BFS Capstone Project Final Submission

- Team:
  - Manasi Khandekar
  - Jay Kumar
  - Rajeev Kumar
  - Rashmi Singh

Executive Summary

Vision:

CredX is a leading credit card provider that gets thousands of credit card applications every year. The CEO believes that the best strategy to mitigate credit risk is to acquire the right customers.

Opportunity:

Recently CredX has experienced an increase in credit loss.

Objective:

To help CredX identify the right customers using predictive models.

Approach summary:

Using past data of CredX's applicants, we will

- determine the factors affecting credit risk
- create strategies to mitigate the acquisition risk
- assess the financial benefit of using the chosen model

Technical Approach

Our approach is underpinned by industry-standard CRISP-DM framework – Business understanding, Data understanding, cleaning and preparation, model building

The problem statement indicates that it is a supervised binary classification problem.

We aim to build a predictive model using classification algorithms to identify delinquent customers who could be potential defaulters for CredX.

We will build multiple models using fundamentals of logistic regression, decision trees, random forests and choose the best model amongst these enabling us to have a better judgement of selecting defaulters.

Data Understanding And Preparation

Data provided:

Demographic data - 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 - 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.

Application ID:

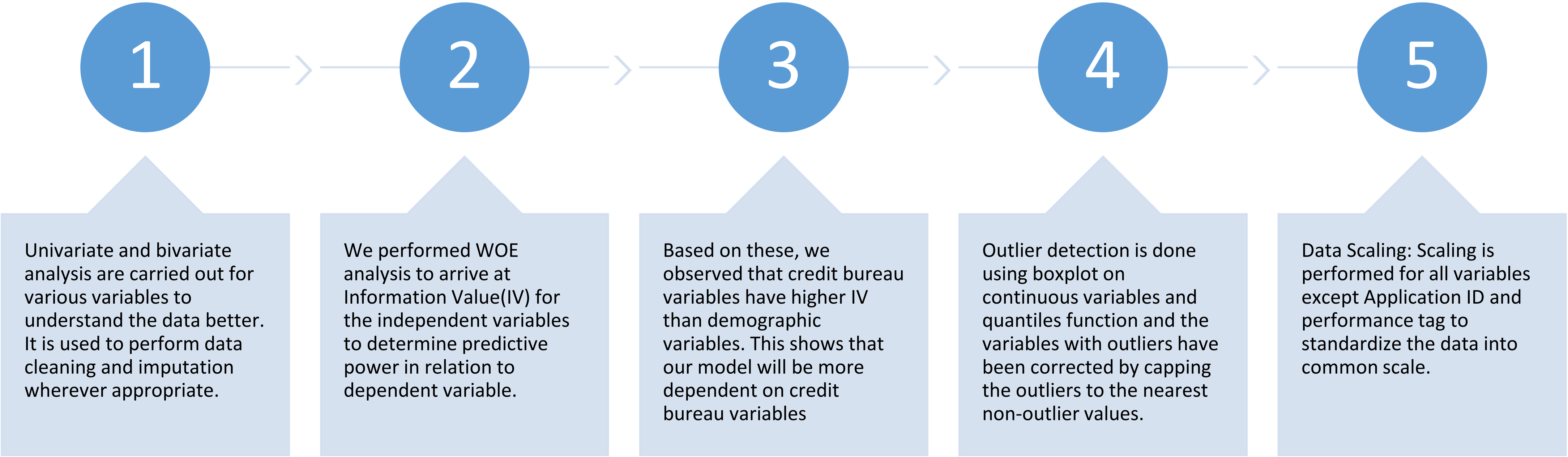
Analysis of Application Id showed there were 3 duplicate ids present. These were identical in both the datasets. The duplicate Application Ids were removed from the datasets.

Remaining Demographic and Credit Bureau dataset is merged based on Application Id into a single dataset.

Performance Tag:

The Performance Tag had 1425 missing values. This implies that these are rejected accounts. The NA values are separated out from both datasets into a rejected Dataset for later analysis

Exploratory Data Analysis - Summary



1

Univariate and bivariate analysis are carried out for various variables to understand the data better. It is used to perform data cleaning and imputation wherever appropriate.

2

We performed WOE analysis to arrive at Information Value(IV) for the independent variables to determine predictive power in relation to dependent variable.

