**Exploratory Data Analysis (EDA) Report**

Credit One has seen an increase in the number of customers who have defaulted on loans. As the credit scoring service, we could lose business if the problem is not solved right away. Using the historical records for customer payments and defaults in addition to customer demographics, Data Analytics team at Credit One seeks to understand whether or not a customer is likely to default on their credit obligations.

This report is intended for the Data Science team. We started our project by defining the business objectives and translating those objectives into goals for the data analytics team. In this task, we fully explore the historical data. The EDA process should be applicable to future projects.

In this work, we imported data in csv format, checked out the number of columns and rows, column names, viewed descriptive statistics such as minimum, maximum, mean values of each attribute, checked for missing values and checked for duplicate rows. These steps were part of the data preparation, data cleaning and transformation tasks. We discretized customer’s age into 6 bins to indicate ages of 20’s, 30’s through 70’s. We replaced ‘Age’ data with an ‘Age Bin’ column. 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 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.
* What are the main lessons we have learned from this experience?
* There is a lot of Python code with examples on data cleaning, data evaluation and visualization. However, there is generally no clear indication of Python version or library version in those references. For example, ‘pylab’ statement in Titanic code would result in errors unless corrected for the latest matplot library and Python version.
* Data visualization is essential to uncover information from any dataset. However, not all plots or charts are applicable to a given dataset. For example, factor plots or linear plots (lmplot) might show valuable insight into Titanic dataset, but not for our dataset.
* Combination and permutation of the attributes do not lend concrete conclusions on default rates based on 2-dimansional visualization.
* What proven methods can we use to uncover more information?
* 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
* Recommendations regarding our findings:
* Apply the equivalent of “variable importance” used in R code to see customer characteristics that are statistically significant and relatively more important in order to predict whether or not a customer will make a payment or be in default next month.
* Define the metrics to follow in predicting customer payment habits, and apply models to data at hand.
* 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.