Group Case Study - BFS

Submitted By:

Harshad Ambekar

Rittik Saha

Business Understanding

- CredX is a leading credit card provider that gets thousands of credit card applications every year but the company is experience a credit loss in the recent years.

- Company want to build best strategy to mitigate credit risk is to acquire the right customers.

- So, we need to identify the right customers for the company using Predictive Models thereby determining the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Solution Approach

- As we understand it is a supervised classification business problems. We need to build a predictive model to identify the customers who are at a risk of defaulting if offered a credit card using Logistic Regression & Random Forest.
- After that we imply x-fold validation technique to find the best approach.
- We consider a series of steps:
  - Business Requirement
  - Data Understanding
  - Data Preparation
  - Work of Evidence (WoE) and Information Value (IV) analysis
  - Model Development
  - Model Evaluation and Conclusion
  - Summary
  - Application Scorecard Analysis
  - Assessing the Financial Benefit

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.

Missing values of Demographic/application data

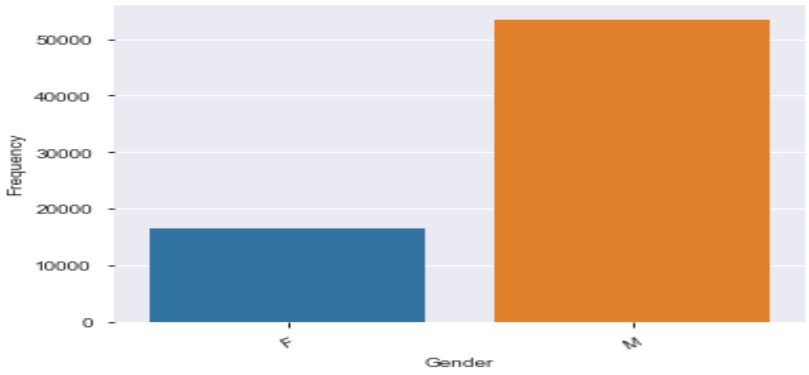
Column Header	Count	Percentage
Performance Tag	1425	2.00
Education	119	0.17
Profession	14	0.02
Type of residence	8	0.01
Marital Status (at the time of application)	6	0.01
No of dependents	3	0.00
Genders	2	0.00

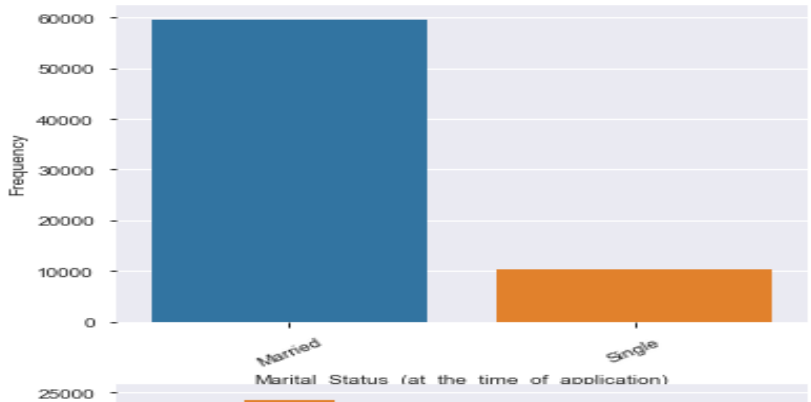
Some more highlights:

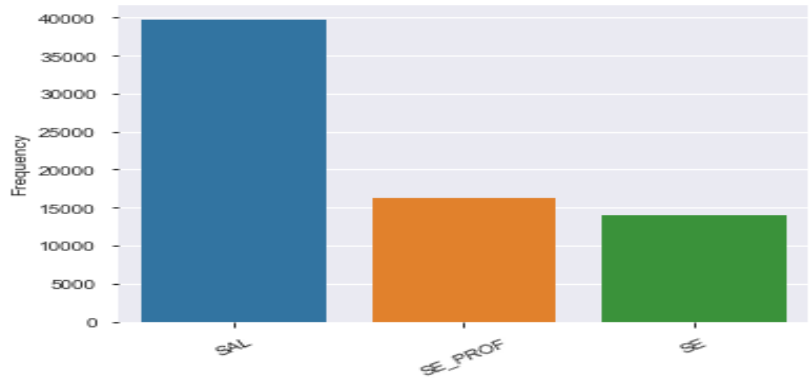
Age: 20 wrong data ranging from -3 to 0

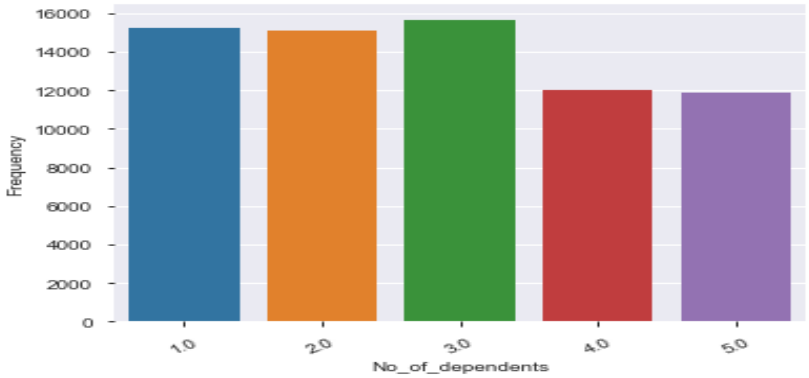
Income: 81 rows have income less than 0

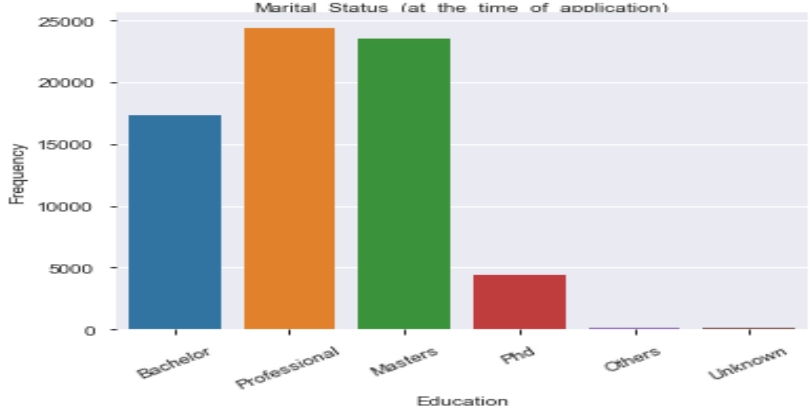
EDA Analysis of Demographic Data

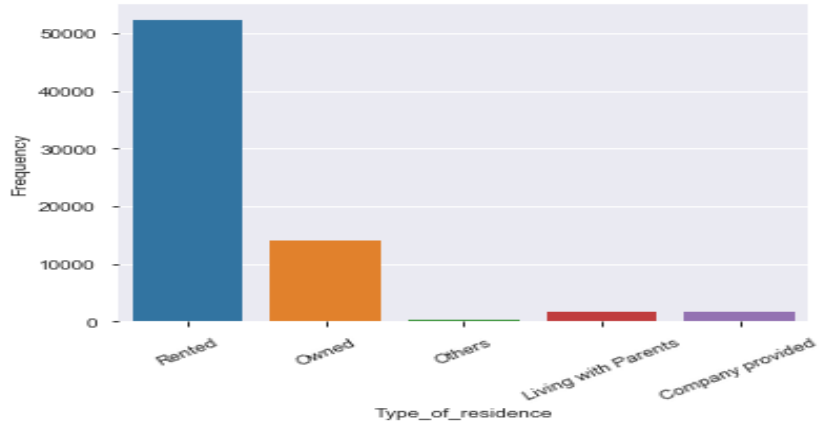












Findings:

- ✓ There are more males than females in the given dataset.
- ✓ More married people than Singles in the given dataset.
- ✓ Most of the people are Salaried employees
- ✓ Almost evenly distributed for all the no. of dependent field, but people with 4 and 5 dependents are lesser than the people with no. of dependents as 1 to 3
- ✓ Most of the people lives in Rented houses, followed by houses owned by them.

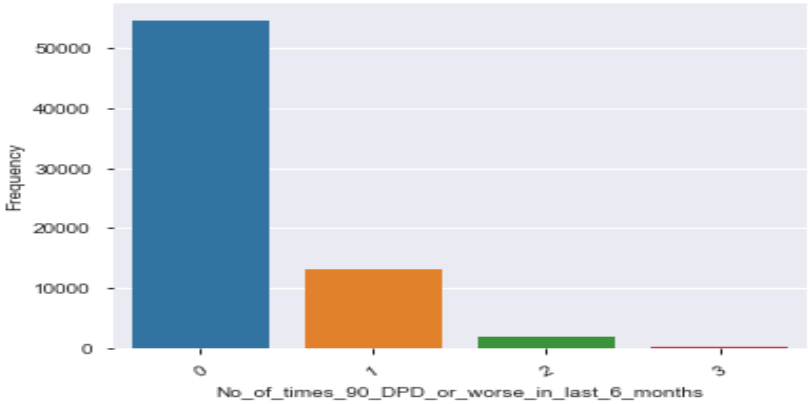
Missing values of Credit Bureau Data

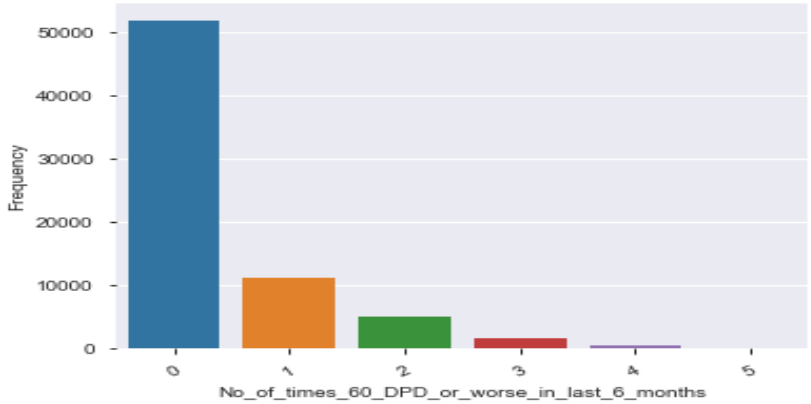
Column Header	Count	Percentage
Performance Tag	1425	2.00
Avgas CC Utilization in last 12 months	1058	1.50
No of trades opened in last 6 months	272	0.38
Outstanding Balance	272	0.38
No of trades opened in last 6 months	1	0.00