3

Based on these, we observed that credit bureau variables have higher IV than demographic variables. This shows that our model will be more dependent on credit bureau variables

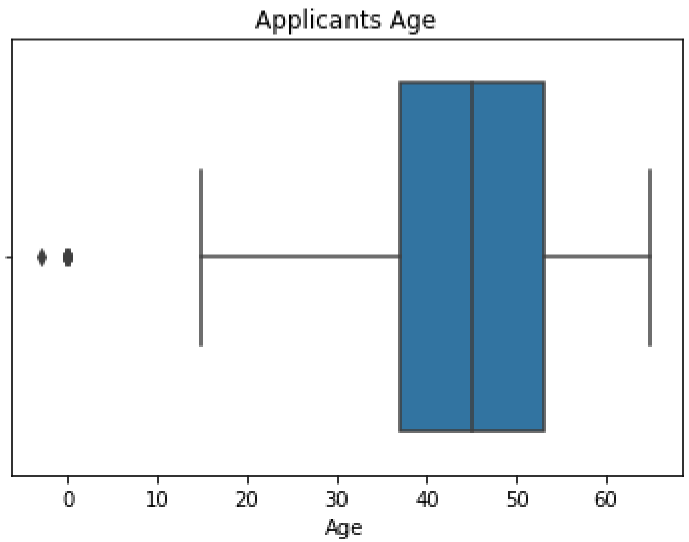
4

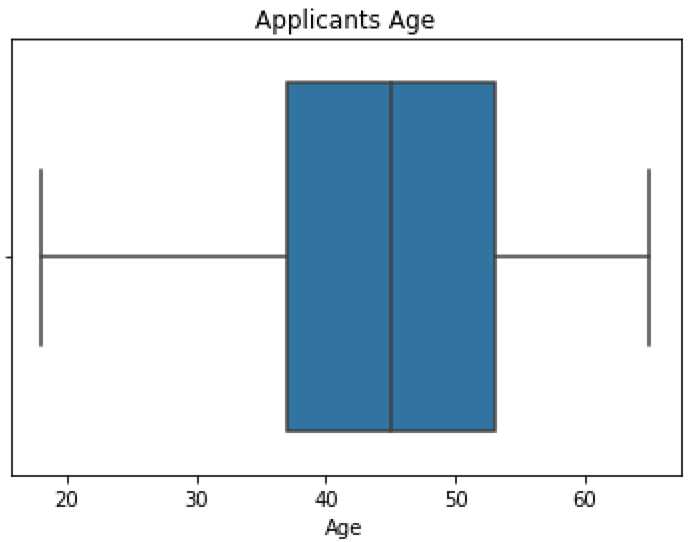
Outlier detection is done using boxplot on continuous variables and quantiles function and the variables with outliers have been corrected by capping the outliers to the nearest non-outlier values.

5

Data Scaling: Scaling is performed for all variables except Application ID and performance tag to standardize the data into common scale.

Exploratory Data Analysis - Detailed





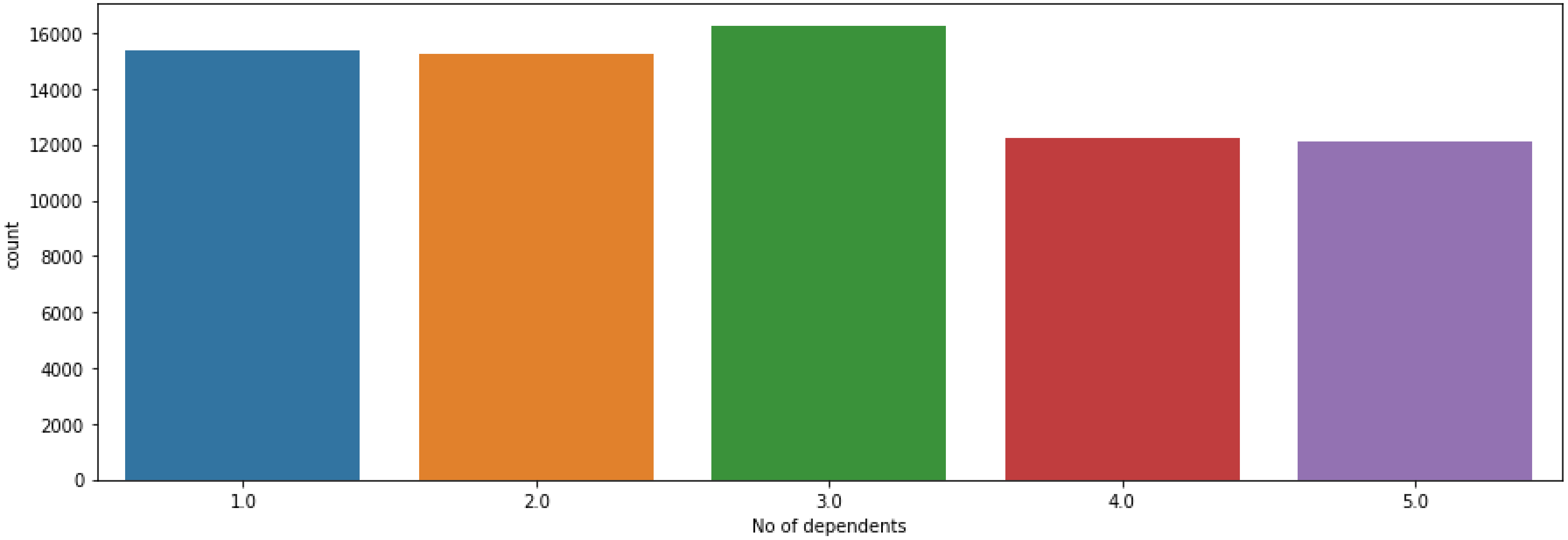
Univariate Analysis on Age:

In chart#1, we can see there are applicants with age less than 18. Those applicants can not be issued credit cards. Also, there are few applicants with 0 or negative value of age. Those records need to be treated/removed from the dataset.

Chart#2 is drawn after removing the above-mentioned applicants from the dataset.

Exploratory Data Analysis - Detailed

Univariate Analysis on Gender: There are 3 times more male applicants than female applicants
Univariate Analysis on Marital Status: There are 6 times more married applicants than single applicants.

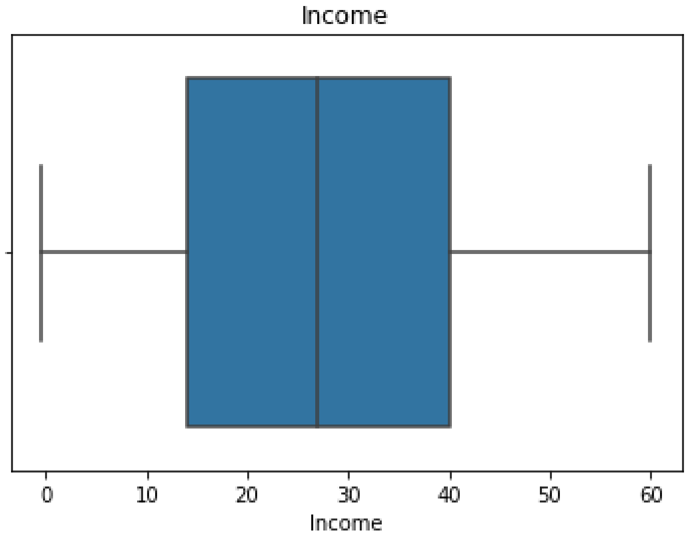


Univariate Analysis on No. of Dependents: The chart shows that the dataset has a good distribution of no of dependents across the applicants.

Exploratory Data Analysis - Detailed

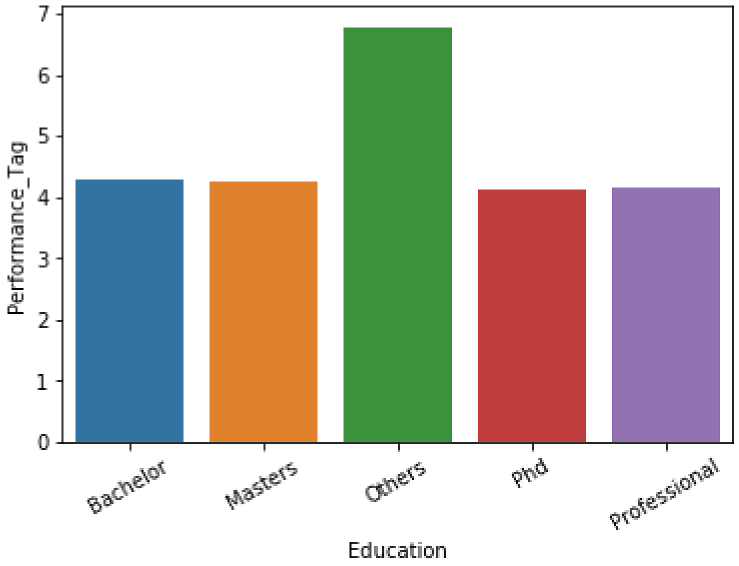
Univariate Analysis on Income:

We can observe from the chart that there are applicants with 0 income or with negative values for income. Those records need to be treated/removed from the dataset.



