Credit Risk Analytics Case Study

- Rushi B
- Suhas Mp
- Raghvendra B
- Shankar V

ABSTRACT:

- This case study is aimed at identify the right customers using predictive models and to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit

BUSINESS OBJECTIVE:

- Identify relevant predictor variables and predict the likelihood of default for the rejected candidates and assess whether the results correspond to your expectations. a response using EDA.
- Identify the most important variables affecting likelihood of default
- Build predictive models and choose the best one.
- Build an application scorecard to identify the cut-off score below which business would not grant credit cards to applicants.
- Estimate the potential financial benefits of using the models for auto-approval of credit cards

Credit Risk Analysis – Work Flow

- The analysis is divided into 5 parts:
- Data Understanding – Demographic and Credit bureau information
- Data cleaning and preparation
- Identifying important variables by WOE and IV and gather important insights by Exploratory Data Analysis
- Predictive modeling
  - Modeling on demographic data and Credit bureau information individually
  - Modeling on combined data of demographic and credit bureau variables
- Application scorecard
  - Identifying the optimal score for rejecting the applicant
- Financial Benefits
  - Assessing the potential benefits of using predictive models for auto-approval

Data Understanding – Demographic and Credit Bureau Data

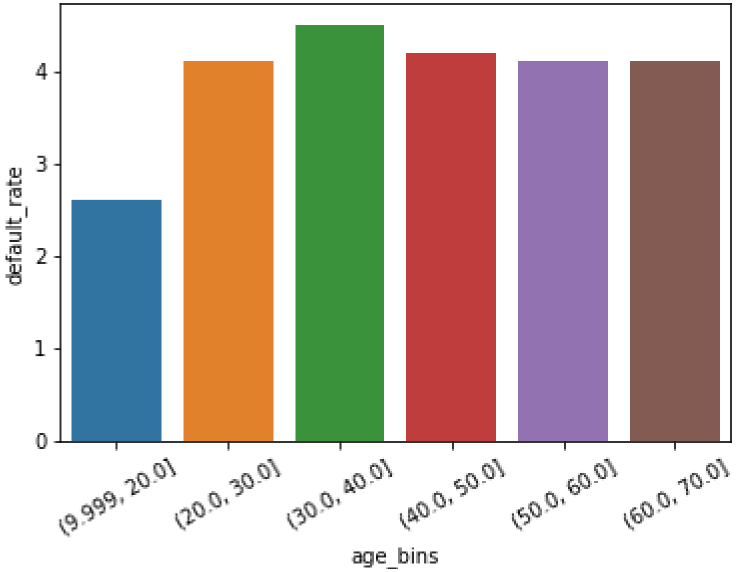
- The data considered for analysis is both demographic data which indicates the customer details and credit bureau data which gives customers behavior and cash flow
- Below are the some of the variables that inherently gives the predictive power to defaulters

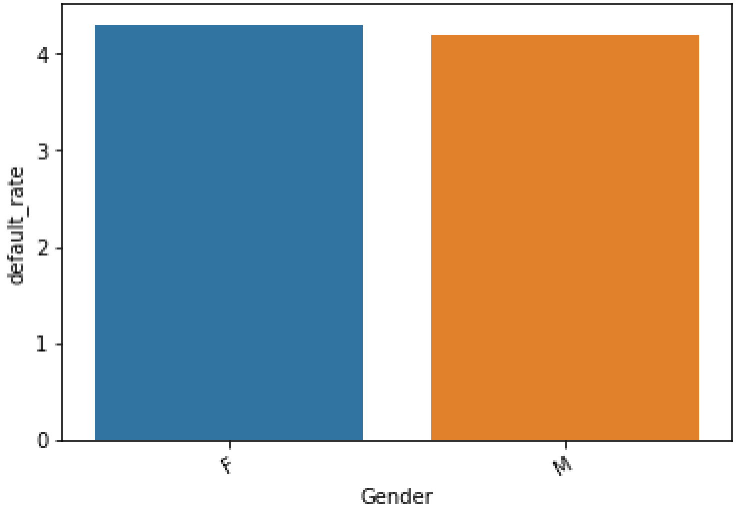
Demographic Data
Age
Income
Gender
Marital Status
Education

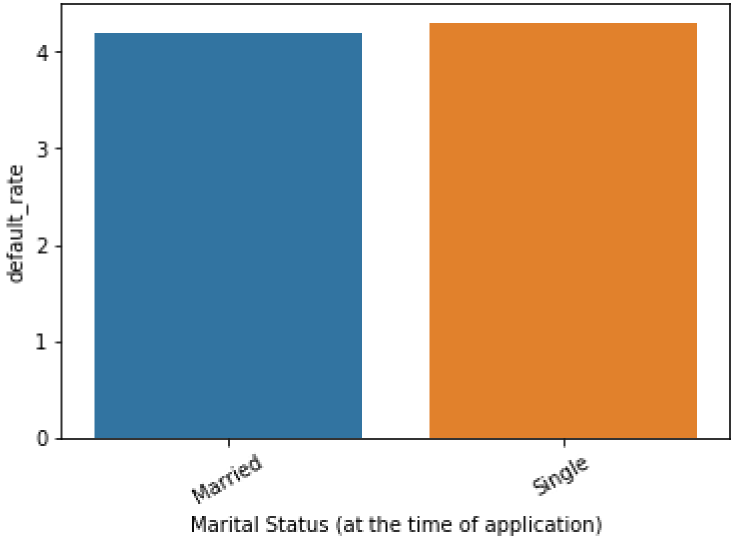
Credit Bureau Data
Outstanding balance
30,60,90 DPD (Past Dues)
Total trades
Number of inquiries
Presence of home or Auto loan