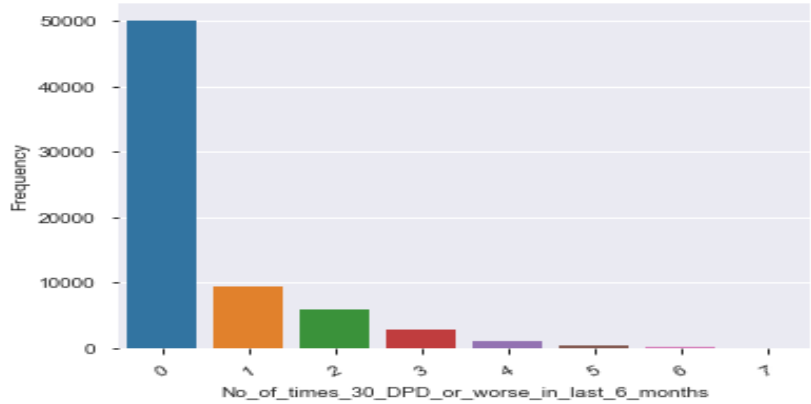
Some more highlights (highly correlated):

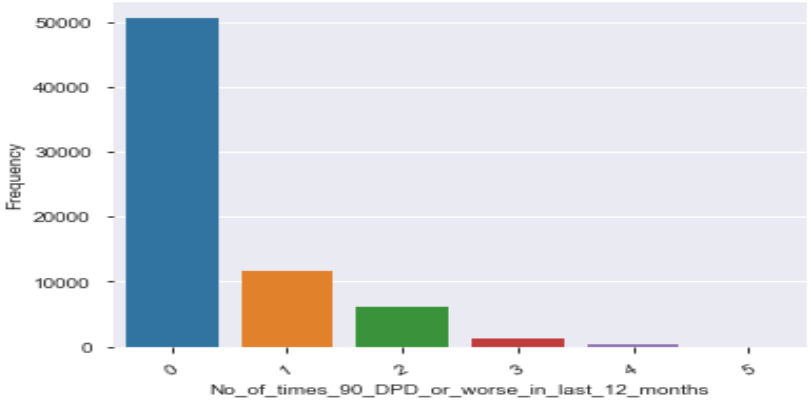
- ✓ Total No of Trades is highly correlated with No of trades opened in last 12 months
- ✓ No of times 30 DPD or worse in last 6 months is highly correlated with No of times 30 DPD or worse in last 12 months
- ✓ No of times 60 DPD or worse in last 12 months is highly correlated with No of times 30 DPD or worse in last 6 months
- ✓ No of times 60 DPD or worse in last 6 months is highly correlated with No of times 60 DPD or worse in last 12 months
- ✓ No of trades opened in last 12 months is highly correlated with No of PL trades opened in last 12 months
- ✓ No of trades opened in last 6 months is highly correlated with No of trades opened in last 12 months
- ✓ Presence of open home loan is highly correlated with Outstanding Balance

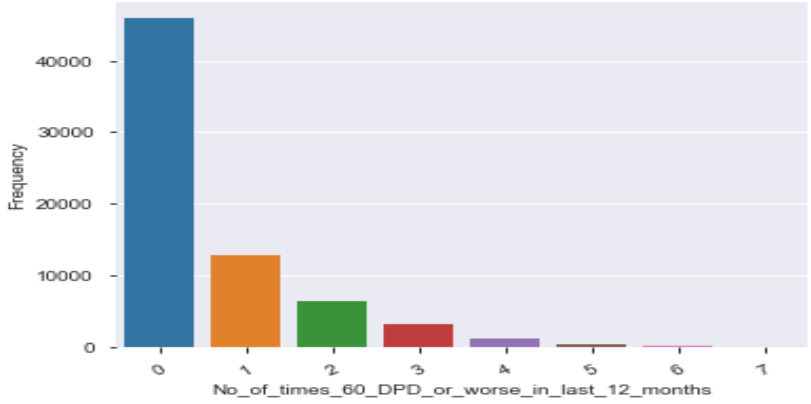
EDA Analysis of Credit Bureau Data

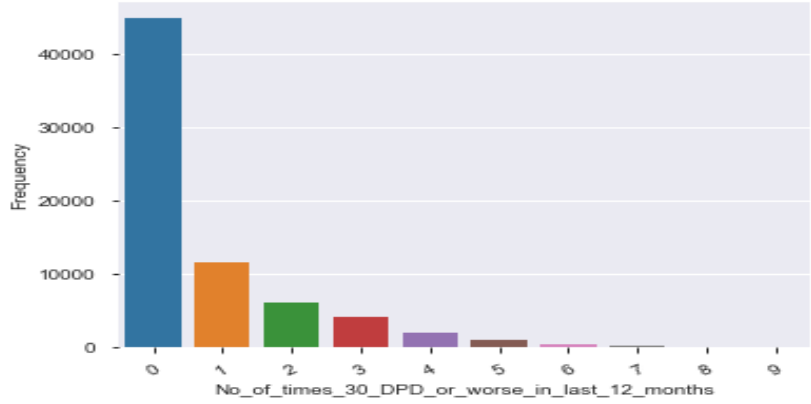












Finding:

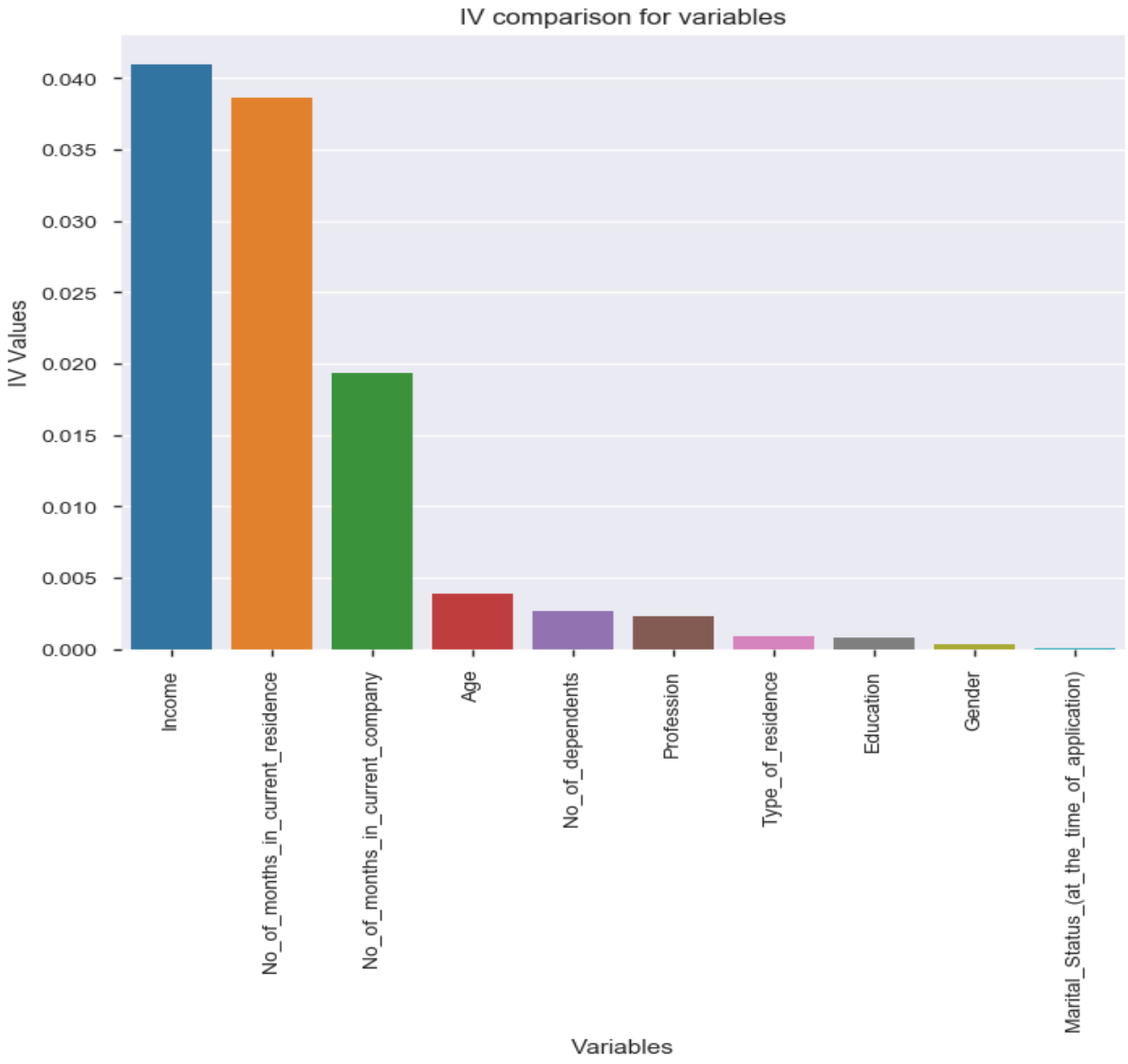
- ✓ Majority of the customers are good customers as they have not defaulted in the last 6 months(value 0 for 90 DPD)
- ✓ Majority of the customers are good customers as they have not defaulted in the last 6 months(value 0 for 60 DPD)
- ✓ Majority of the customers are good customers as they have not defaulted in the last 6 months(value 0 for 30 DPD)
- ✓ Majority of the customers are good customers as they have not defaulted in the last 12 months(value 0 for 90 DPD)
- ✓ Majority of the customers are good customers as they have not defaulted in the last 12 months(value 0 for 60 DPD)
- ✓ Majority of the customers are good customers as they have not defaulted in the last 12 months(value 0 for 30 DPD)