Univariate Analysis on Education:

We can observe from the chart that most of the applicants are well educated. The highest number of defaulters are in the Others category



Exploratory Data Analysis -Correlation Plot

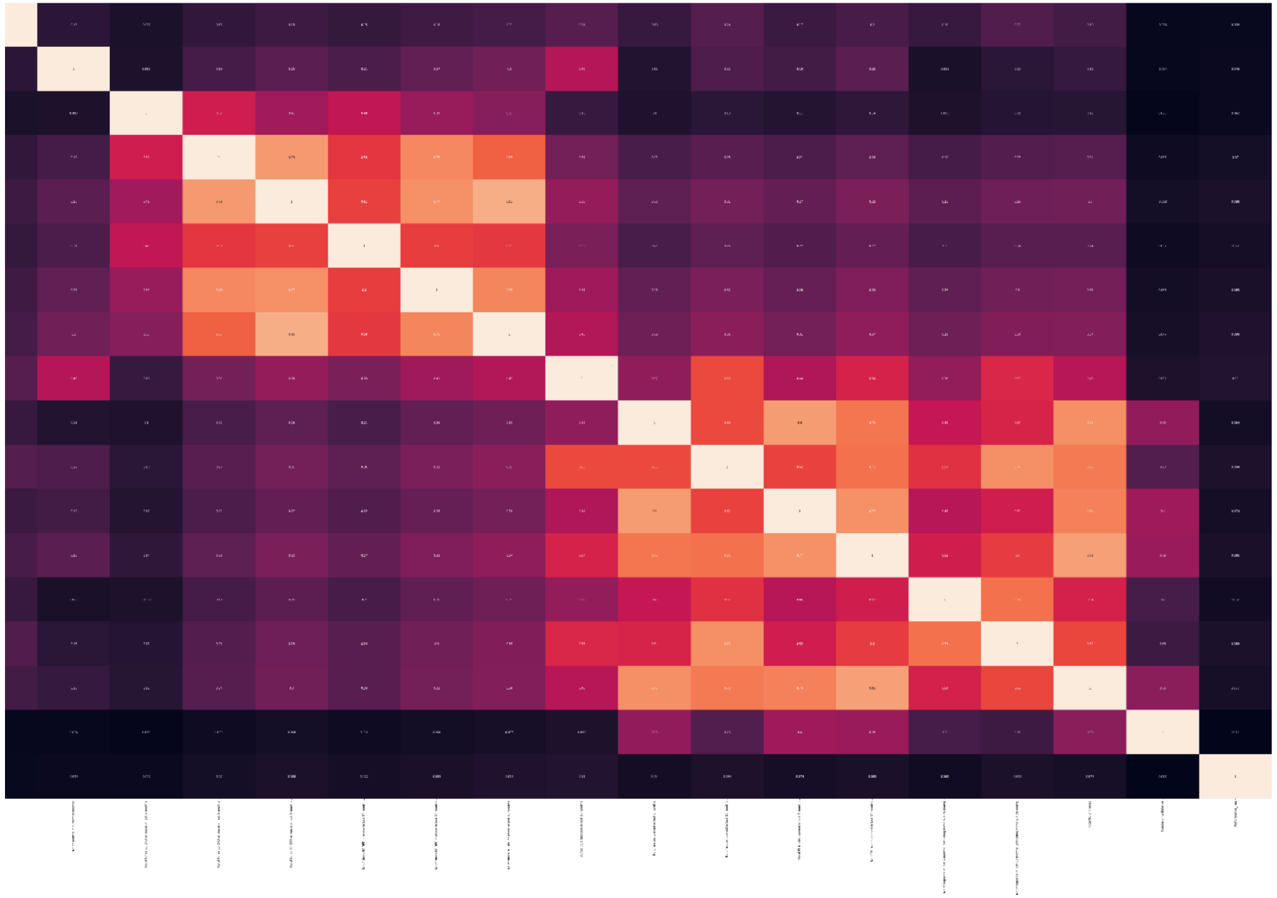
We see two groups of variables being correlated with variables within the dataset

