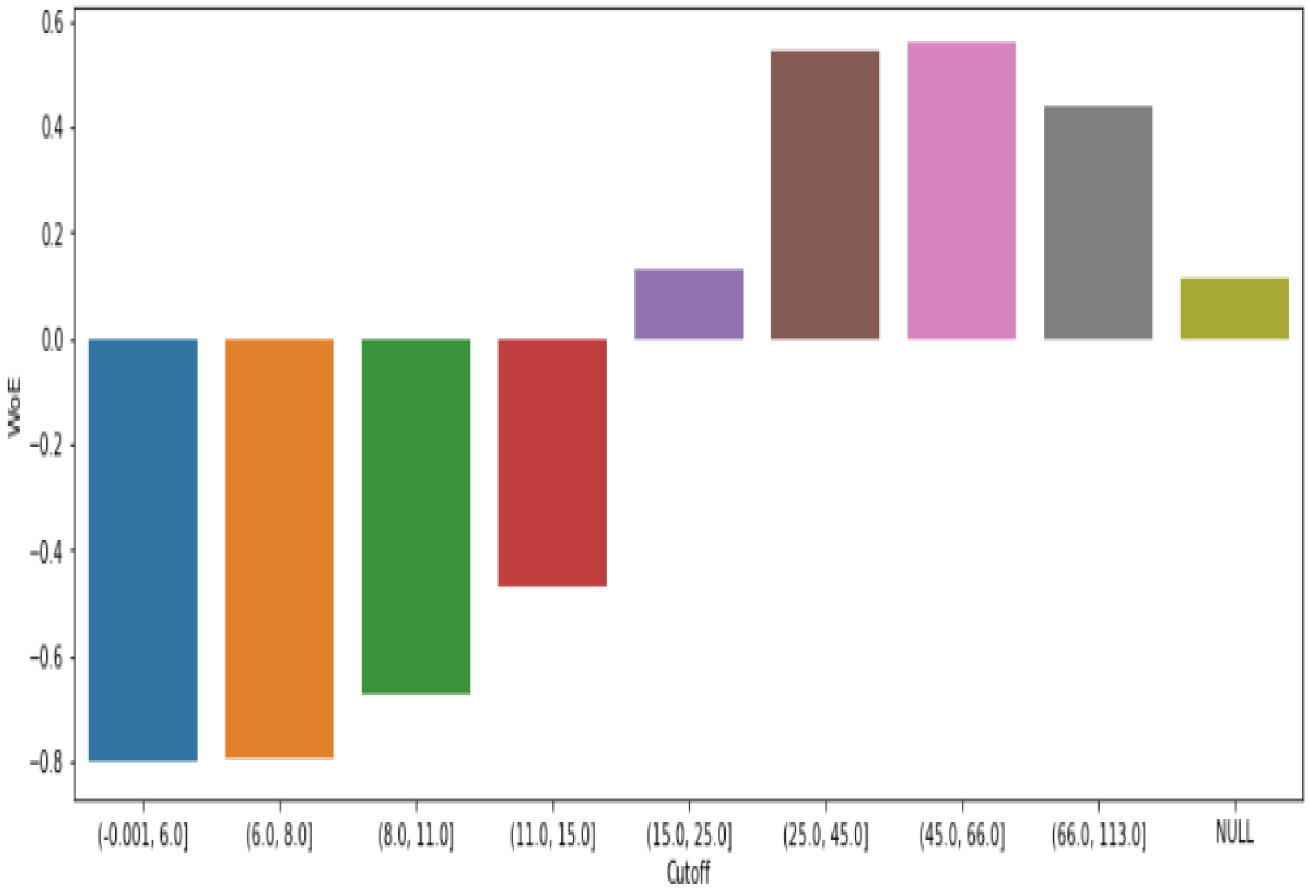
Understanding Avgas_CC_Utilization_12_months as predictor Credit Bureau variable

The WoE values across rising in bins show monotonic increase in WoE as Avgas_CC_Utilization_12_months increase across bins (except highest values bin).

Similar trend of monotonic increase in defaulter's percent is also observed in the bar plot for the bins created for Avgas_CC_Utilization_12_months.

```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_creditbureau[woe_creditbureau.Variable=='Avgas_CC_Utilization_12_months'])
plt.show()
```