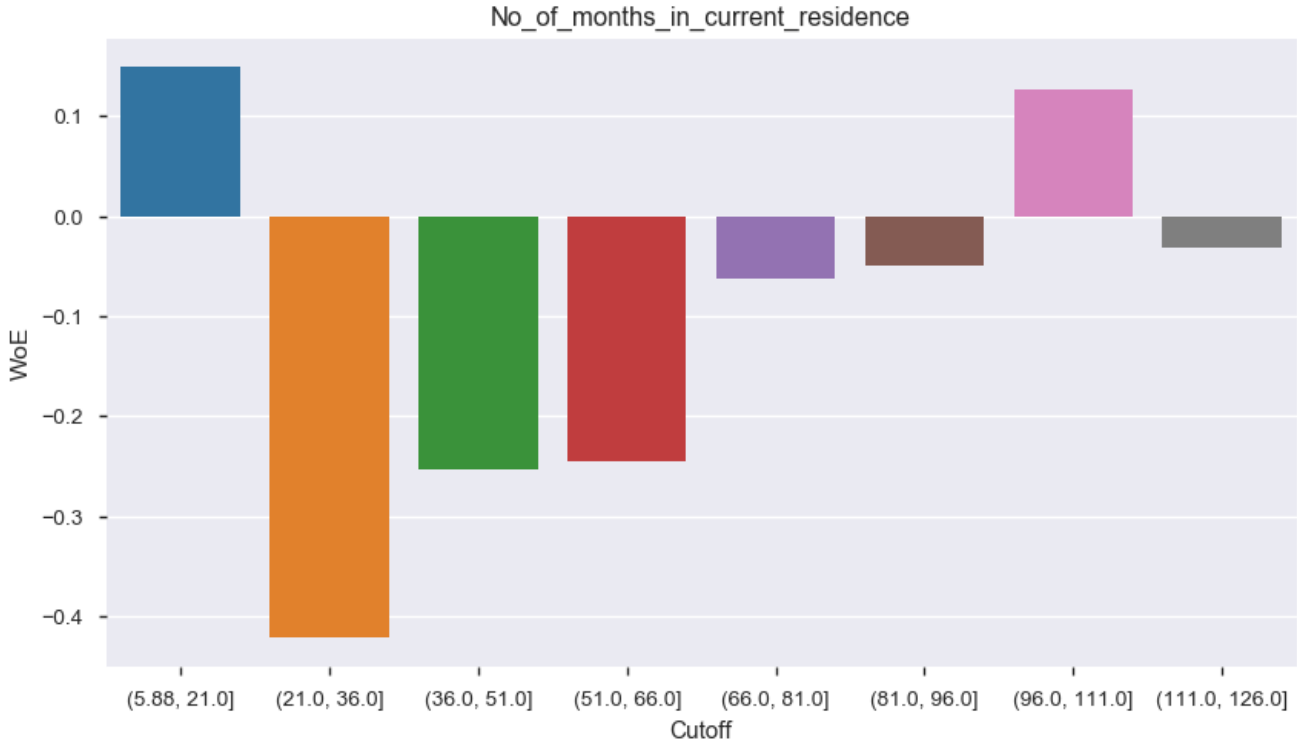
WoE / IV Analysis for Demographic Data

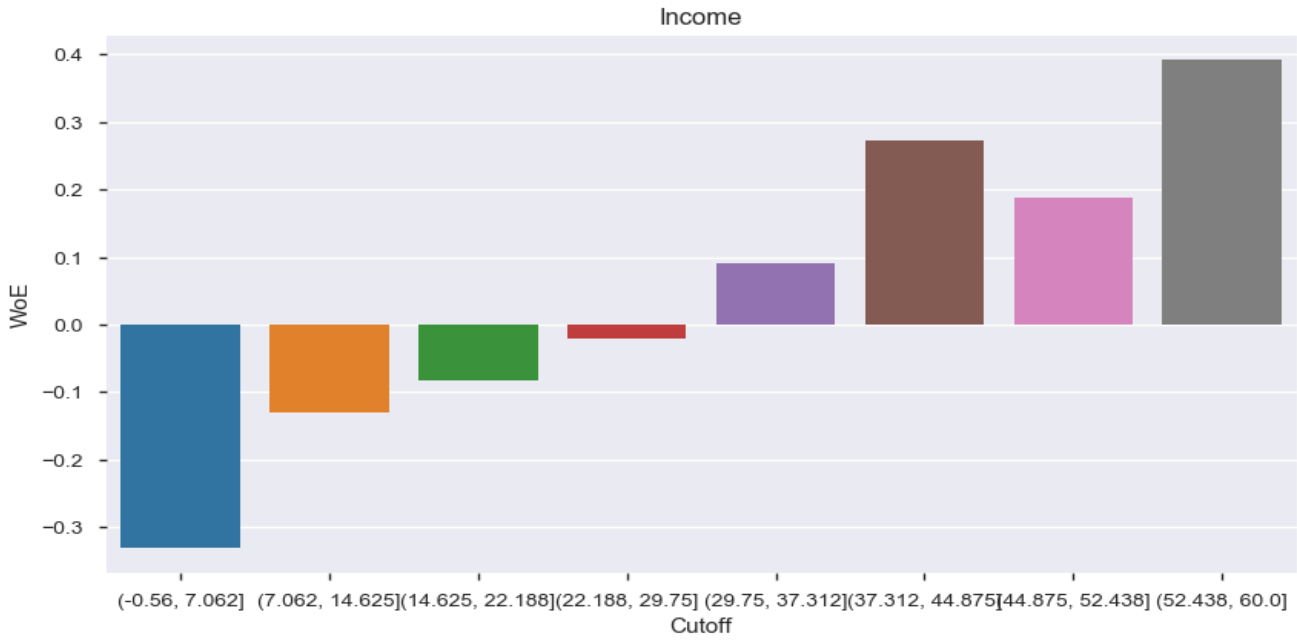
As per IV Rules, IV value of less than 0.02 is not useful so **Income** & **No of months in current residence** seems to be the useful predictor variables here,

Variable	IV
Income	0.040995
No_of_months_in_current_residence	0.038627
No_of_months_in_current_company	0.019302
Age	0.003863
No_of_dependents	0.002653
Profession	0.002289
Type_of_residence	0.000918
Education	0.000765
Gender	0.000320
Marital_Status_(at_the_time_of_application)	0.000093

