CREDIT RISK ANALYTICS

(CAPSTONE PROJECT- BFS)

SUBMISSION

SUBMITTED BY:

Swami Prem Pranav Kayashyap (APFE1786831)

UNDER THE GUIDANCE OF:

Sankaran Karthikeyan

Abstract – Credit Risk Analysis

Business Objective:

➢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.

➢To identify the right set of customers for leading credit card provider “CredX” using Predictive models thereby determining the factors affecting credit risk.

Proposed Solution :

➢To build a predictive model which will use the past data of credit card applicants and help CredX identify the right customers to minimize the credit loss.

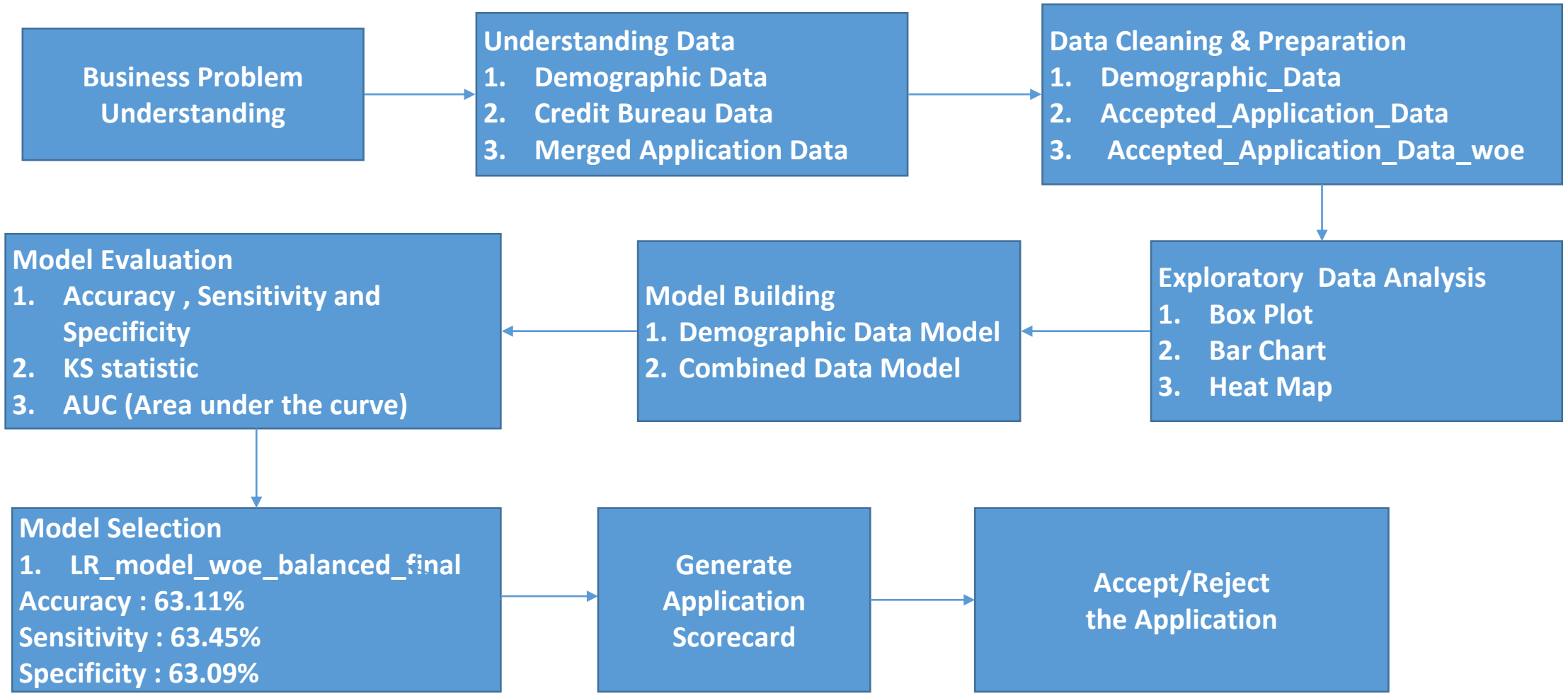
Data sources

➢ Demographic Data - Applicant’s traits

➢ Credit Bureau Data - Applicant’s credit and financial data

➢ Final Data - Demographic + Credit Bureau (merged)