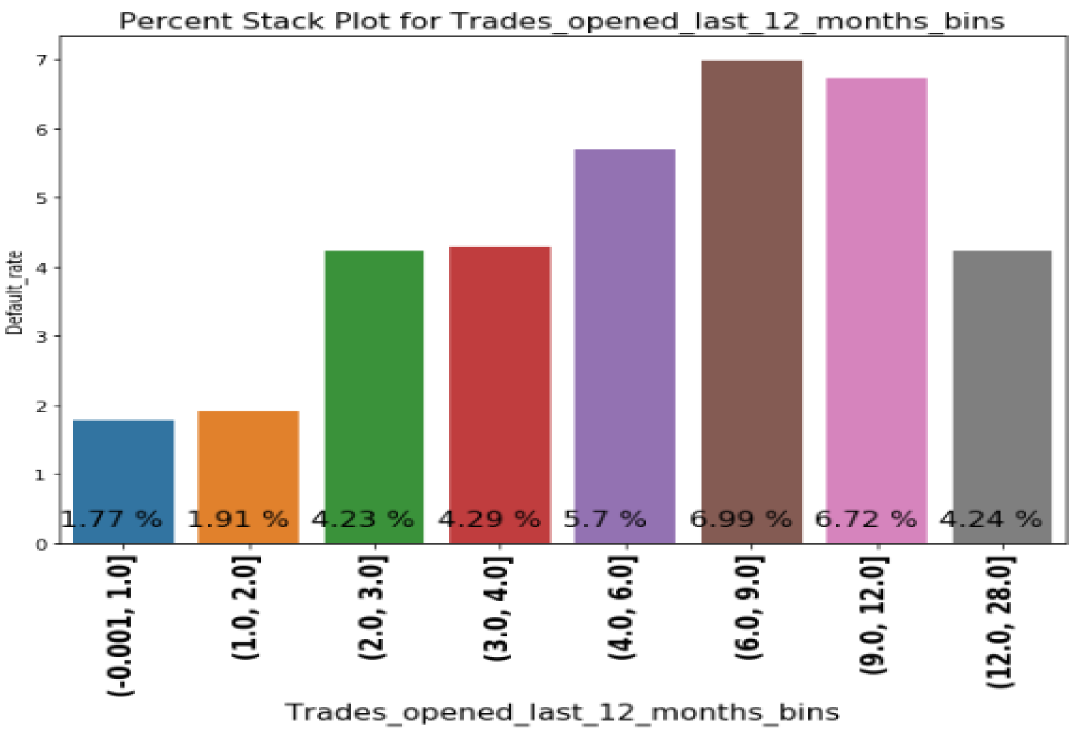


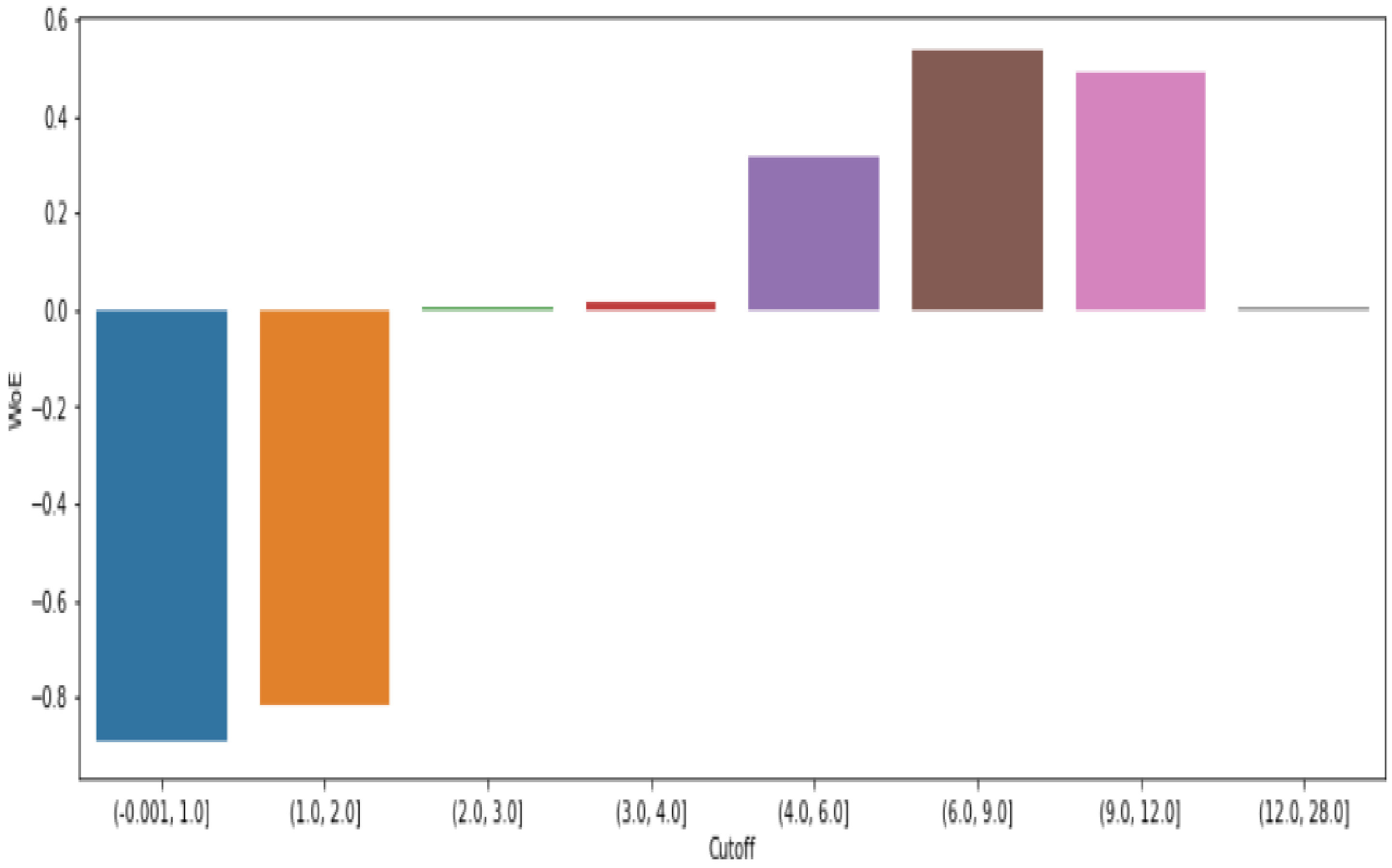
Understanding Trades_opened_last_12_months as predictor Credit Bureau variable

The WoE values across rising in bins show monotonic increase in WoE as Trades_opened_last_12_months increase across bins (except the highest value bins).

Similar trend of monotonic increase in default percent is also visible in the bar plot for the bins created for Trades_opened_last_12_months.

```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_creditbureau[woe_creditbureau.Variable=='Trades_opened_last_12_months'])
plt.show()
```