Graphical Details





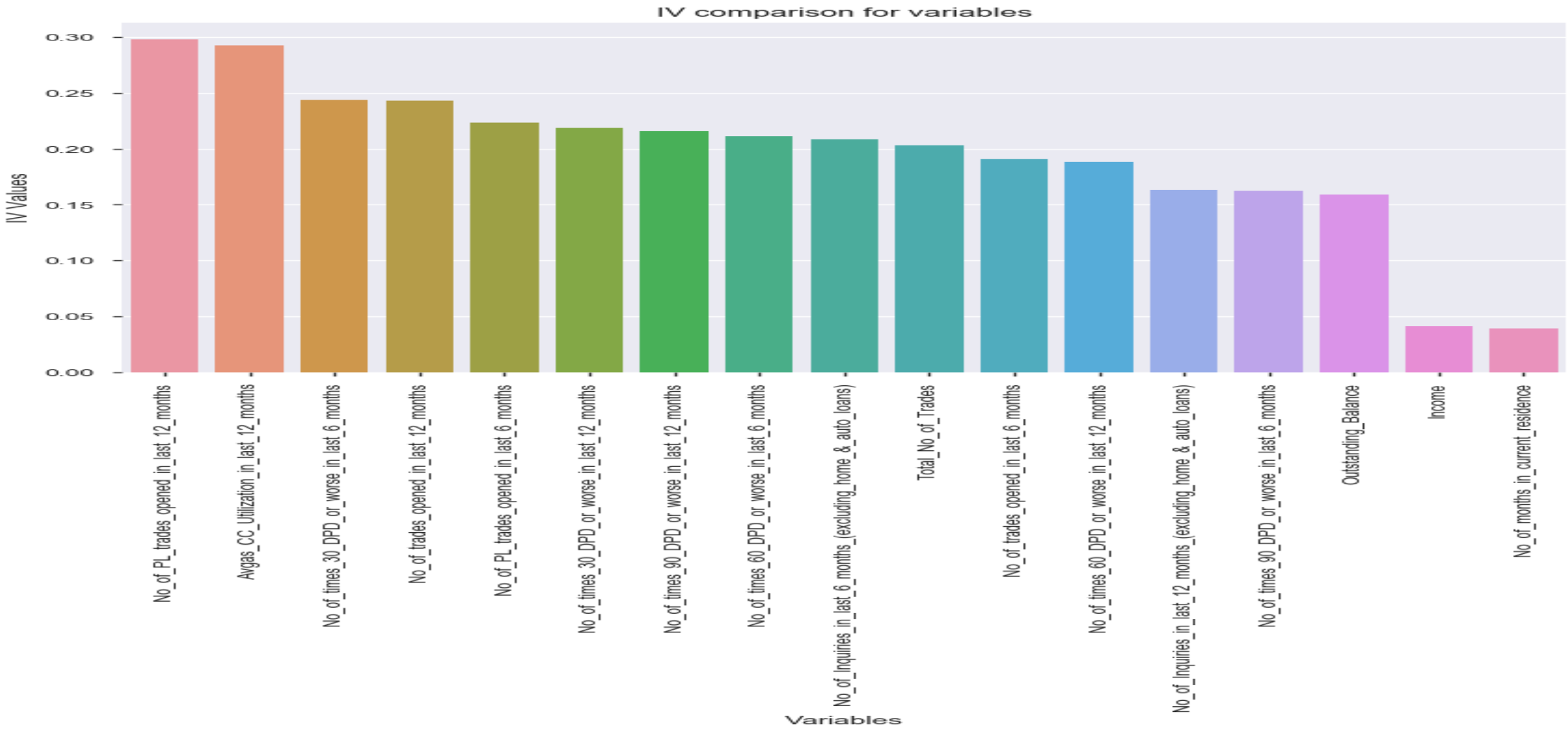


WoE / IV Analysis on Combine Dataset

The list of Variables in Master dataset and as per IV rules, the IV values are consider IV values higher than 0.02

Variable	IV
No_of_PL_trades_opened_in_last_12_months	0.298316
Avgas_CC_Utilization_in_last_12_months	0.292881
No_of_times_30_DPD_or_worse_in_last_6_months	0.244069
No_of_trades_opened_in_last_12_months	0.243041
No_of_PL_trades_opened_in_last_6_months	0.223782
No_of_times_30_DPD_or_worse_in_last_12_months	0.218508
No_of_times_90_DPD_or_worse_in_last_12_months	0.215683
No_of_times_60_DPD_or_worse_in_last_6_months	0.211201
No_of_Inquiries_in_last_6_months_(excluding_ho...	0.208701
Total_No_of_Trades	0.203117
No_of_trades_opened_in_last_6_months	0.191031
No_of_times_60_DPD_or_worse_in_last_12_months	0.188191
No_of_Inquiries_in_last_12_months_(excluding_h...	0.163104
No_of_times_90_DPD_or_worse_in_last_6_months	0.162738
Outstanding_Balance	0.159139
Income	0.040893
No_of_months_in_current_residence	0.038784

Graphical Details



Model Building Approach

- **We have 2 dataset:**
  - WOE Transformed Demographics Dataset
  - WOE Transformed Combine Dataset

- **4 Models for Each dataset:**
  - Logistic Regression Model
  - Logistic Regression with Regularization Model
  - Decision Tree Model
  - Random Forest Model

Model Building Results for WoE transformed Demographics Dataset

Model	Accuracy	Precision	Recall	Specificity
	(Test data)	(Test data)	(Test data)	(Test data)
Logistic Regression	79.40%	4.79%	20.80%	82.00%
Regularized Logistic Regression	38.30%	3.72%	54.55%	37.60%
Decision Tree	70.04%	5.71%	39.31%	71.80%
Random Forest	61.70%	5.35%	49.37%	62.10%