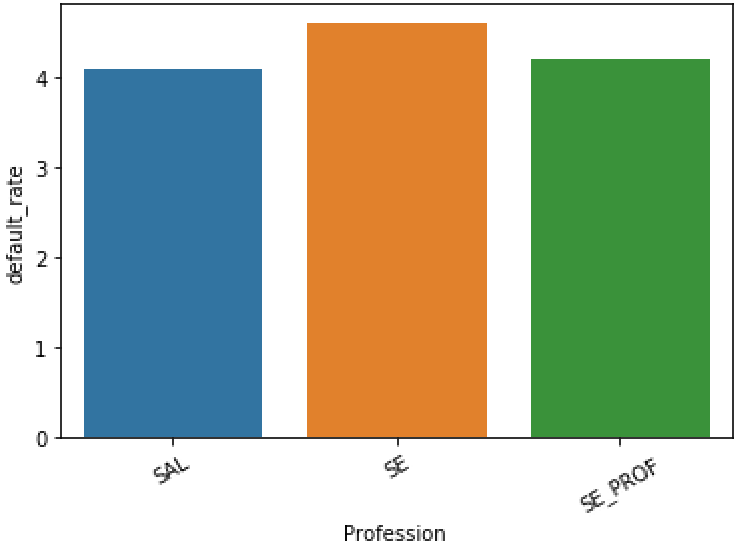
EDA Insights:

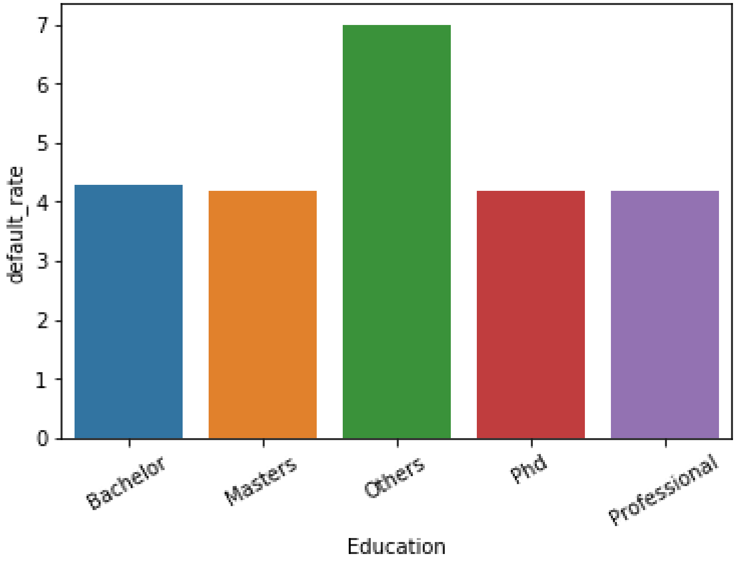
- The average default rate is 4% by all its customers in which credX has granted creditCard.