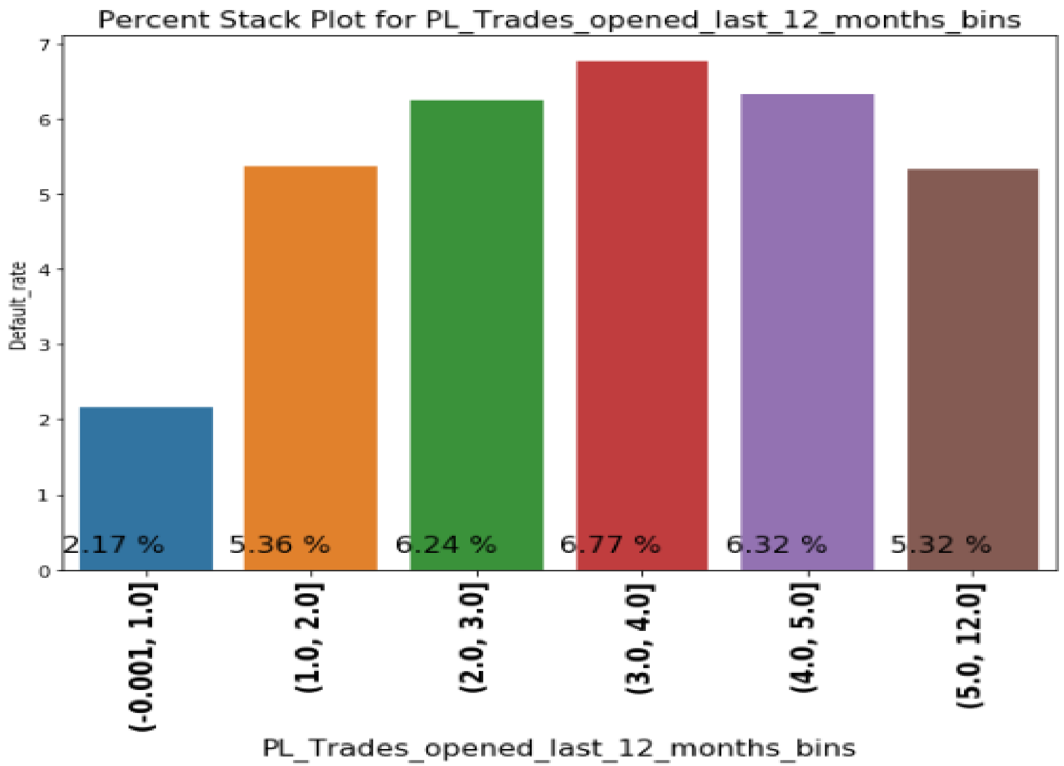


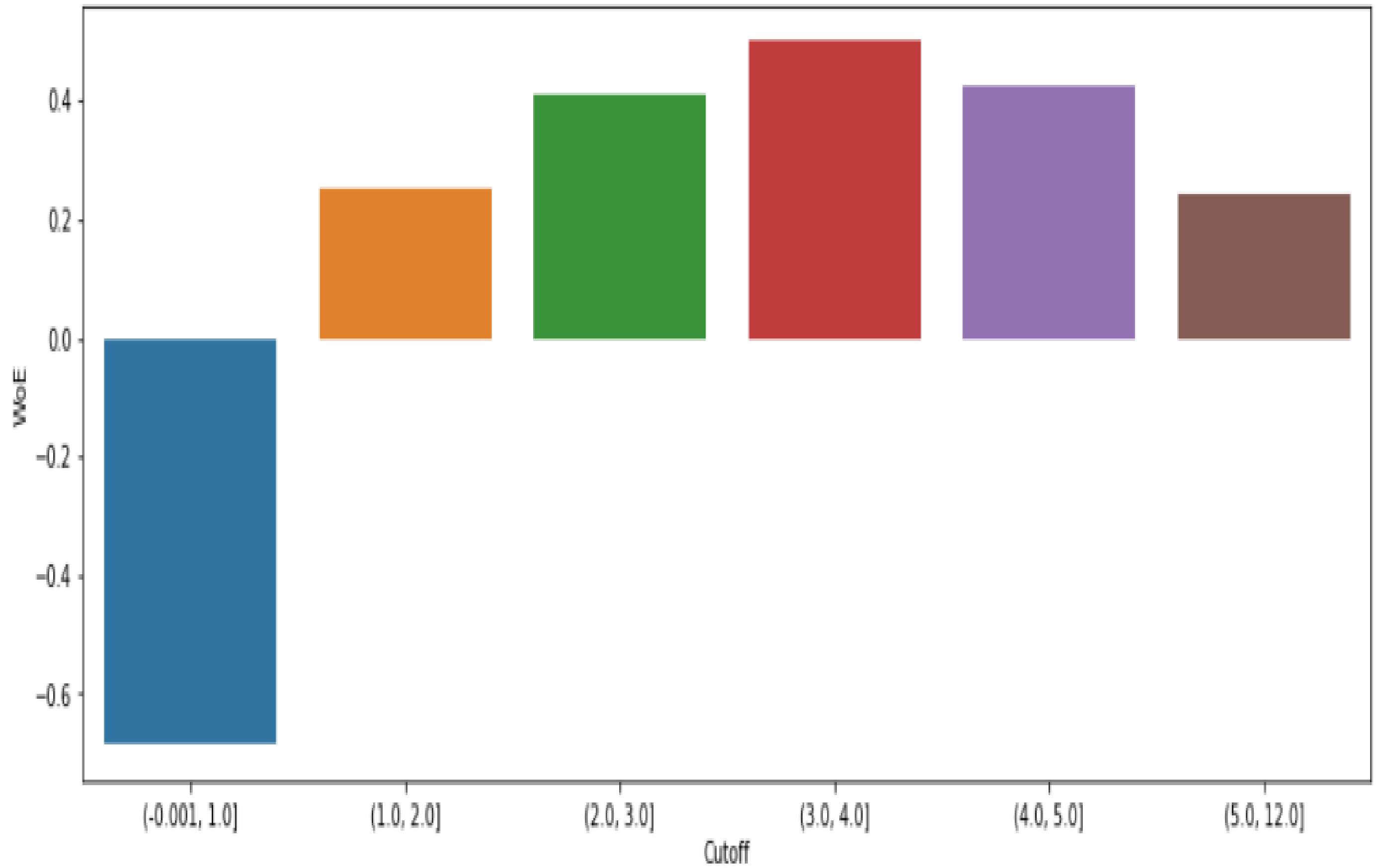
Understanding PL_Trades_opened_last_12_months as predictor Credit Bureau variable

The WoE values across rising in bins show monotonic increase and then monotonous decrease as PL_Trades_opened_last_12_months increase across bins.

Similar trend of monotonic increase and decrease in default percent is also observed in the bar plot for the bins created for PL_Trades_opened_last_12_months.

```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_creditbureau[woe_creditbureau.Variable=='PL_Trades_opened_last_12_months'])
plt.show()
```