Hyperparameters for the models:

- Logistic Regression Model
  - AUC: 0.54
  - Cut off point: 0.05
- Decision Tree Model
  - max_depth : 4
  - min_samples_leaf : 200
  - min_samples_split : 100
  - Criterion : entropy
  - class_weight : balanced
- Regularized Logistic Regression
  - C : 0.001
  - Penalty : L1
- Random Forest Model
  - max_depth : 5
  - min_samples_leaf : 400
  - min_samples_split : 400
  - n_estimators : 900
  - max_features : 10

Model Building Results for WoE transformed Combine (Demographics and Credit Bureau) Dataset

Model	Accuracy (Test data)	Precision (Test data)	Recall (Test data)	Specificity (Test data)
Logistic Regression	62.40%	6.67%	60.36%	66.18%
Regularized Logistic Regression	52.40%	6.33%	73.68%	51.50%
Decision Tree	59.10%	6.70%	66.63%	58.70%
Random Forest	60.30%	6.96%	67.41%	60.00%

Hyperparameters for the models:

- Logistic Regression Model
  - AUC: 0.66
  - Cut off point: 0.05
- Decision Tree Model
  - max_depth : 6
  - min_samples_leaf : 150
  - min_samples_split : 100
  - criterion : entropy
  - class_weight : balanced
- Regularized Logistic Regression
  - C : 0.001
  - Penalty : L1
- Random Forest Model
  - max_depth : 4
  - min_samples_leaf : 500
  - min_samples_split : 500
  - n_estimators : 400
  - max_features : 10

Summary

Model Building Results for WoE transformed Demographics Dataset:

- ✓ All the models built on the demographics dataset seems to have week predictive power.
- ✓ Best Recall we have got is from the Random Forest Model which is around 54% on the training dataset and 49% on the test dataset. This obviously can not be considered to be a good model.

Model Building Results for WoE transformed Combine (Demographics and Credit Bureau) Dataset:

- ✓ Models on master dataset are performing much better than that of models built on demographics dataset.
- ✓ Decision Tree and Random Forests with hyperparameter tuning have produced decent results.
- ✓ Regularized Logistic Regression model has produced the best results with recall greater than 70% on the training and the test dataset.

Model Evaluation and Conclusion

Model Evaluation Parameters:

- ✓ Always Looking for Higher Accuracy and optimize Recall.
- ✓ Showing good result in Test Dataset.
- ✓ Accuracy score along with confusion matrix should be prepared for each model.
- ✓ Gini-Index needs to be evaluated for ensumble methods models.
- ✓ There is a class imbalance in the dataset. This needs to be handled using balanced class during each model preparation.
- ✓ AUC-ROC curve for the model using cut-off values.
- ✓ Work of Evidence (WoE) and Information Value (IV) analysis and developing WoE transformed dataset for Demographic/application data and combine with Credit bureau data

Conclusion:

- ✓ Best Model (Regularized Logistic Regression on Master Dataset)
- ✓ Since Regularized Logistic Regression has produced the best results, we would go ahead with this model to create the scorecard and to calculate the financial benefits of the model.