Group1

- No.of.times.90.DPD.or.worse.in.last.6.months
- No.of.times.60.DPD.or.worse.in.last.6.months
- 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.30.DPD.or.worse.in.last.12.months
- No.of.times.60.DPD.or.worse.in.last.12.months
- Avgas.CC.Utilization.in.last.12.months

Group2

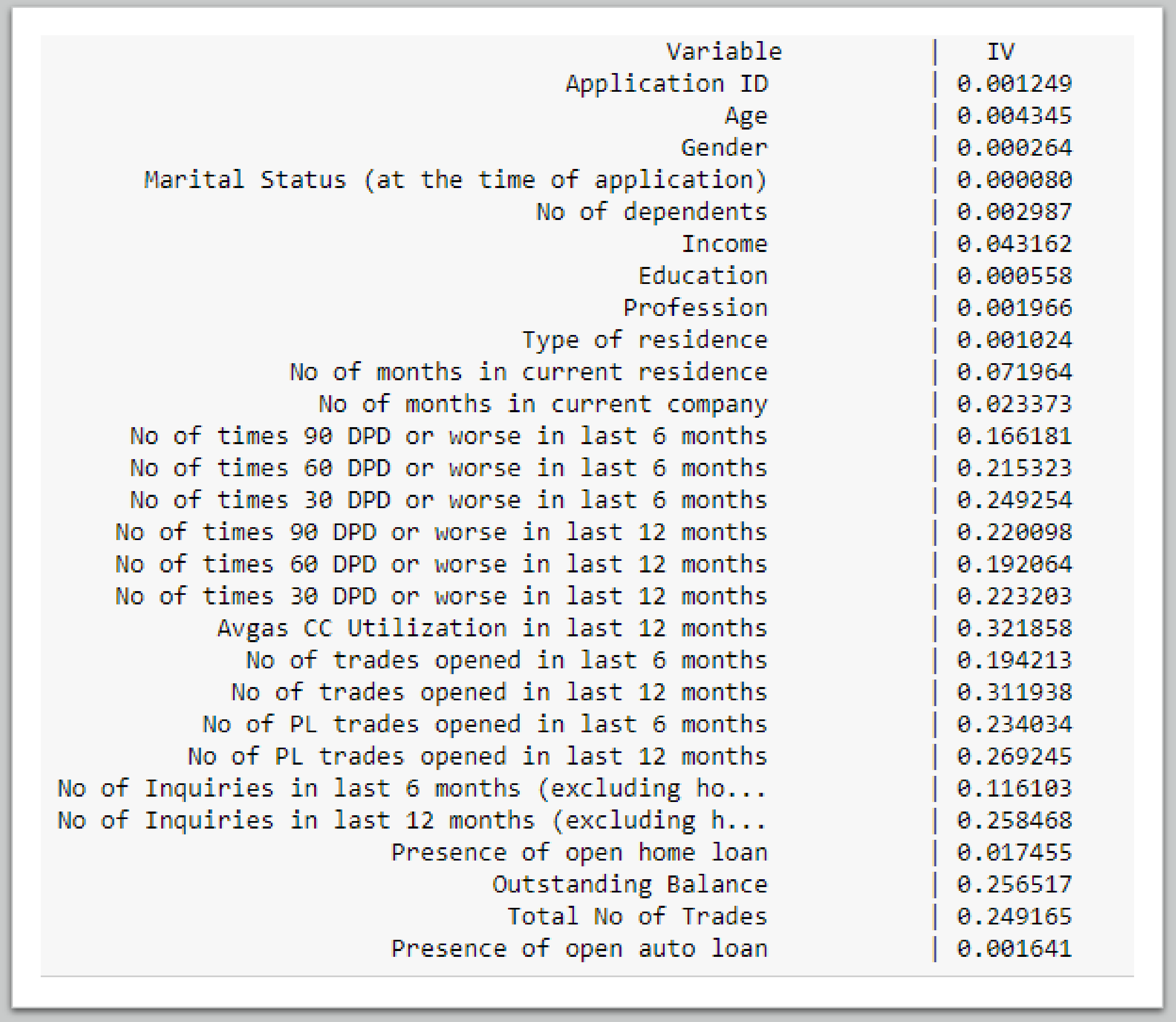
- No.of.trades.opened.in.last.6.months
- No.of.PL.trades.opened.in.last.6.months
- No.of.PL.trades.opened.in.last.12.months
- No.of.trades.opened.in.last.12.months
- Total.No.of Trades
- No.of.Inquiries.in.last.12.months..excluding.home... auto.loans.
- No.of.Inquiries.in.last.6.months..excluding.home...a uto.loans.