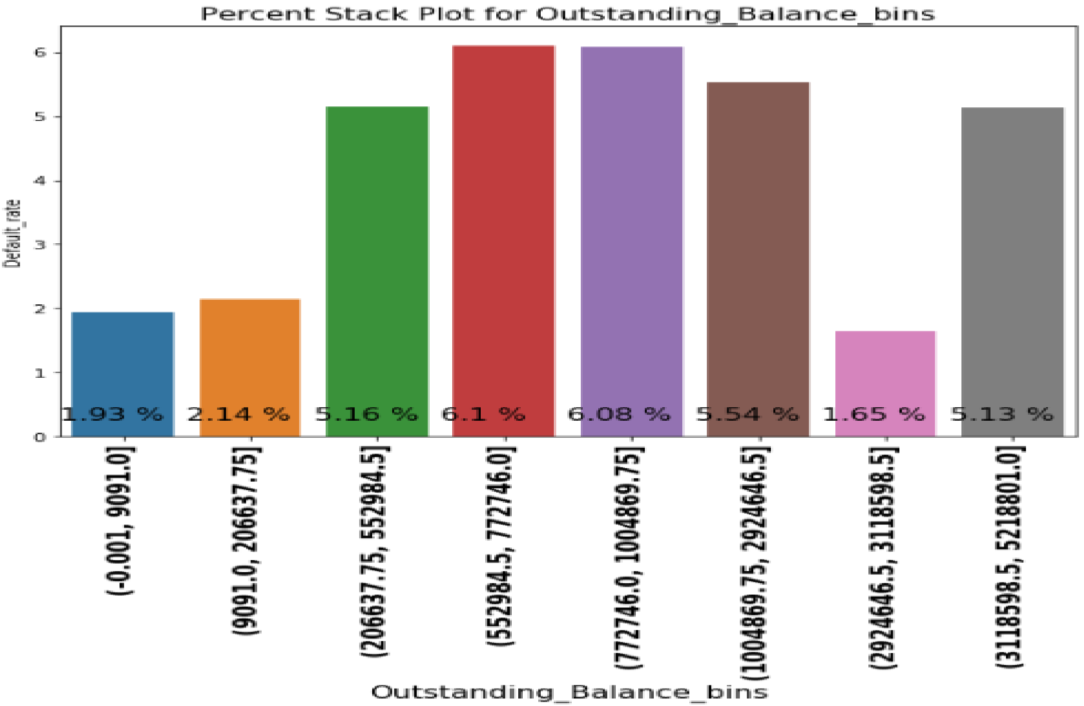


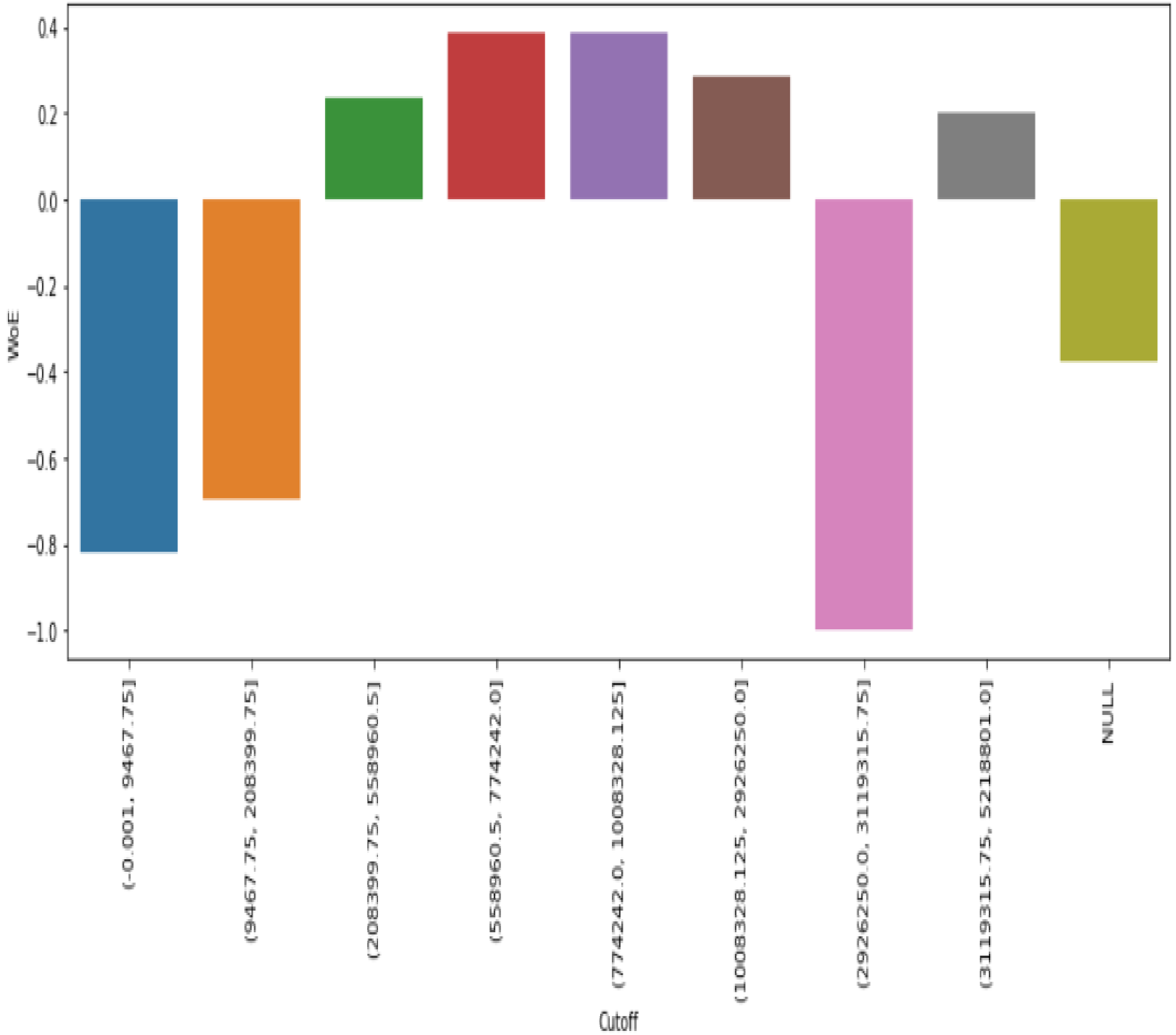


Understanding Outstanding_Balance as predictor Credit Bureau variable

The WoE values across rising in bins show monotonic increase in WoE and then monotonous decrease as Outstanding_Balance increase across bins.

Similar trend of monotonic increase and decrease in default percent is also observed in the bar plot for the bins created for Outstanding_Balance.