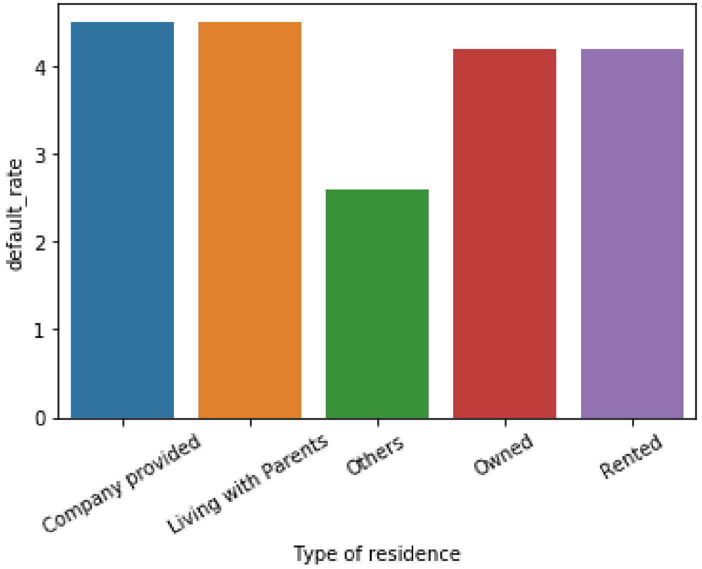






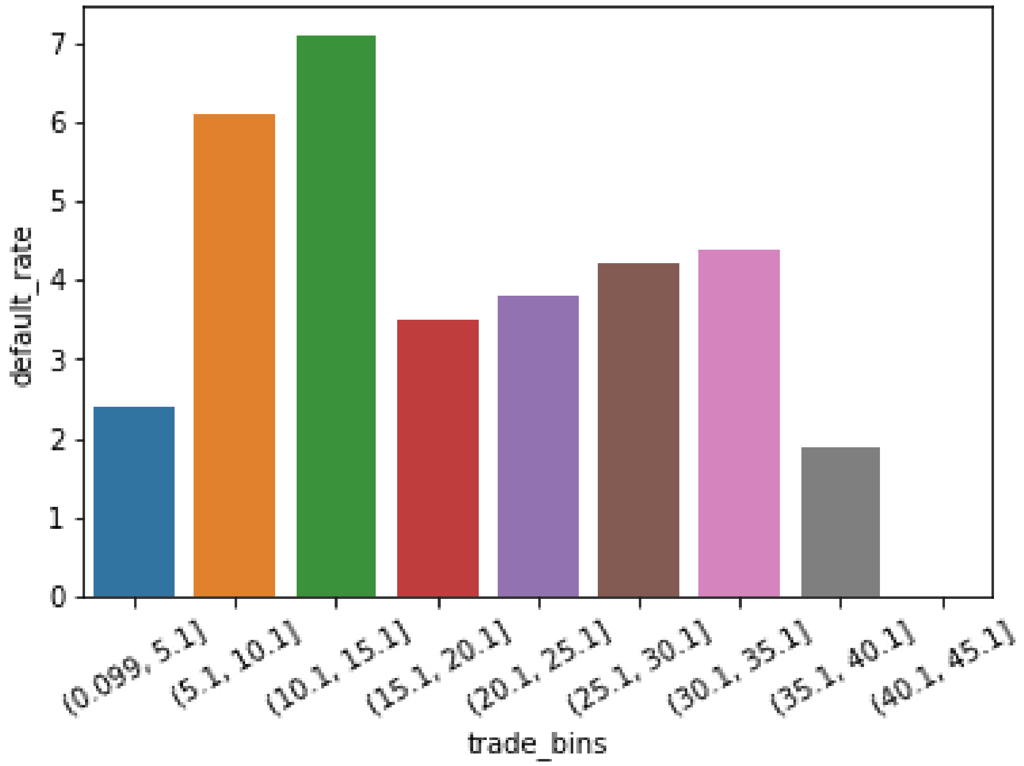


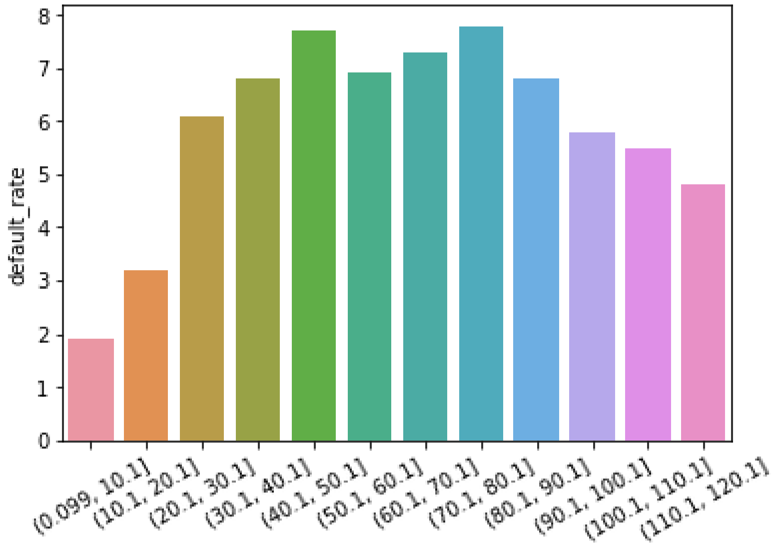




EDA Insights:

- Customers who are having home loans, high average credit utilization and the people who trade between 5-15 times have high default rate which is approximately 6.6% default rate

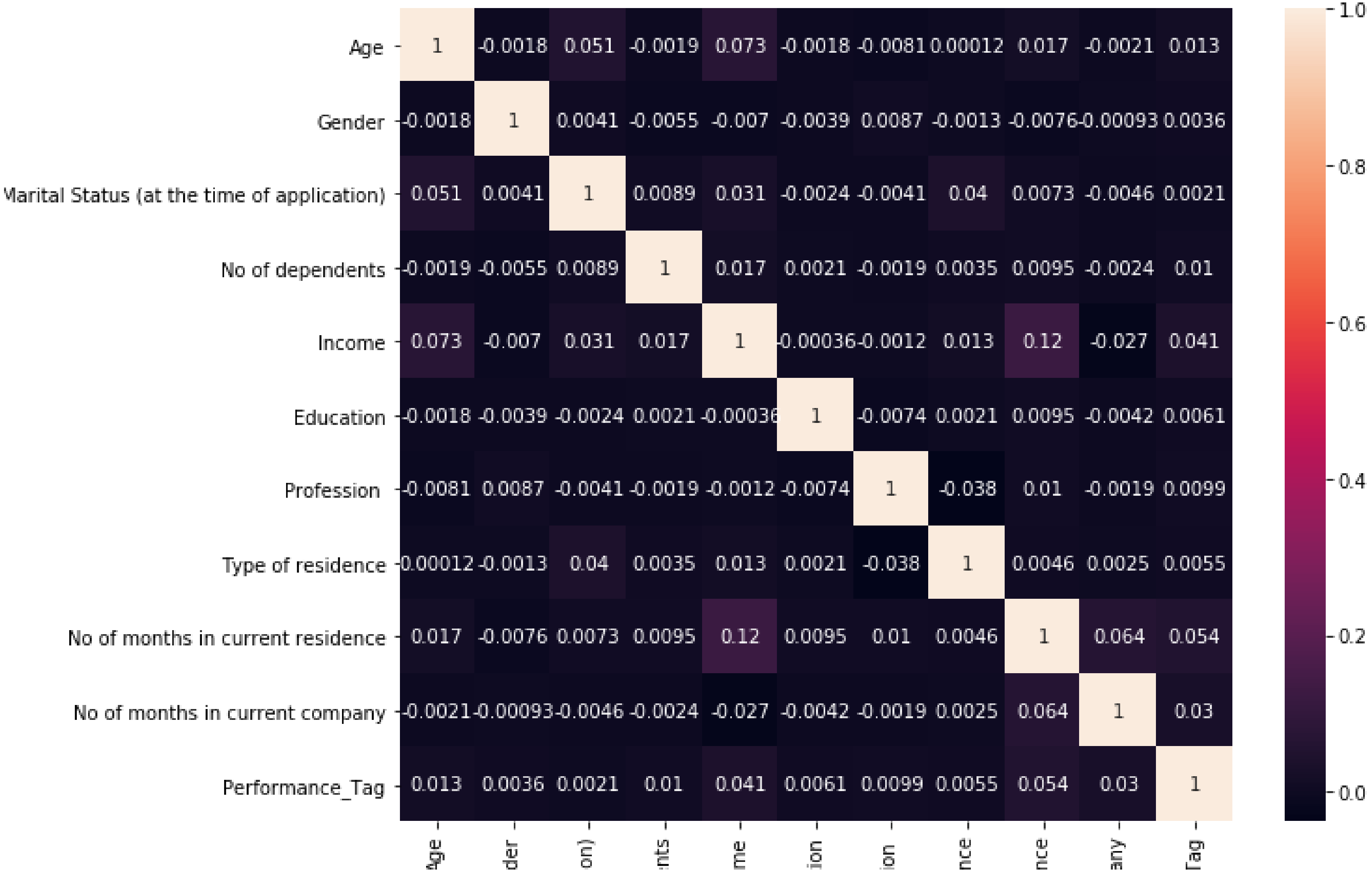




income_bins	No_of_customers	count_defaulters	default_rate
(0.999, 11.0]	13817	747	5.4
(11.0, 21.0]	13214	610	4.6
(21.0, 31.0]	13425	588	4.4
(31.0, 41.0]	13334	464	3.5
(41.0, 51.0]	9895	344	3.5
(51.0, 61.0]	4916	141	2.9