WOE and IV analysis

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

- Selecting features with Predictive Powers in the range of 0.1 to 0.5 and above as they are of Medium and strong predictive power



```
                                                Variable  |    IV
                                          Application ID  | 0.001249
                                                     Age  | 0.004345
                                                  Gender  | 0.000264
          Marital Status (at the time of application)  | 0.000080
                                        No of dependents  | 0.002987
                                                  Income  | 0.043162
                                               Education  | 0.000558
                                              Profession  | 0.001966
                                       Type of residence  | 0.001024
                      No of months in current residence  | 0.071964
                        No of months in current company  | 0.023373
           No of times 90 DPD or worse in last 6 months  | 0.166181
           No of times 60 DPD or worse in last 6 months  | 0.215323
           No of times 30 DPD or worse in last 6 months  | 0.249254
          No of times 90 DPD or worse in last 12 months  | 0.220098
          No of times 60 DPD or worse in last 12 months  | 0.192064
          No of times 30 DPD or worse in last 12 months  | 0.223203
                  Avgas CC Utilization in last 12 months  | 0.321858
                    No of trades opened in last 6 months  | 0.194213
                   No of trades opened in last 12 months  | 0.311938
                 No of PL trades opened in last 6 months  | 0.234034
                No of PL trades opened in last 12 months  | 0.269245
    No of Inquiries in last 6 months (excluding ho...  | 0.116103
    No of Inquiries in last 12 months (excluding h...  | 0.258468
                            Presence of open home loan  | 0.017455
                                     Outstanding Balance  | 0.256517
                                      Total No of Trades  | 0.249165
                            Presence of open auto loan  | 0.001641
```

Model Building

Predictive model building approach will include the following:

- Data Split: The final dataset is split into Train and Test in 70:30 ratio for model building.
- The cutoff value for the probability of default will be chosen such that model evaluation metrics like accuracy, sensitivity and specificity are almost equal.
- Logistic regression, Decision tree and Random Forest Algorithm will be built once on Demographic and once on merged data of Demographic and Credit Bureau.
- The rejected population will be used to assess model performance.
- Metrics such as accuracy, sensitivity and specificity will be used to evaluate the models.

Model Evaluation Metrics

After building and evaluating multiple models, we find that logistic regression is performing better than Random Forest or Decision Tree. We recommend to use the logistic regression model as the final model to prepare the application scorecard.

Models	Accuracy	Sensitivity	Specificity
Logistic Regression on Demographic dataset	95.8%	55.1%	58.18%
Decision Tree on Demographic dataset	95.75%	0%	1%
Random Forest on Demographic dataset	95.14%	0%	1%
Logistic Regression on Merged dataset (with rejected 1425 records)	91.33%	70%	55.5%
Decision Tree on Merged dataset (with rejected 1425 records)	95.75%	0%	1%
Random Forest on Merged dataset (with rejected 1425 records)	94%	4%	97%

Application scorecard – Our Approach

- An application scorecard will be built using the chosen model and a suitable score cut-off will be identified.
- The scores of the rejected population will be compared with those of the approved ones.
- Based on the scorecard, financial benefit assessment will be performed and presented. This assessment will include assumptions, incremental benefit in terms of credit loss avoided and potential loss of revenue due to rejection of good customer.

Application scorecard

The application scorecard has been built using the final model of logistic regression

The logistic regression model was chosen since its evaluation metrics were slightly better to other models as well it’s an easily interpretable simple model.

Score = Offset + (Factor * log(odds))

- Higher scores indicate less risk for defaulting
- CUTOFF SCORE is equal to 360.95

Application scorecard – Rejected population

- Average Score of Rejected Application is less than that of approved Application
- Rejections by Bank:1425
- Rejections correctly predicted by Model:451
- 40% of correct rejections are predicted by the model

Financial Analysis

There were 253 customers which were wrongly identified as Default which led to 253*x loss of revenue

There were 8788 risky customers which may have identified as defaulters if current model was used. So 8788*x amount was exposed to risky customers. This may have been saved using current model.

Key Takeaways

- Logistic regression model is chosen as the final Model.
- Optimal score cut-off value of 360.95 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.

Thank you!