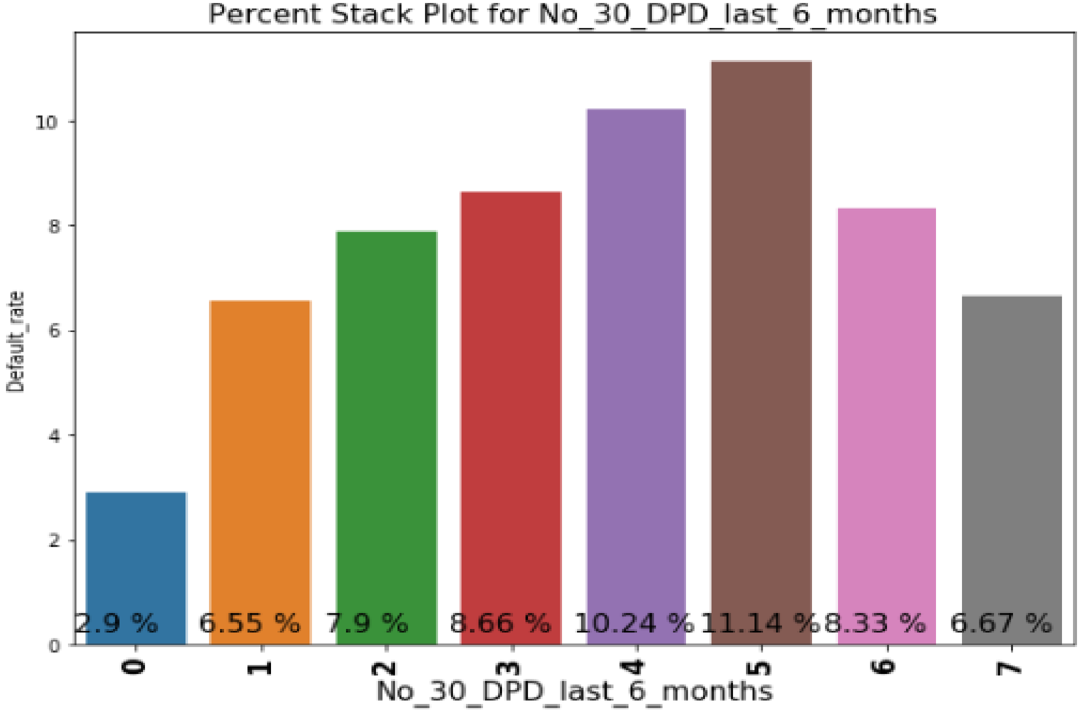




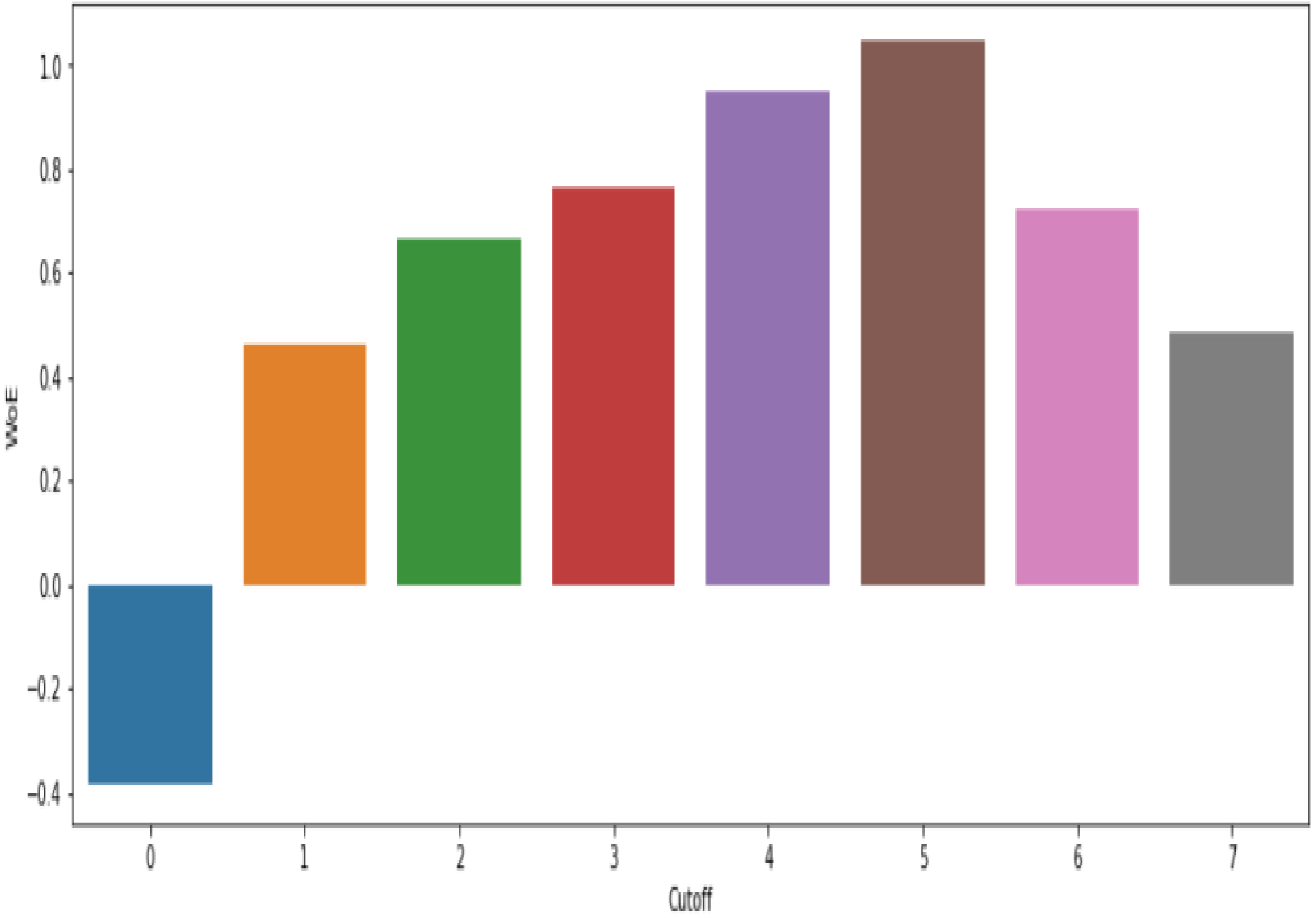
Understanding No_30_DPD_last_6_months as predictor Credit Bureau variable

The WoE values across rising in bins show monotonic increase and then monotonous decrease as No_30_DPD_last_6_months increase across bins.

Similar trend of monotonic increase and decrease in default percent is also observed in the bar plot for the bins created for No_30_DPD_last_6_months.




```
plt.figure(figsize=(15,5))
sns.barplot(y='WoE', x="Cutoff", data=woe_creditbureau[woe_creditbureau.Variable=='No_30_DPD_last_6_months'])
plt.show()
```



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

- **4 models for each data set**
  - Logistic regression with RFE
  - Logistic regression with regularization
  - Decision tree
  - Random forest

- **Total 16 models**

Model Building Results for Demographics Dataset

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%
Logistic Regression with Regularization	54.16%	5.1%	56.06%	100%	66.31%
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
- Logistic Regression with Regularization
  - C : 0.1
  - Penalty : L1
- 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%
Logistic Regression with Regularization	59.89%	5.85%	57.03%	100%	77.26%
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
- Logistic Regression with Regularization
  - C : 0.1
  - Penalty : L1
- 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%
Logistic Regression with Regularization	58%	6.45%	66.70%	100%	99.36%
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
- Logistic Regression with Regularization
  - C : 0.01
  - Penalty : L1
- 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%
Logistic Regression with Regularization	58.43%	6.78%	68.75%	100%	97.54%
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
- Logistic Regression with Regularization
  - C : 0.01
  - Penalty : L1
- Random Forest
  - max_depth : 4
  - min_samples_leaf : 350
  - min_samples_split : 400
  - n_estimators : 900
  - max_features : 10

Model Evaluation Techniques

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.
- Choice of Regularization (L1 or L2) for Regularization 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 :**
  - The model gave good recall values on Test data (which is our objective).
  - The model was able to reject almost all the manually rejected applications (work like humans).
  - It is an ensemble model which means it aggregates various diverse models adding almost all available information across the data.
  - The mode of operation of this model is very complex. Hence, guessing the prediction logic to forge data will not be possible both by staff members or deliberate defaulters.
  - The model is very stable. 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 is expected to have comparatively long life over others and is expected to have less modifications with time.
  - It requires no outlier treatment.
  - 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 / $\log_{10}(2)$

  offset = target_score - factor × $\log_{10}$(target_odds)