Problem Solving Approach Flow Chart



Data Cleaning and Preparation

Data Cleaning and Preparation

- Identified and deleted duplicate applicants
- Imputed null and erroneous values for numerical data with mean, median… as applicable
- Introduced new category as fit for categorical variables replacing empty space and null
- Excluded applicants with a rejected application
- New data frame with all WOE values has been created
- Both Final Data(original) and WOE Data have been split into 2 based on the PerformanceTag
- Features have been renamed for better readability

Final Data for EDA and Model Building

✓ Accepted_Application_Data

✓ Rejected_Application_Data

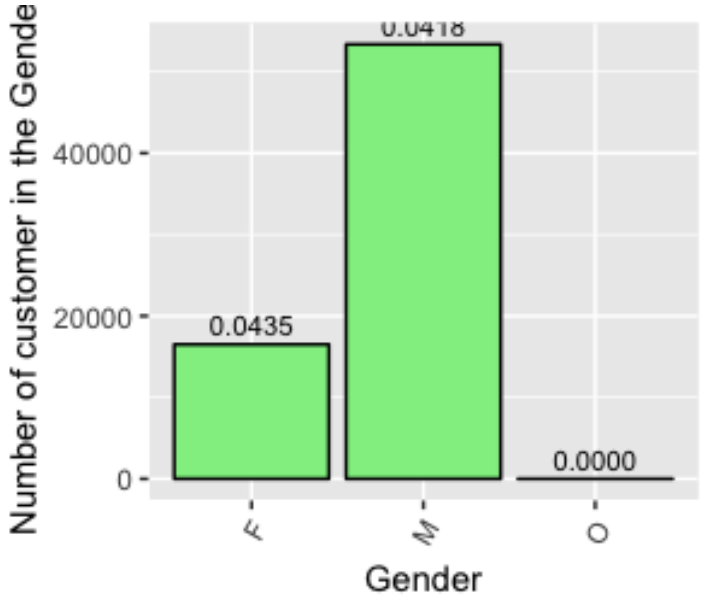
✓ Accepted_Application_Data_woe

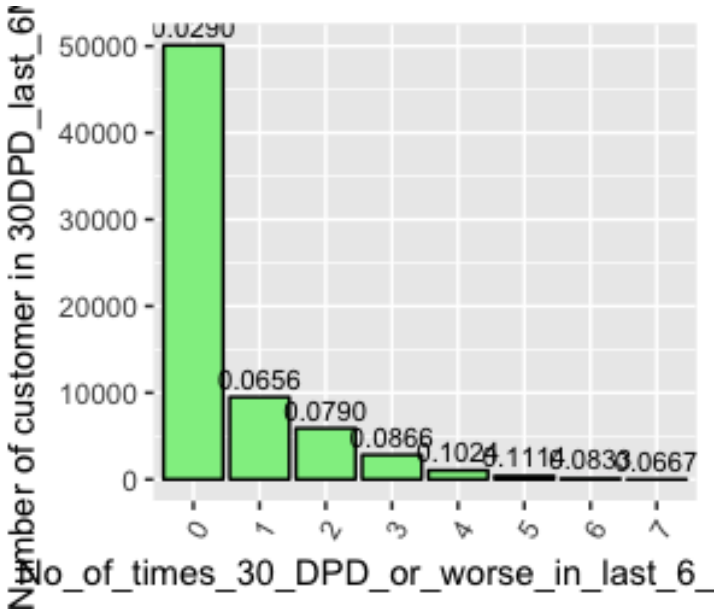
✓ Rejected_Application_Data_woe

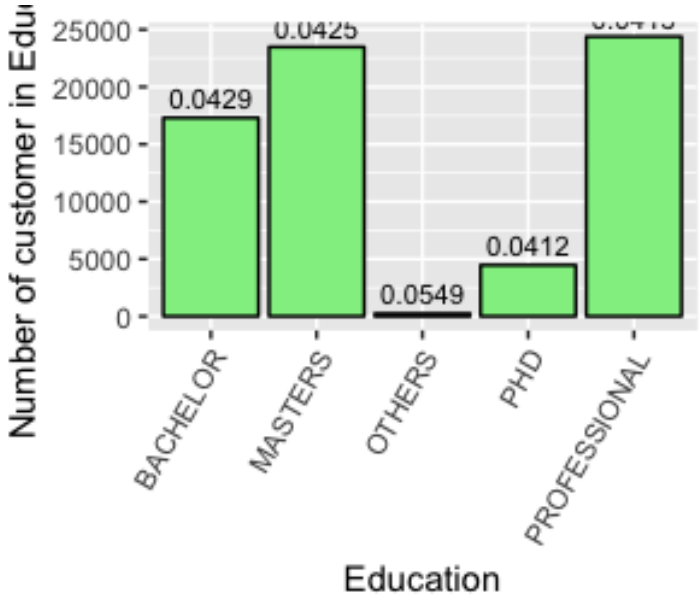
Exploratory Data Analysis

Univariate Analysis

- For most part, data is balanced except for gender
- Irrespective of number of number of days in dpd(30,60,90) and time period(6,12 months) all the dpd variables had clear effect on Performance Tag and follow same pattern
- Most of the applicants are well educated, salaried and live in rental houses
- CC_utilization is maximum in the bins (10-20] and (100-110] but the default rate is on par with mean