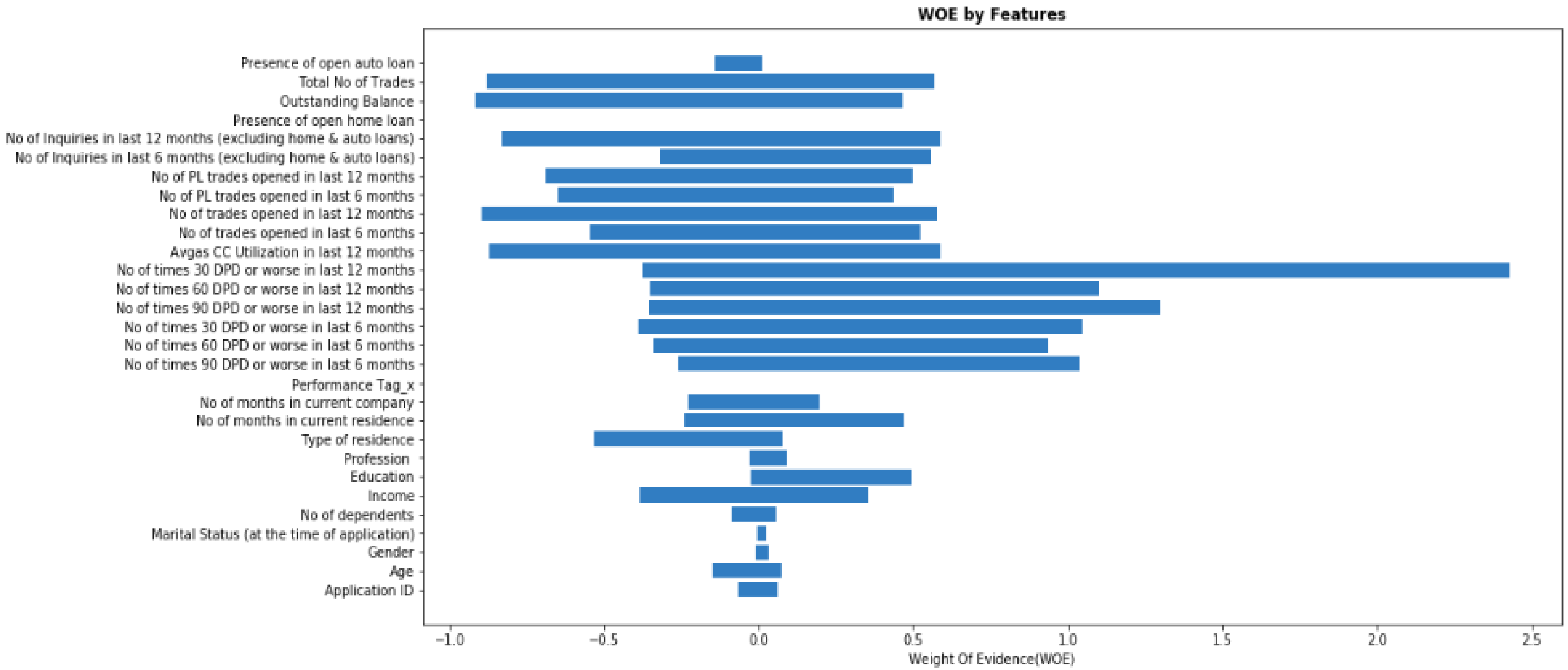
Correlation matrix:

- Correlation between various fields of the demographic and credit bureau data

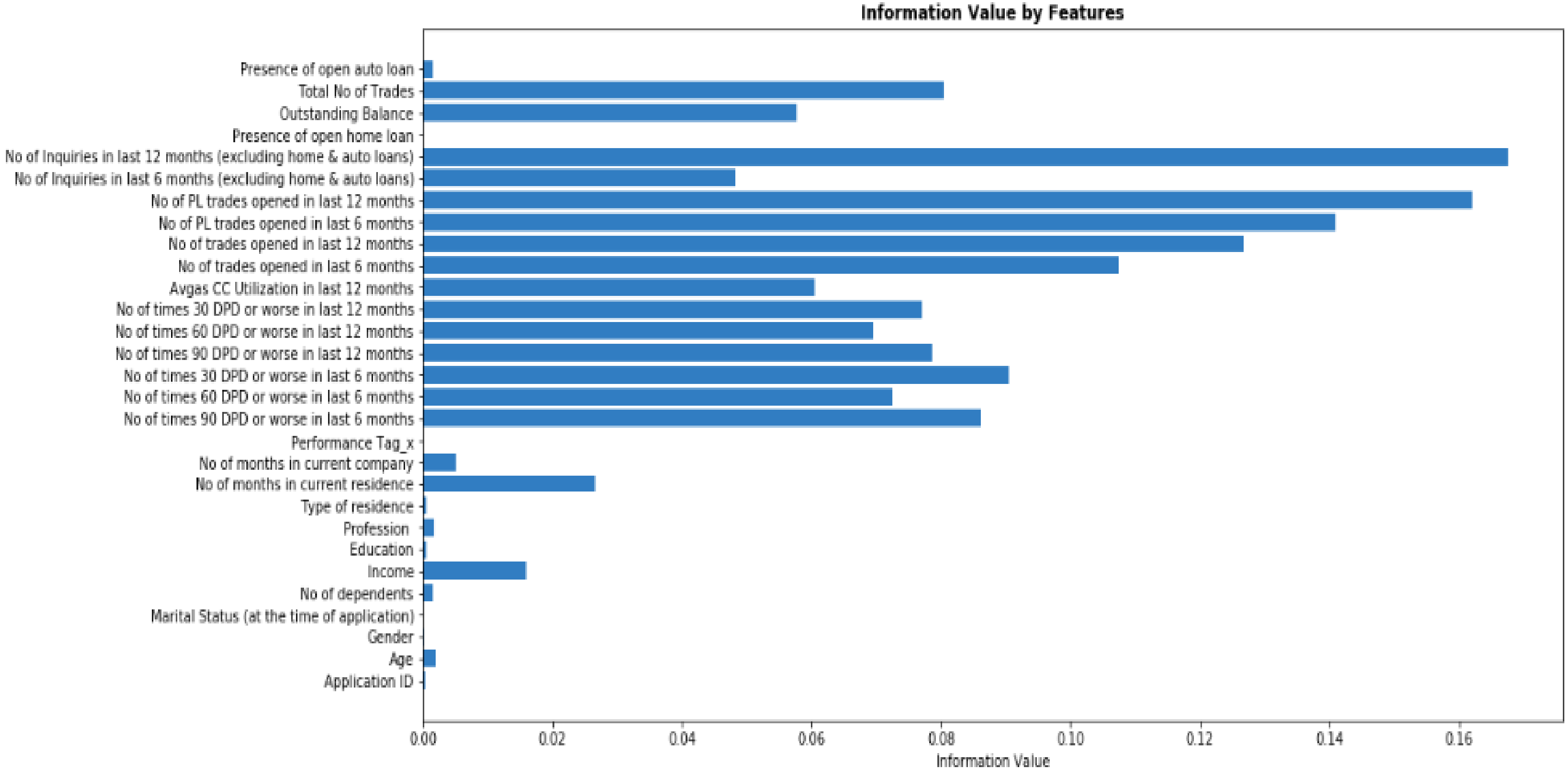


Weight of Evidence:

- **Weight of evidence** (WOE) is a measure of how much the **evidence** supports or undermines a hypothesis.
- WOE measures the relative risk of an attribute of binning level.



Information value by features:



Model building:

- Identifying all the suitable variables by PCA, RFE
- Since we are predicting whether a customer defaults or not which is a categorical type below modeling techniques can be used:

  1. Logistic Regression
  2. Decision Trees
  3. Random Forest

- As part of model validation, predict the likelihood of default for the rejected candidates and assess whether the results correspond to the business expectations.
- Tuning the hyper parameters based on the results of model evaluation score cards and confusion matrix that has high specificity and recall

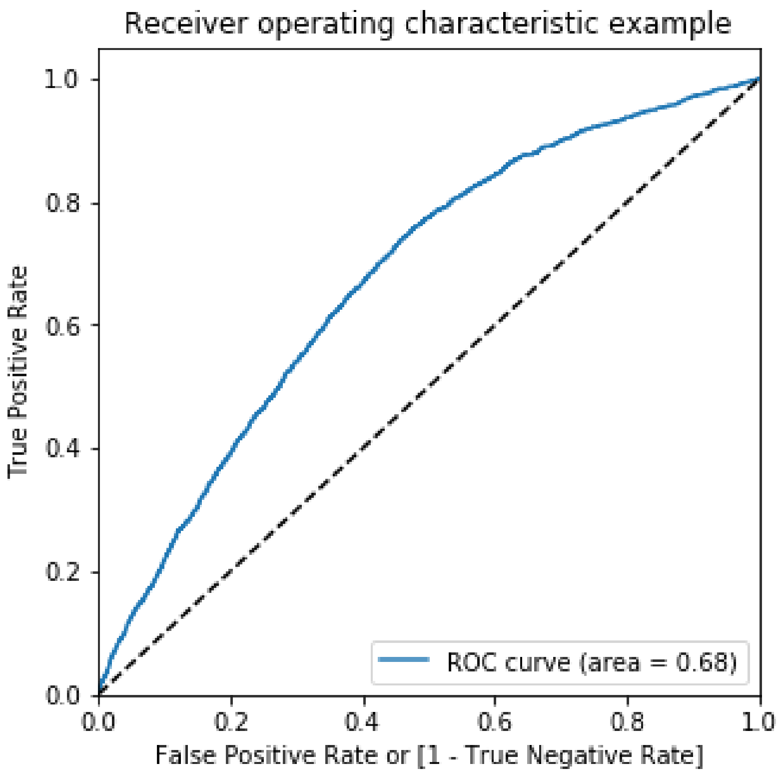
Logistic Regression Results:

- Below are the important predictor variables that are determined by Logistic Regression