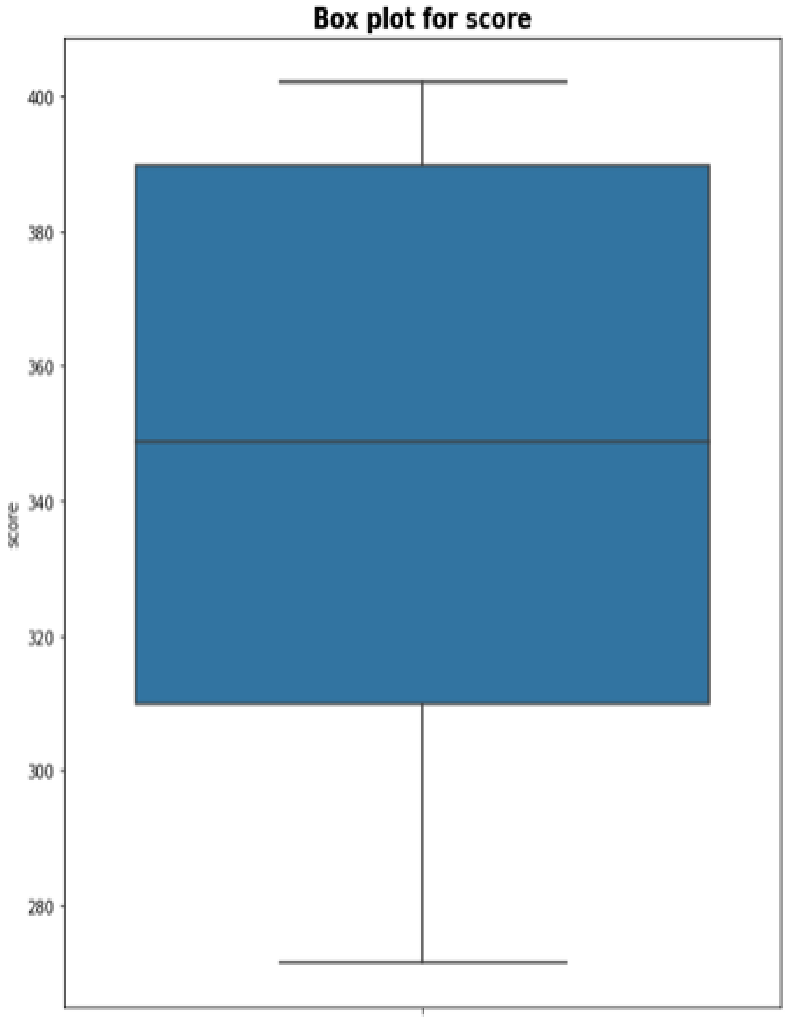
  scorecard['logit’] = $\sum (\beta \times \mathrm{WoE}) + \alpha$

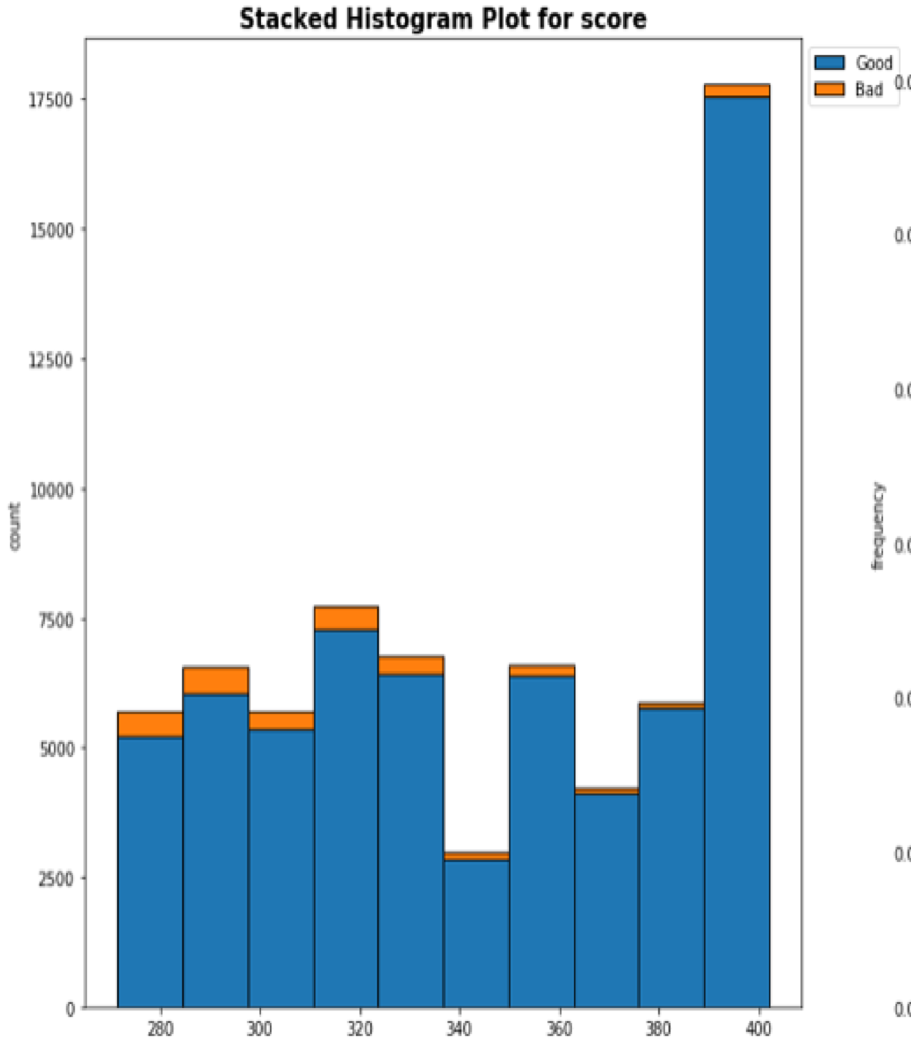
  (where β — logistic regression coefficient and α — logistic regression intercept)

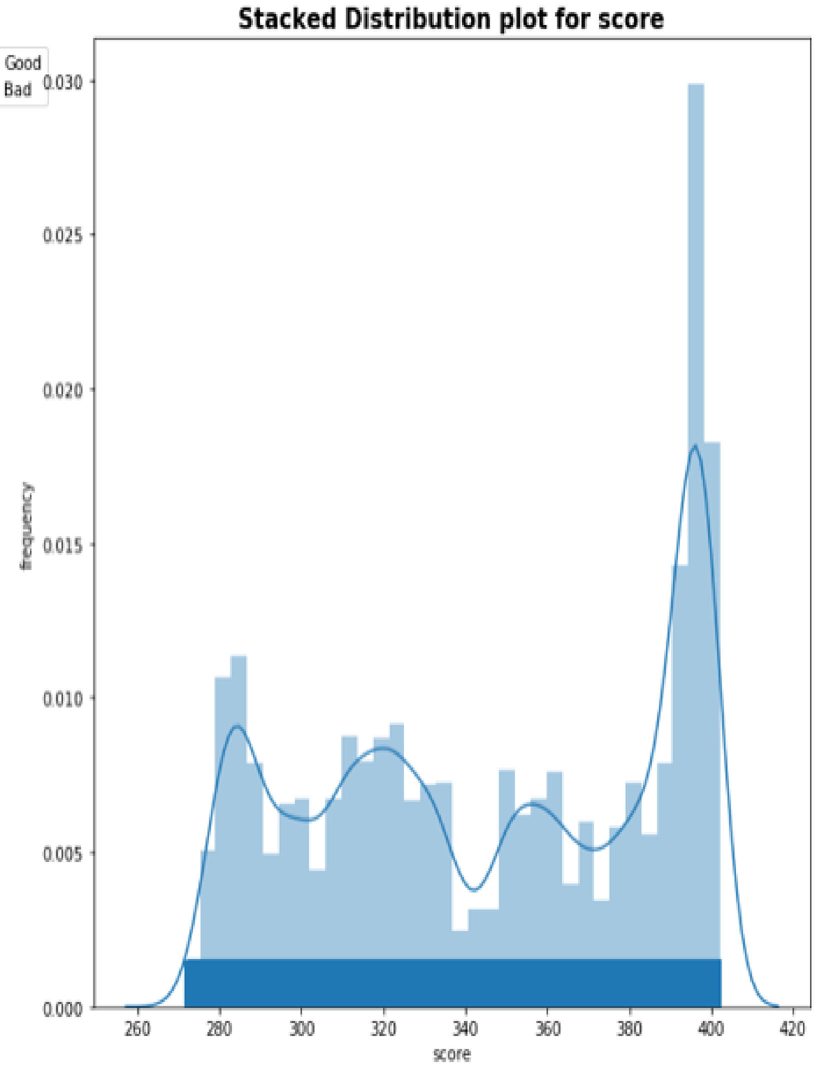
  Finally, **scorecard['score'] = offset - factor × scorecard['logit']**