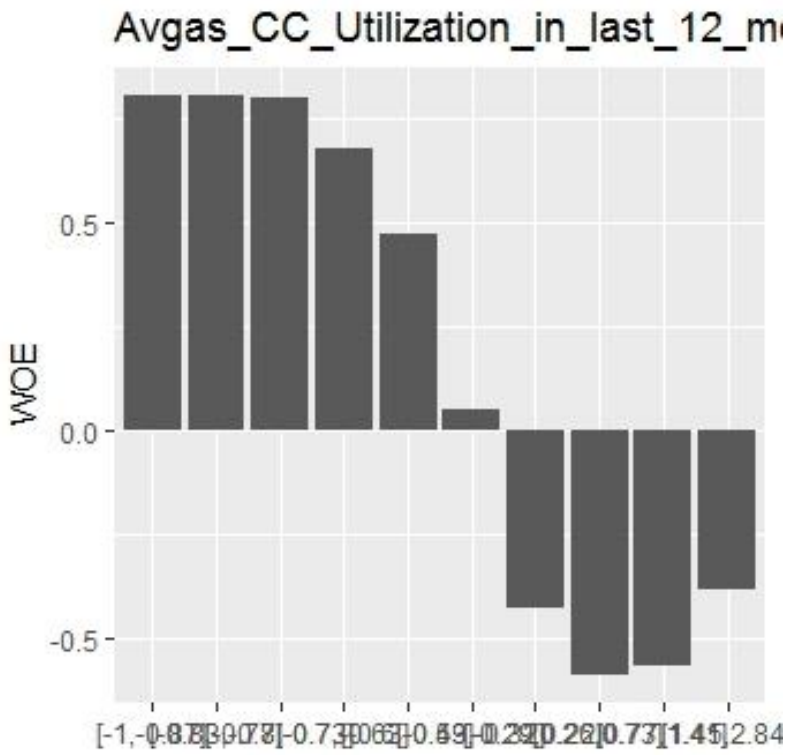


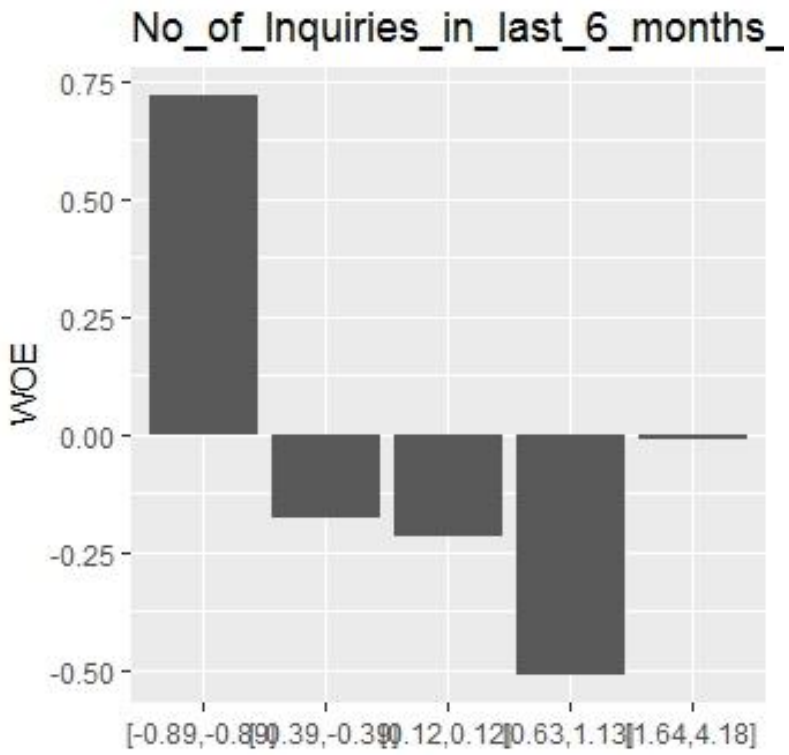


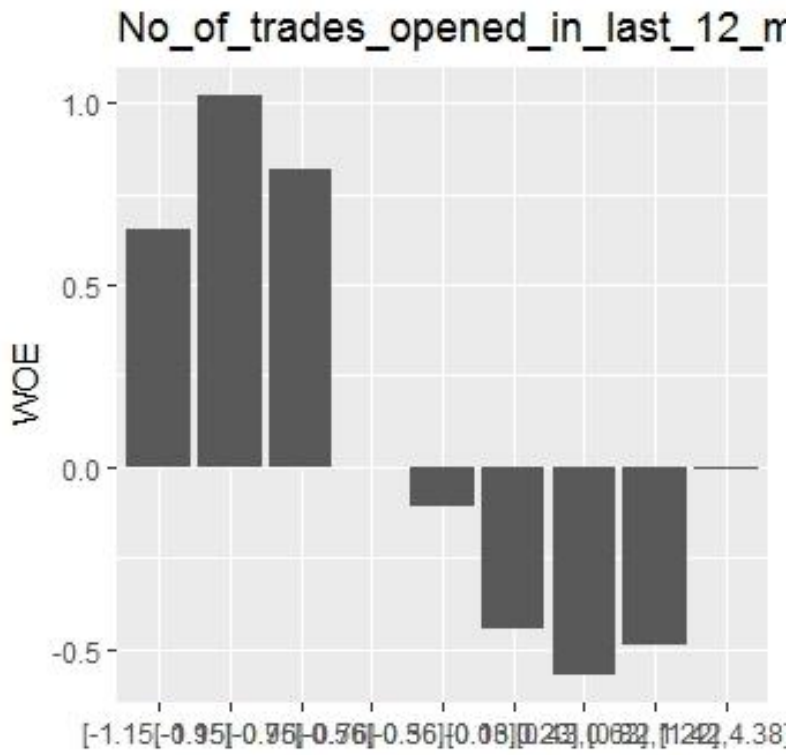
Exploratory Data Analysis

The clear distinction in avgs_cc_utilization implies its a significant predictor and has been confirmed by the Information Value

Along with being significant predictor No_of_inquiries and No_of_trades_opened_in_last_12_months show a pattern that divides data points at approximately same bin