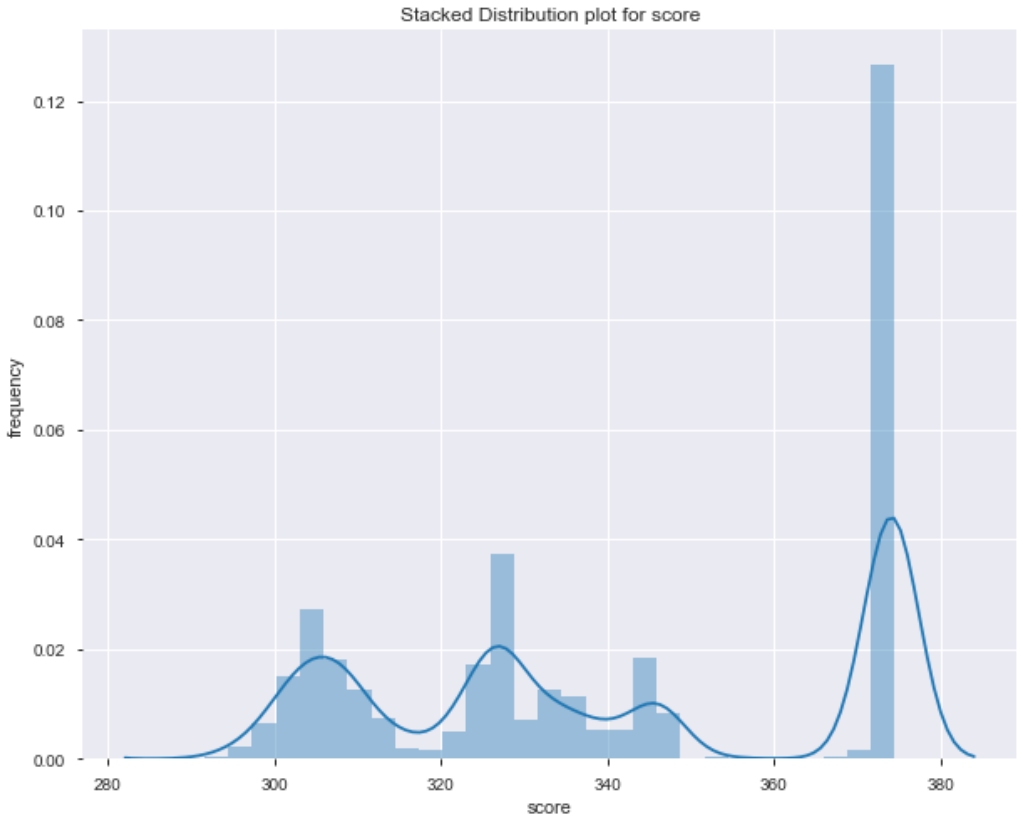
Application Scorecard Analysis

Recommended Cut off : 310

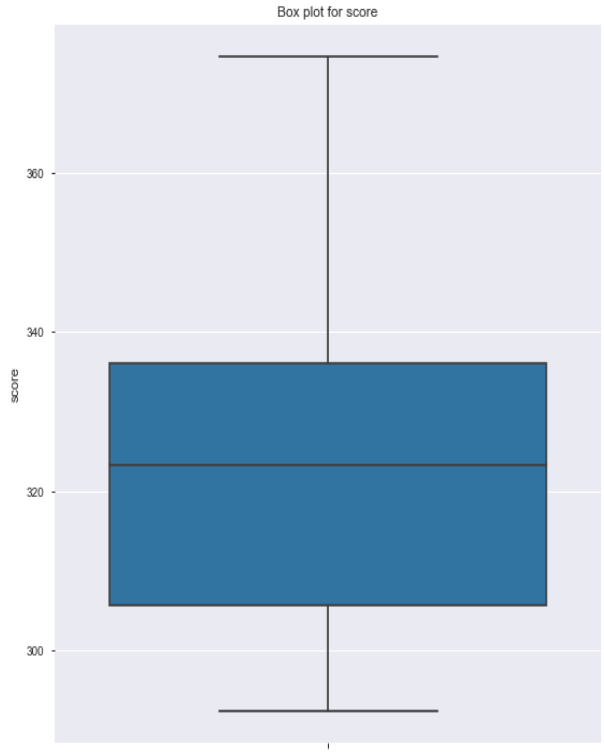
- ✓ Customers with a score less than 310 would not be granted credit card.
- ✓ Majority of the customers falls in the range of 290 to 350
- ✓ Cutoff of 310 correctly identifies almost 89% of the bad customers.
- ✓ If we consider the scorecard built for the master dataset, then almost 21% of the good customers are not going to get the credit card.
- ✓ If we reduce the cutoff from 310 to a lower number then it will defeat the purpose of doing this exercise of identifying the bad customers.
- ✓ Though, if Bank is ready to take the risk they may reduce the cutoff by 5 points, keeping it to 305. A cutoff of 305 would correctly identify 76% of the bad customers, and will impact around 2.5% good customers.

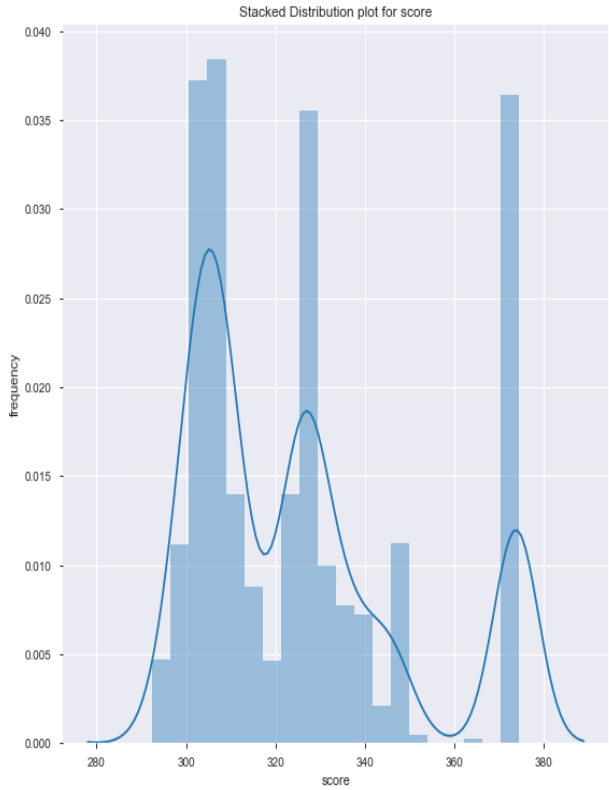
Graphical Details

Plotting for the Whole Dataset



Plotting for the Default Cases





Assessing the Financial Benefit

Application Scorecard:

- ✓ Total number of Customers = 71292 (remember we removed three duplicate reports)
- ✓ Approved Customers = 69867 (there were 1425 records with null values for performance tag, 71292-1425 = 69867)
- ✓ Default Customers = 2947 (Customers with Performance Tag 1)

Assumption to compute P & L :

- ✓ Customer Acquisition Cost (including paper work, phone calls cost, service tax etc.)
- ✓ 50 USD Credit Card Limit = 49,950 USD (taking odd number so that the money at risk is a round figure)
- ✓ Money at Risk per customer = 49,950 + 50 = 50,000 USD
- ✓ Total Money at Risk (Defaulted Customers) = 50,000 × 2947 = 14,73,50,000 USD

Potential Financial Benefit:

- ✓ Model we built has a recall of 74%, hence it can save 74% of 14,73,50,000 USD

Thank You