Application Scorecard Variation Plots

Overall Scorecard Variation Plots

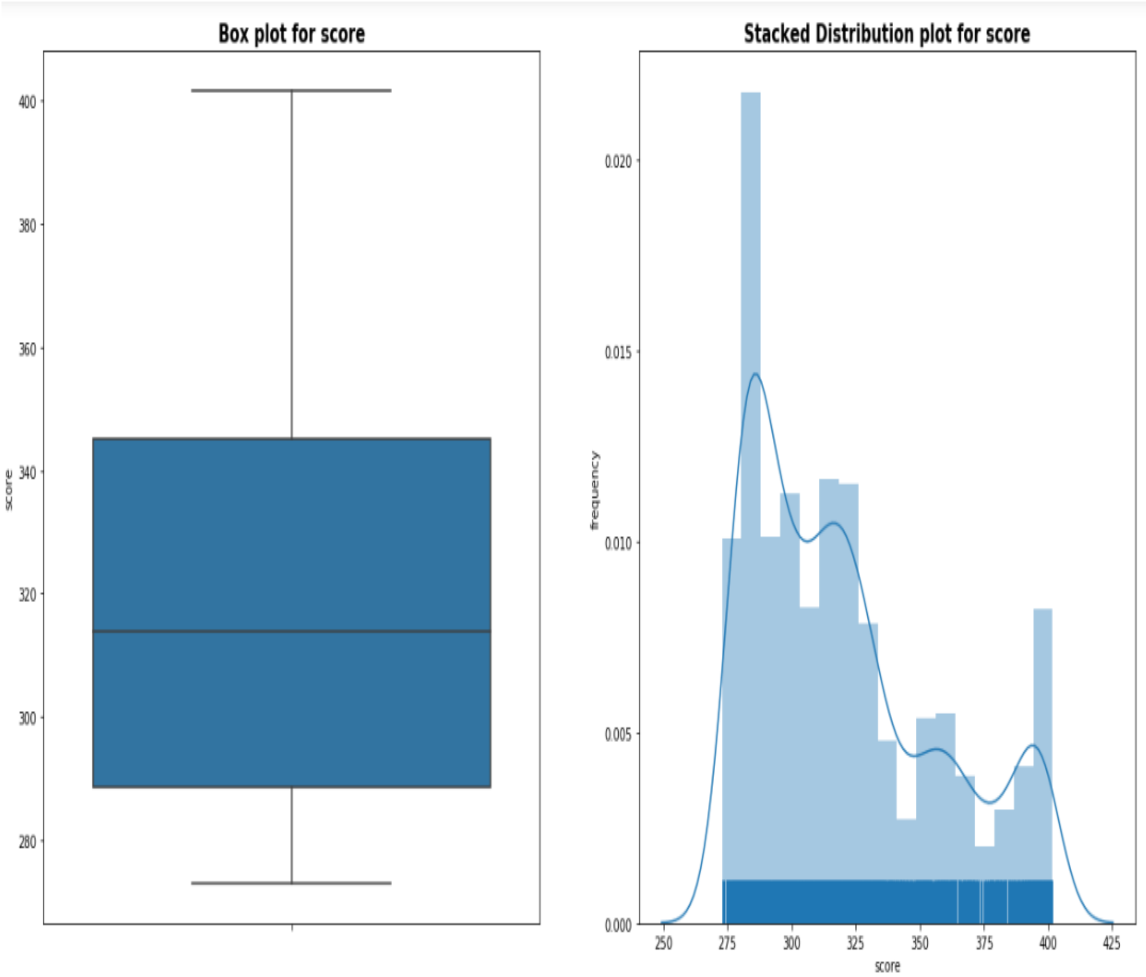




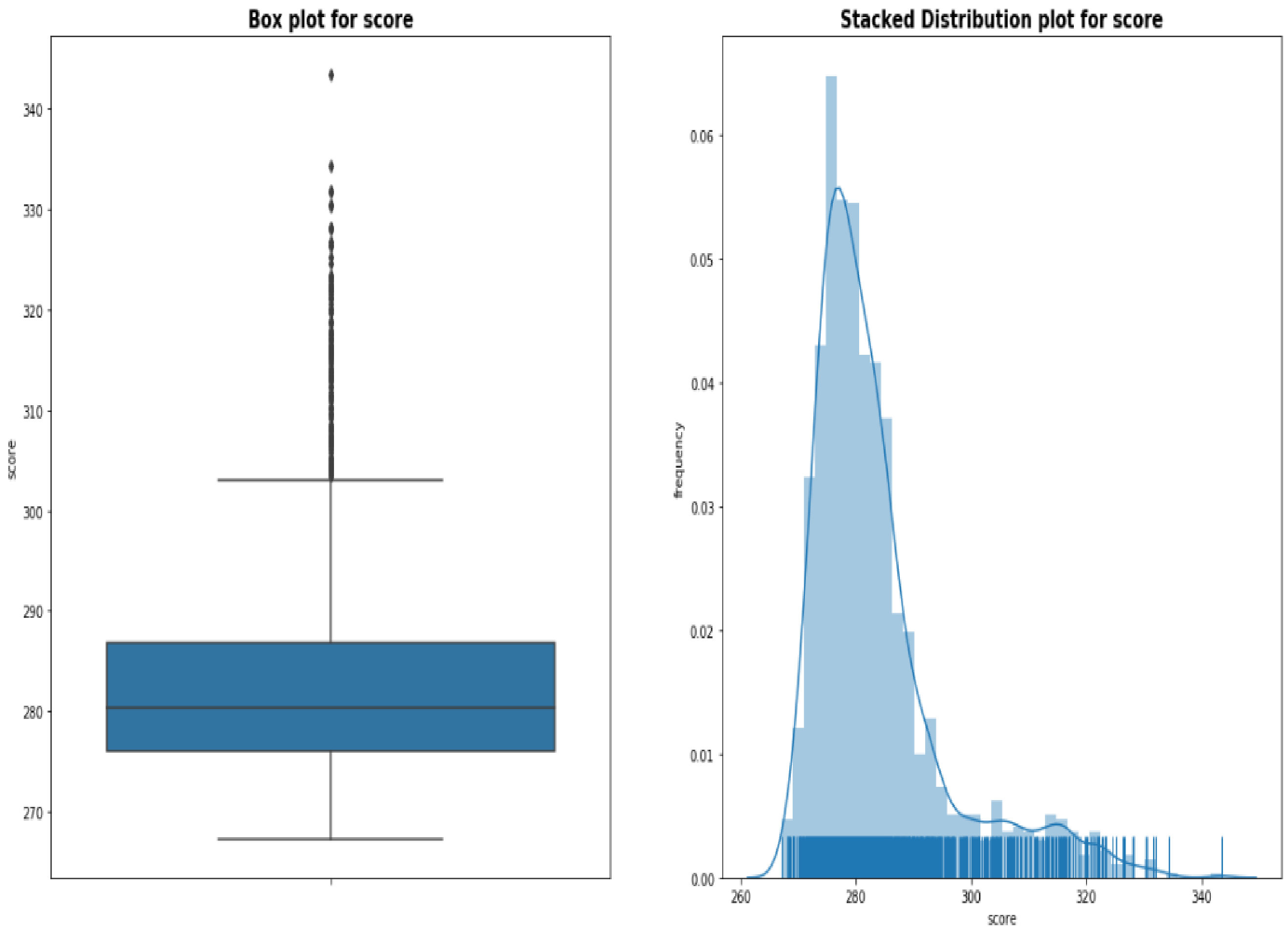


Application Scorecard Variation Plots continued …