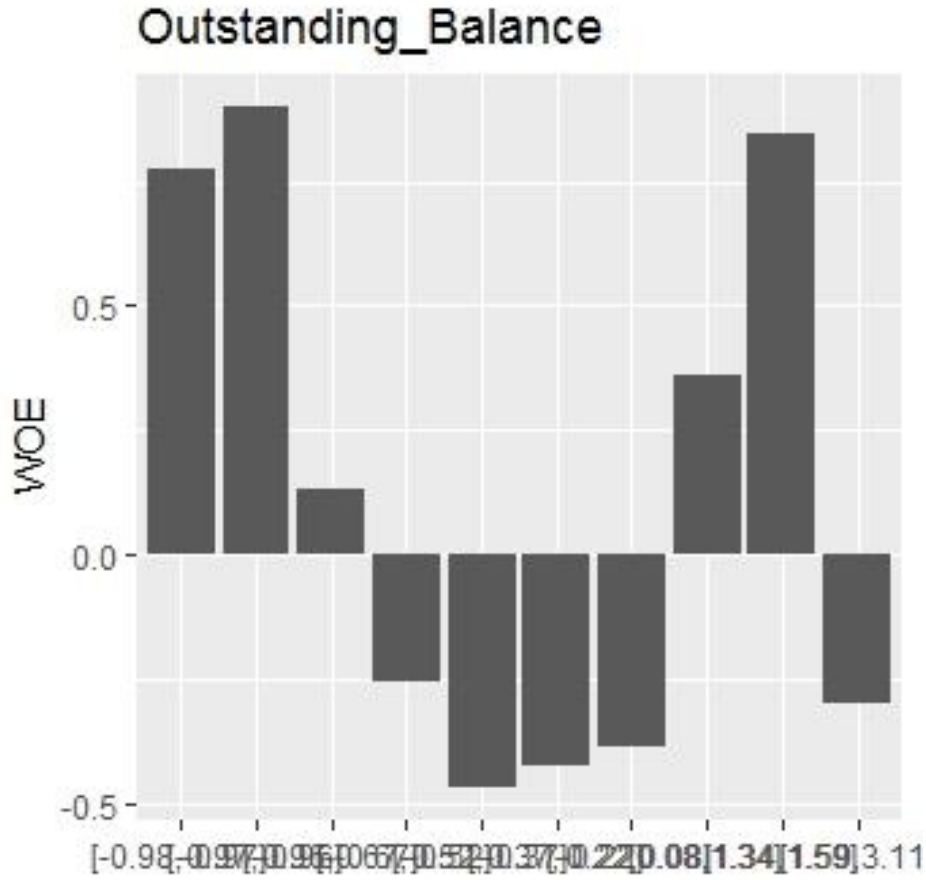






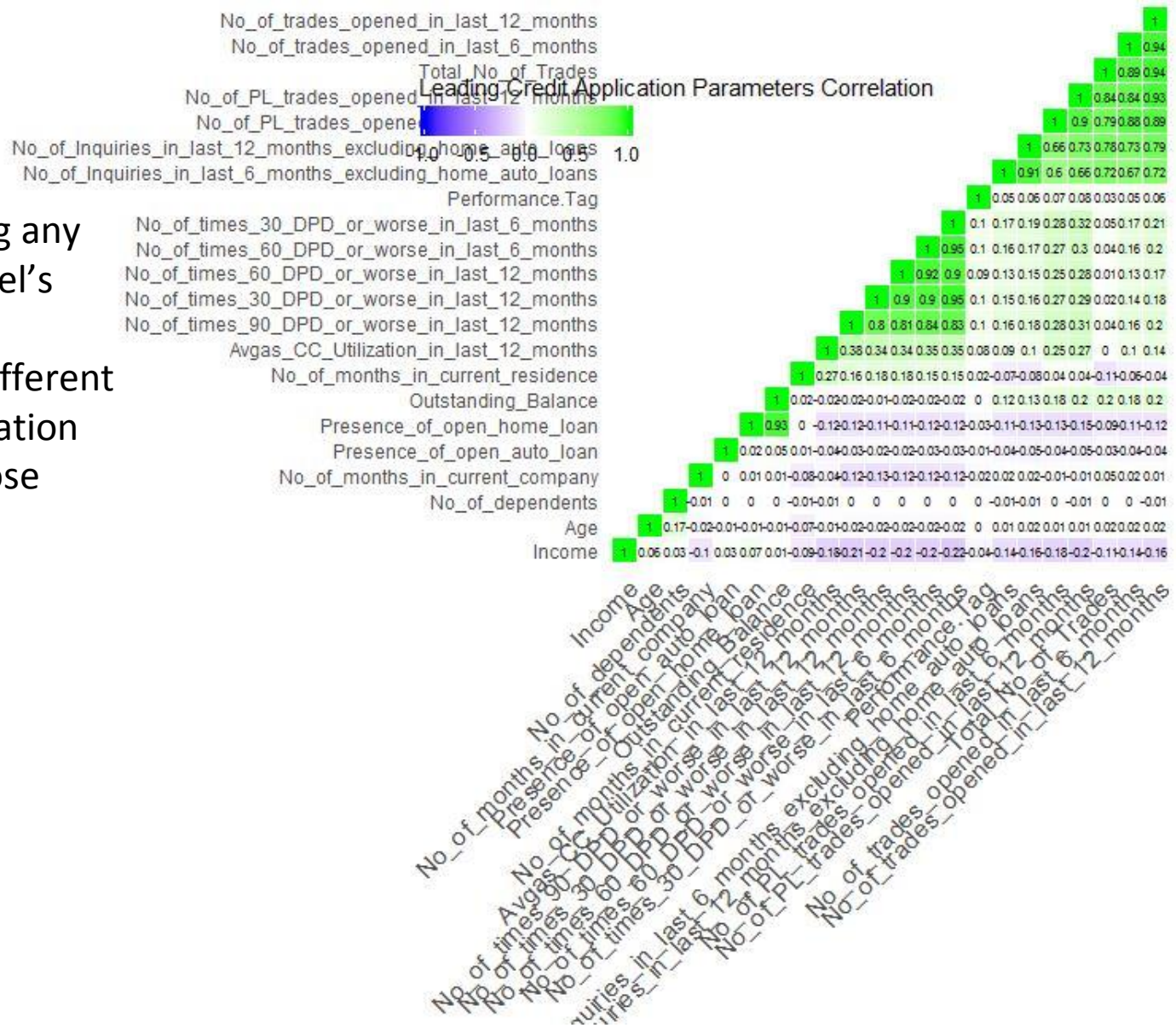
Eccentric Outstanding Balance

- Outstanding balance is often expected to show a progressive pattern as the value increases or decreases
- In the current data sample despite not having an expected structure, it still remained a significant feature to the Performance Tag



Correlation

- The variables with most correlation will not be adding any benefit and so using all of them will hamper the model's performance
- Variables like DPD, trades_opened, inquiries in the different time periods are the ones that have maximum correlation
- Eventually we might be using 1or2 variables from those groups so as to improve the performance



Performance comparison

Metric\Model	Original	Original_Balanced	WOE	WOE_balanced	Demographic_original	Demographic_balanced
Accuracy	0.6504	0.6184	0.6071	0.6311	0.4902	0.5205
Sensitivity	0.65400	0.61847	0.67555	0.63499	0.48513	0.53115
Specificity	0.56547	0.61645	0.60414	0.63091	0.60219	0.52005
KS-Statistic	0.0873	0.2328	0.2796	0.2659	0.0873	0.0512
AUC	0.543	0.6164	0.6398	0.6329	0.54366	0.5256007