- **Confusion Matrix:**

```
[[28277 18455]
 [  701  1375]]
```

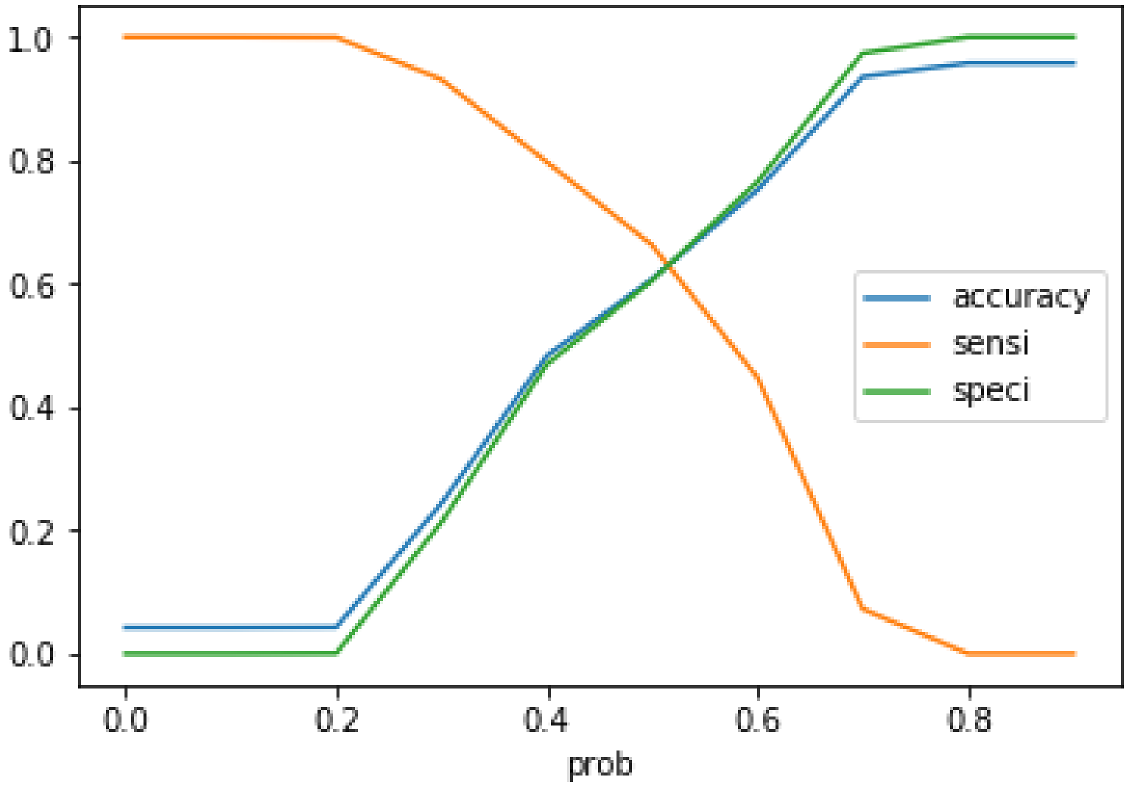
	coef	std err	z	P>|z|	[0.025	0.975]
Age	0.6923	0.146	4.733	0.000	0.406	0.979
Gender	1.0942	0.539	2.030	0.042	0.037	2.151
No of dependents	0.8045	0.182	4.411	0.000	0.447	1.162
Profession	0.9299	0.204	4.560	0.000	0.530	1.330
Type of residence	0.8729	0.295	2.963	0.003	0.296	1.450
No of months in current company	0.2828	0.065	4.369	0.000	0.156	0.410
No of times 30 DPD or worse in last 6 months	0.4012	0.026	15.284	0.000	0.350	0.453
Avgas CC Utilization in last 12 months	0.6112	0.022	27.416	0.000	0.568	0.655
No of Inquiries in last 12 months (excluding home & auto loans)	0.3243	0.022	14.691	0.000	0.281	0.368



Model Evaluation:

- Plots and values of specificity, sensitivity and accuracy for different probabilities

prob	accuracy	sensi	speci	preci	recall
0.0	0.042534	1.000000	0.000000	0.042534	1.000000
0.1	0.042534	1.000000	0.000000	0.042534	1.000000
0.2	0.042985	0.999518	0.000492	0.042534	0.999518
0.3	0.243567	0.931118	0.213023	0.049935	0.931118
0.4	0.482954	0.797206	0.468993	0.062524	0.797206
0.5	0.607523	0.662331	0.605089	0.069339	0.662331
0.6	0.751619	0.447977	0.765107	0.078105	0.447977
0.7	0.935461	0.072736	0.973787	0.109738	0.072736



Application Score Card:

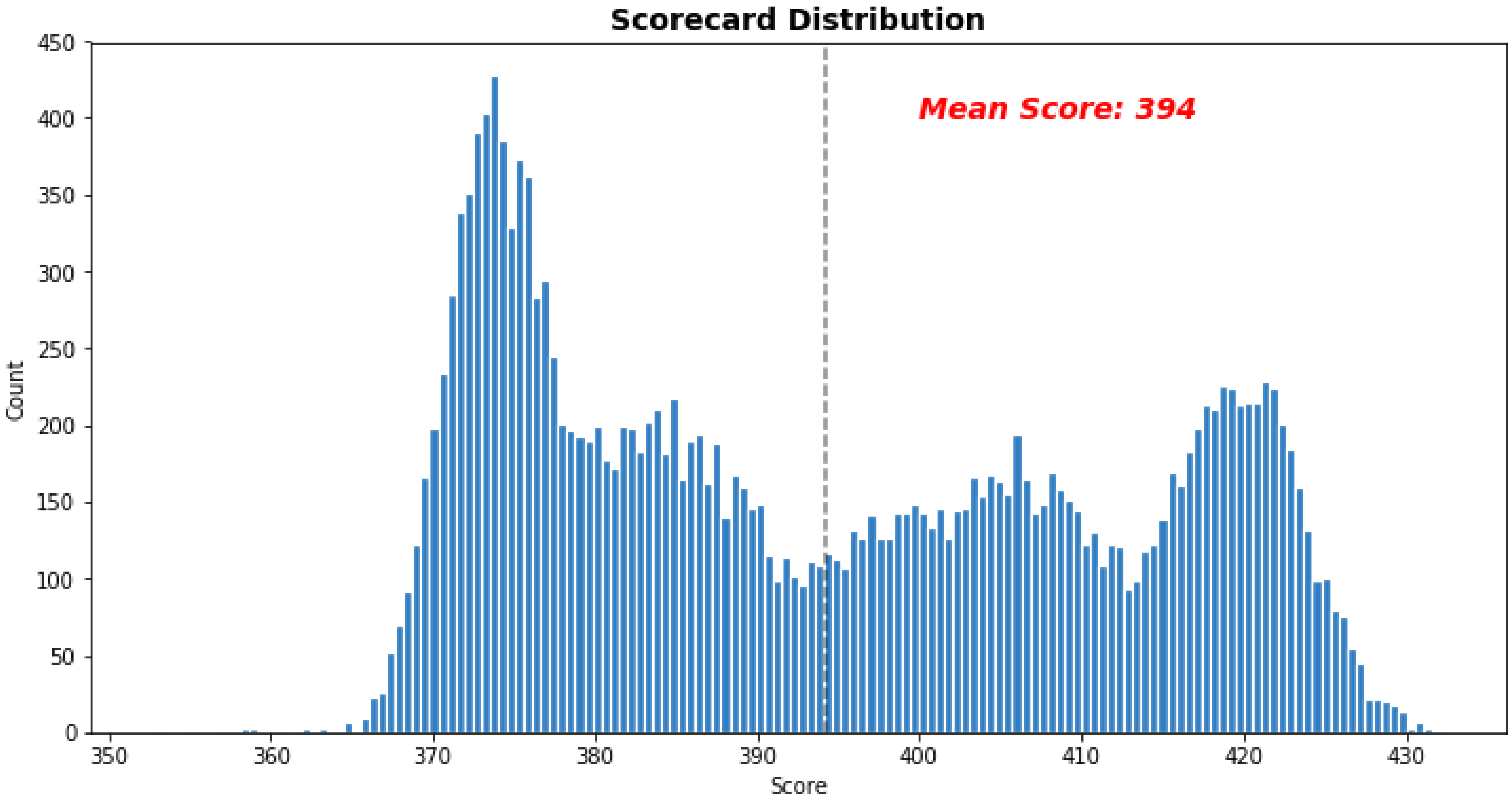
Score is calculated using the following expression:

$$Score = 400 + \left(20 * \frac{log(odds)}{log(2)}\right)$$

- To build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.
- For the rejected population, calculate the application scores and assess the results. Compare the scores of the rejected population with the approved candidates and comment on the observations.
- On the basis of the scorecard, identify the cut-off score below which you would not grant credit cards to applicants.

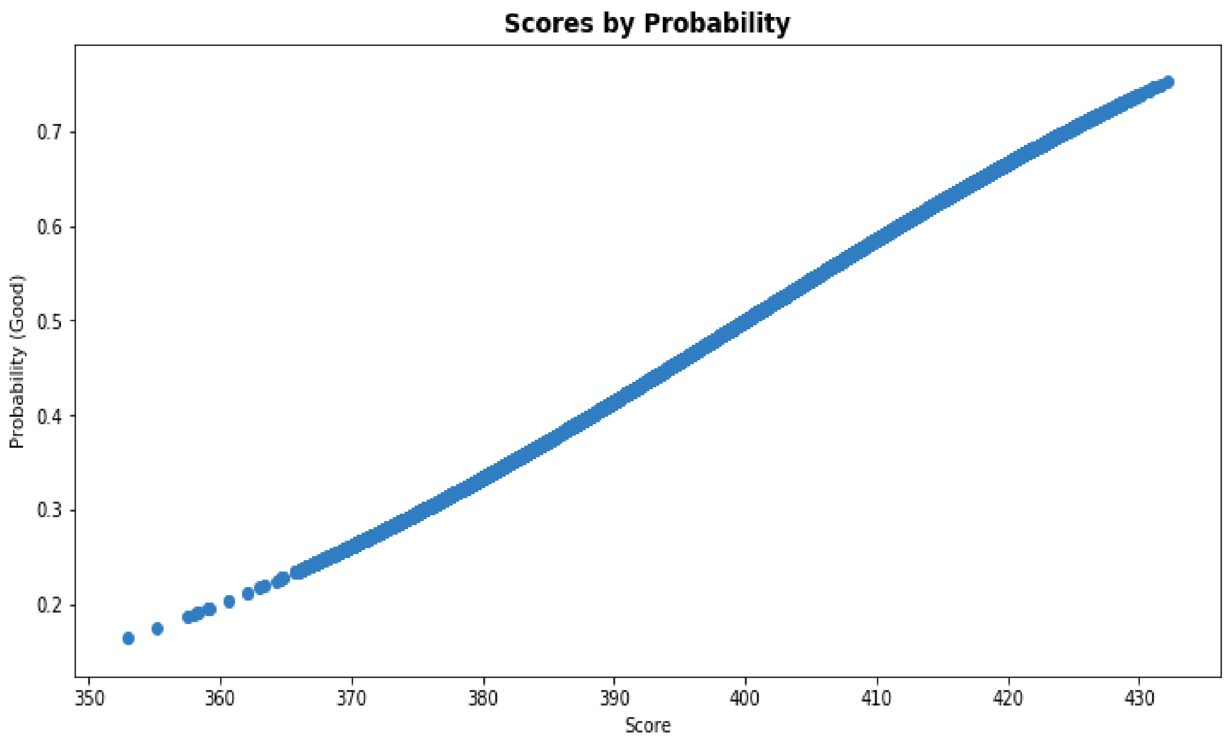
Score Card:

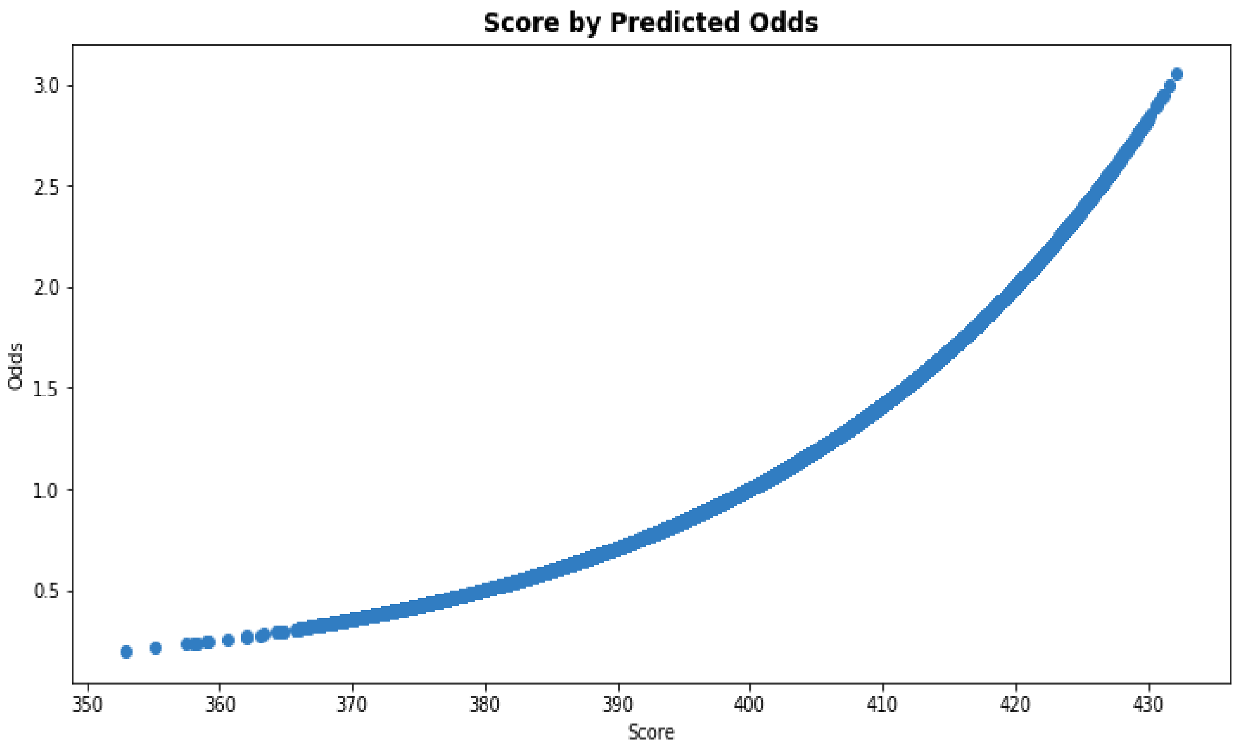
- Mean score is 394 for the combined demographic and credit bureau data