Defaulters Scorecard Plots



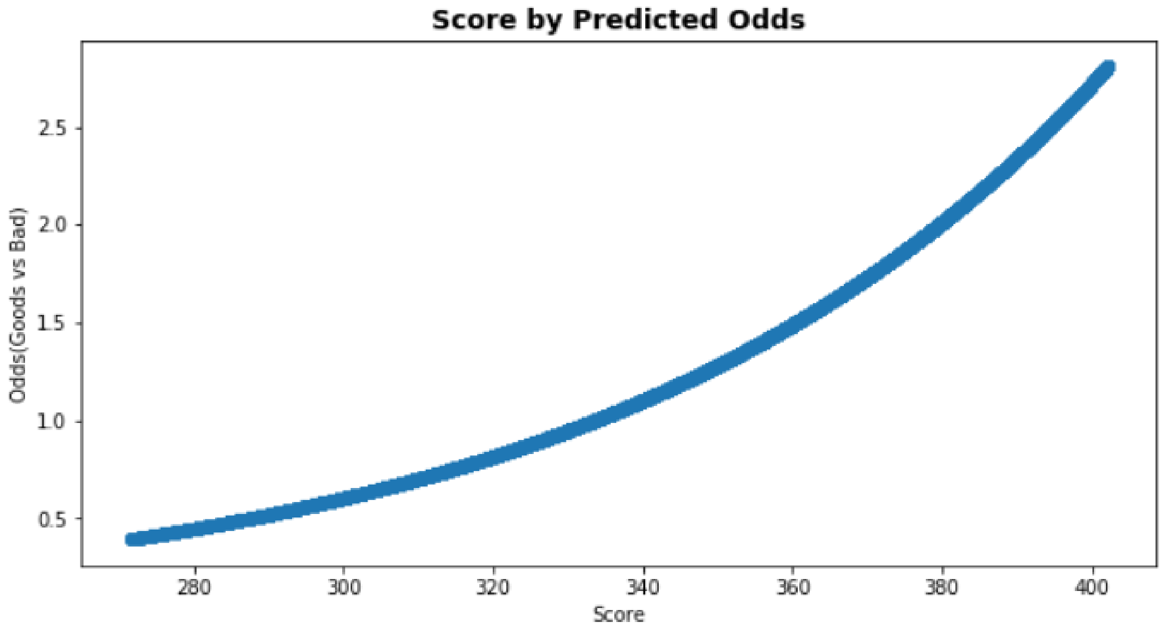
Rejected Population Scorecard Plots

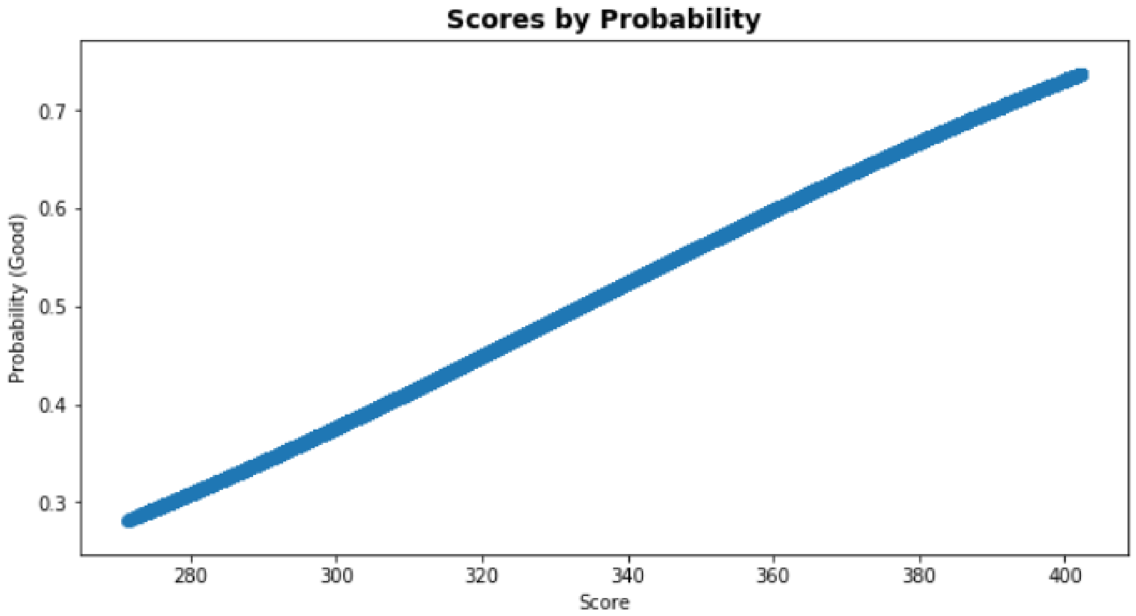


Application Scorecard

- **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