Credit One Project
Progress Report

After doing an extensive exploratory data analysis (attached), I have some key observations about the data to share that includes some potential business value and some recommendations:

- The overall current default rate is about 22%
- Bill Amounts and Payment Amounts have very similar distributions within, hence the collinearity between the months. We can assume from this pattern that customers tend to have consistent behaviors.
- Overall, default status does not seem to have a strong relationship with any demographic or variable other than payment history.
- There is a positive correlation between Credit Limit and Education-graduate school. This means that customers with higher degrees are more likely to get higher credit limits.
    - This needs to be examined, is it because people with higher degrees are more likely to have better qualifications for higher credit? Or is it an unsupported anecdotal assumption that people with higher degrees are more likely and/or able to pay back funds?
- The data is showing a negative correlation between credit limit and payment history, meaning that the higher the credit limit, the more likely customers pay the balances.
    - The caveat is, when looking at the scatter plot, larger amounts do not *tend* to "not default". It seems that the large million-dollar credit limit is a major outlier that is clearly affecting the rest of the data. 75% of the data has a Credit limit of up to $240,000.
- When running a mathematical check for outliers (outside 3 standard deviations or outside the 99.7% range), there were no lower outliers. However, there were 130 upper outliers!
- For now, I have removed only the million-dollar credit limit outlier. I do not recommend removing all larger amounts in the meantime. Discretization might solve this issue before running machine learning models.