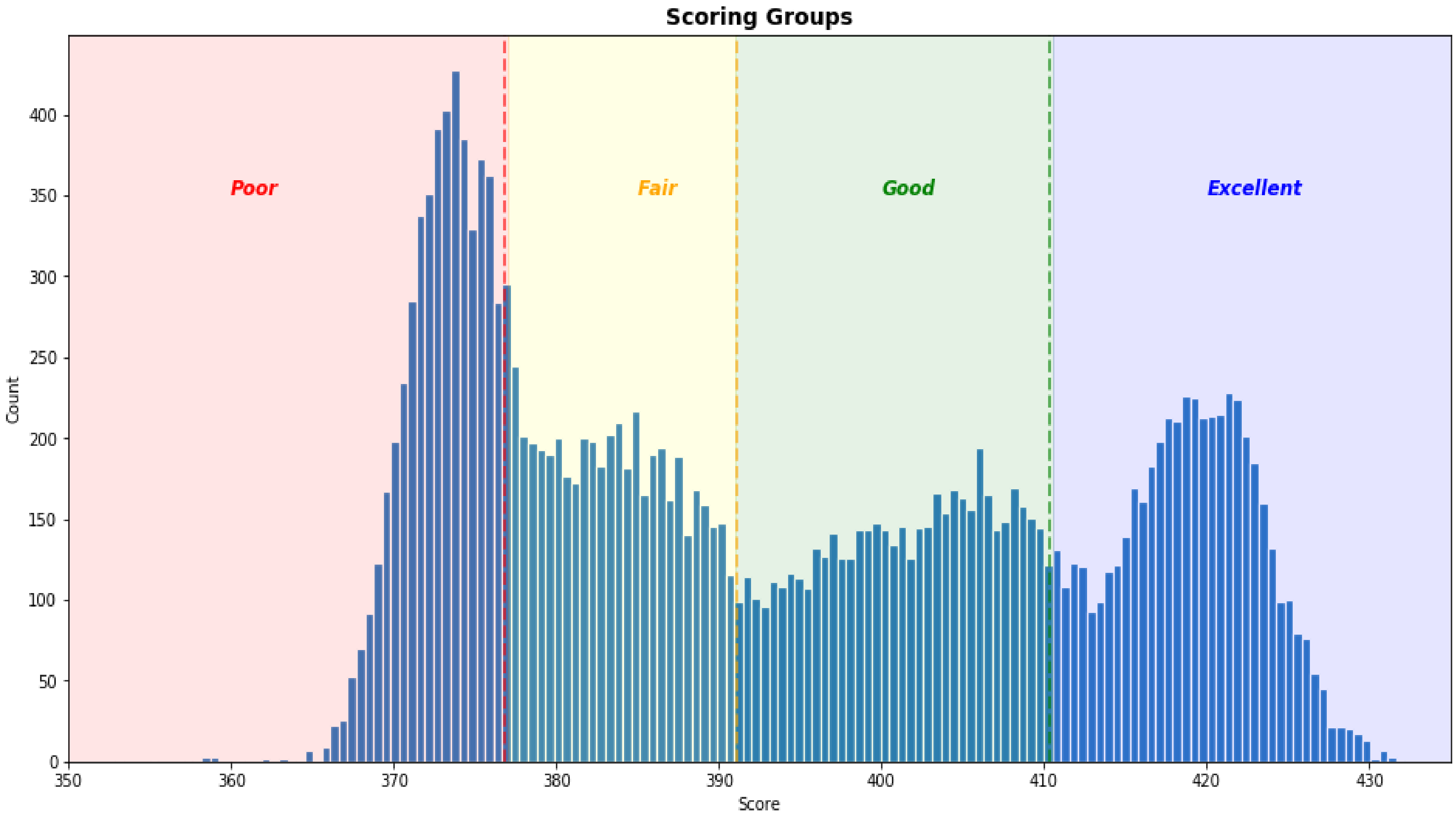
Score Plots:

- Plots to show the nature of probability determined and the predicted odds





Cut Off Scores



- Scoring groups

Financial Reporting:

- Age, gender, number of dependents, profession, type of residence, number of months in current company, No. of times 30 DPD or worst in last 6 months, average credit utilization in last 12 months and number of inquires in last 12 months are some of the variables which business has to focus to get the value.
- Customers whose credit score is less than 377 can be truly rejected, whose score is between 377 – 391 can be re-considered, 391 -411 are good and reliable and above 411 are assets for the company.

- Assumptions based on which the model was built :
  - Since the data obtained has 4% default rate, we should be performing SMOTE analysis to make it a balanced data to perform the regression and statistical analysis.
  - Since the data is updated with WOE values on which analysis is done, we must be careful in specifying the bin to focus the actual prediction

THANK YOU