Team Mulberry: Executive Summary

The Problem and Stakeholders

Our goal was to create a model that could predict who would default on their credit card. This information is especially valuable to credit card companies and credit bureaus, as they could use this information to analyze risk when approving future credit cards, or determine policy for leniency with regards to late payments. The model would also be of use to the United States Treasury, as credit card non-payment has been used as an indicator for upcoming recessions. We used accuracy and recall as our metrics because our stakeholders would be most interested in who would be in danger of defaulting.

The Data

We were given data on 30,000 individuals. Our information included demographic information such as age, education, marital status, and biological sex. We were also given credit history, including payment status, payment amount, and bill amount over 6 months, as well as the individuals' credit limits. We used target encoding to process the categorical information.
We also created additional features from these that we thought might correlate to defaulting.

Approaches

After splitting the data between training and validation sets so that it was balanced between those who defaulted and those who did not, we created the following models:

- A nearest neighbors model.
- A random forest model.
- A logistic regression model. We created models on each feature separately, and found that only the models that used the total number of unpaid months, average payment status, and most recent payment status did not predict that no one would default on their credit card. So, we created a model that only incorporated these 3 features.
- A vector classifier that, like the logistic regression model, worked best when we only used the above stated features.
- A multilayer feed forward neural network.

Using the first four models above, we created a voting model which assumed that if the votes were tied, the person defaulted.

Final Model

The voting model had the best metrics, with an accuracy of 82.03% and a recall of 37.80% on the validation data.
On the training data, we had an accuracy of 82.00% and a recall of 38.15%, which is very similar.

Conclusion and Recommendations

Compared to random guessing, our model better predicts credit card defaults. It could be used, in conjunction with other factors, to make policy decisions.