Performance Comparison Explained

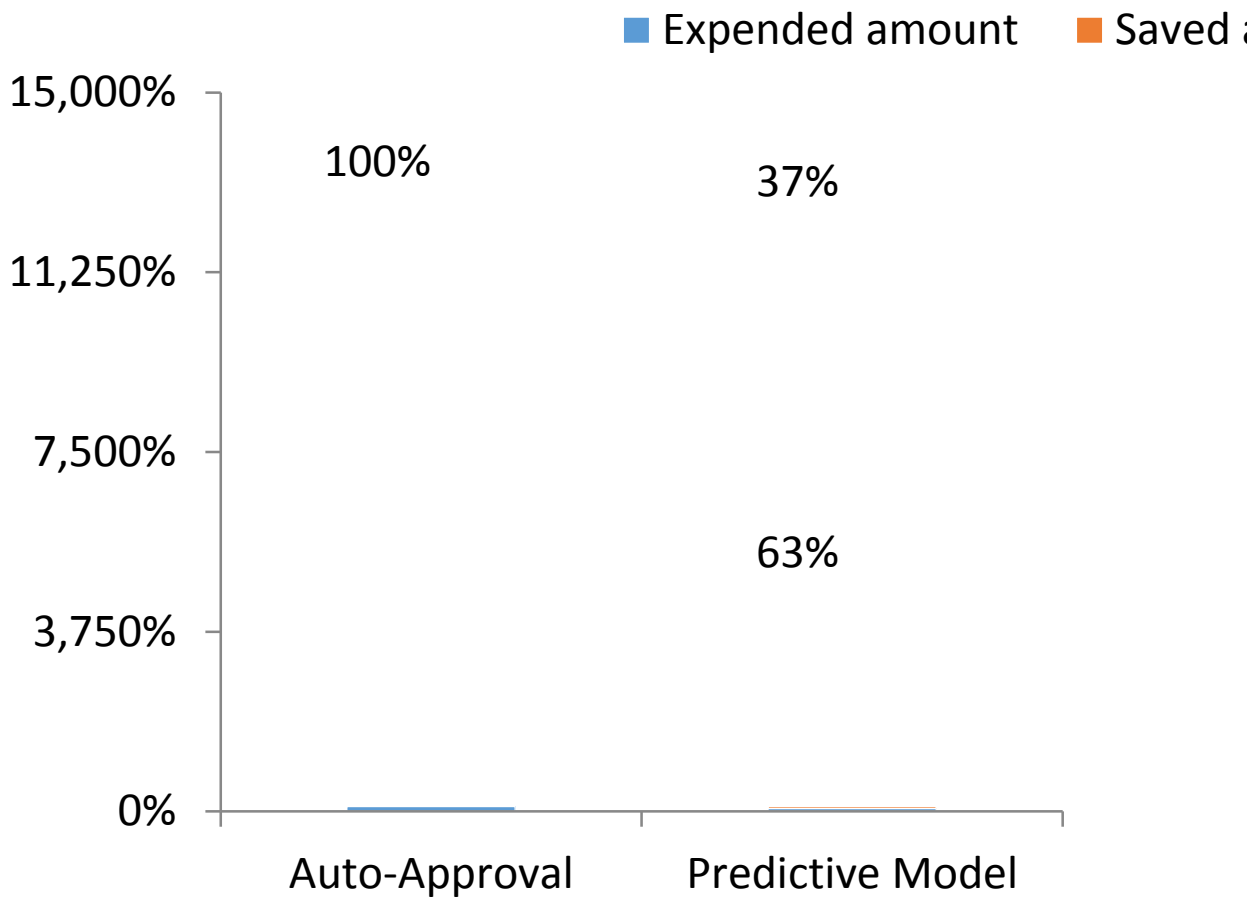
- Logistic Regression has been used as the predictive model to solve problem at hand because of its power with classification data
- Model with only Demographic data is weaker than the one with Demographic and credit data
- The model with highest Accuracy is the one with all original data but it lacks the balance between other statistics
- Final model decided upon is the one with WOE values and balanced data

Application Scorecard

- This application scorecard was prepared with the odds of 10 to 1 at a score of 400 doubling every 20 points. The scorecard is can be calculated using below equation.
- Score = 333.56 + 28.8539 ln (odds) Where, odd = predicted_probability/(1-predicted_probability)
- As per our scorecard, it is implied that any applicant with score less than 300 is a potential defaulter.

Financial Implications

- For a Financial institution that does not use any predictive model, credit cost will be 100%(since there is no model all applicants are approved of a credit)
- But by using our predictive we would only expend 63% of total 100% cost



***Graph was not created with live data, it is just a representation of the model's performance**