**Exploratory Data Analysis**

**Credit One**

**Marko Gorenc**

The exploratory data analysis exercise for Credit One covers 24 features which can be grouped into five principle main categories:

* Group 1 (features 1-5) - demographic data for each customer (curent account balance, gender, education, age, and marital status
* Group 2 (features 6-11) - historical payment information of each customer in months counting back from September, 2005.
* Group 3 (features 12-17) - historical record of the amount of bill statement covering the same time period as above.
* Group 4 (features 18- 23) -amount of the previous payment remitted from the customer.
* The final feature (24) describes whether or not the customer has defaulted on his/her loan.

The task was to perform exploratory data analysis on the data provided to understand those most likely to default on a loan. By understanding the general demographic groups and higher risk customers, Credit One could make informed decisions when making loans to individuals deemed as high risk. For this EDA work, we focused on identifying high risk individuals by demographic data because we’re focused on predicting a new customer’s ability to pay off their loan. Although historical payment, bill amount and previous payment amount are useful in understanding an individual’s likelihood of default, this information is collected for existing customers with existing loans.

The process of EDA is centered around establishing and understanding cursory relationships in the data. EDA dominantly utilizes charts and plots to visually analyze for relationships. Additionally, numerical relationships are established through the generation of correlation and covariance matrices and heat maps. The key findings from this EDA are that there are no resoundingly clear relationships that can be drawn between demographic information and those likely to default on a loan from Credit One. However, we do make a few key observations about those who are defaulting as an overrepresentation of their population, relative to their proportion of the total population. These individuals appear to be dominantly: (1) men, and/or (2) single, and/or (3) have no more than a high school education and/or (4) older (45-79) and younger (<30). These basic observations are likely pointing to the complexity of identifying high risk individuals and the difficulty in identifying trends in 1D or 2D plots.

My recommendation to the team is to move forward with machine learning. By utilizing ML, we may be able to understand the complex demographics of individuals that are more likely to default so that we may make more informed decisions about who receives future loan money.