Exploratory Data Analysis

Credit One has seen an increase in the number of customers who have defaulted. As a result, they are losing business. To mitigate the damage, we were given the task to solve this problem promptly. Using the historical records for customer payments and defaults in addition to customer demographics, our team will seek to understand whether a customer is likely to default.

In this work, we imported the data in csv format. We reviewed the descriptive statistics such as the minimum, maximum, and mean value of each attribute. Moreover, we also checked if there were any missing values (of which there were none) and duplicate rows. We also discretized age into 6 bins which were (21 - 30), (31-40), (41-50), (51-60), (61-70) , (71-79). After cleaning and preparing the dataset, we explored further by visualization of the dataset. We created various histograms, line plots, scatter plots, box plots and distribution plots. We created various histograms, line plots, scatter plots, box plots and distribution plots. We also created a correlation matrix, observed highly correlated attributes and removed redundant features.

* What did we learn anything of potential business value from this analysis?

We learned that customer demographics could play an important role in customer behavior including their payment habits. Concrete information derived from the data:

* Credit One has more female customers than male customers.
* Majority of the customers (82%) have college education, where 36% went to graduate school.
* Majority of customers are in their 20’s and 30’s. The numbers decrease by age.
* Less than one-quarter of the customers tend to be in default.
* Customers in their 70’s have the highest credit limit, followed by those in their 60’s and 30s, but credit limit has little to do with customer default.
* Interesting data analysis on individuals who defaulted/non-default:
* Potential defaulters show higher frequency of lower limit of amount of credit limit.
* The more education, the less likely a customer is to default.
* Married people more likely to default than singles, but the difference is not that large. Also, 'Others' are more likely to default than marrieds, weakening the correlation.

We cannot control customer spending habits, nor can we change customers’ spending or payment habits. EDA gave us some characteristics of the current customer base. We can apply predictive analytics using the customer demographics and historical payment & default data to determine the type of customers we should extend credit. Based on our findings, we can reduce loan limits to existing customers that are more at risk of default and add low risk customers to expand business. Recommendations regarding our findings is to apply variable importance to see customer characteristics that are statistically significant and an important feature in order to predict whether a potential customer will default. Lastly, we can set up a “customer test group” to evaluate our findings and prediction methods, and track improvement in loan defaults rates in the next 6 months to one year.