Machine Learning for Public Policy
HW2 Report

This report looks at data relating to the delinquency rates of an individual. It attempts to find the best predictors to determine the chance that an individual will experience financial distress in the next two years.

Methodology:

The following methodology was used to develop the model:

1. Data Exploration and Visualization

At this stage, I explored the data in order to find trends, and extreme values in the data. On average, the number of credit lines and loans was 8, with certain extreme cases. There were approximately 15 individuals who had over 50 open credit and loan lines. Number of dependents was mostly concentrated around 0 or 1, but with some outliers at 20 dependents. The Revolving Utilization of Unsecured Lines, which is the total balance on credit card, and personal lines of credit except real estate and not installment debt (e.g. car loans divided by the sum of credit limits) also contained some surprising values; namely, there was one individual in particular with 50708 lines of unsecured lines. This is possibly an extreme case, or a human error mistake in data entry. I also found that there wasn't a significant difference amongst zip codes when it comes to income, number of people who are delinquent in two years, age, or debt ratio.

Grouping by those individuals who are likely to be delinquent in two years, we find that there is an income gap. Those not delinquent had an average monthly income of 6748, and those that were had an average income of 5631. The gap persisted when we looked at the median, which singles that the gap was not caused by some outlier values. Similarly, debt ratio was also different amongst those who were delinquent and those who were not with the former group have a higher debt ratio than the latter. In addition, those who were delinquent were on average 2 times 90 days late on payments. What we found, also, was that those who had around 6 or 7 dependents were more likely to be delinquent.

Lastly, what we found from looking at the correlation matrix, which calculates the correlation between each pairs of variables (i.e. how much or how little variables covary with each other), found, not surprisingly, that Number of times an individual was 30-50 days past their due date, was highly correlation with an individual being 60-89 days late, and 90 days late. This will have implications for the variables we choose to include in our model.

2. Pre-processing data.

At this stage, we deal with missing values in the data, and how best to replace them. From what we can see, there are only two variables with missing values: monthly income and Number of dependents. Given that the average number of dependents in the data was 0.75, we will round up to 1, and replace each missing value entry with 1 for the number of dependents.

Monthly income also had missing values. The mean for monthly income was 6670, whereas the median was 5400. Indeed, by looking at a histogram of monthly income, we can see that it is heavily skewed; therefore, using the mean would be inappropriate as outliers heavily influence it. As such, we will replace the missing values with the mean.

3. Building the Model

Given what we learned from the correlation matrix, we decided to remove Number of Time an individual was 90 days late, and the number times an individual was 60-89 days late in order to avoid a multicolinearity problem in our model. In addition, we remove personID as that is a useless predictor variable.

Results

I find that the best trained model has an accuracy of 0.9348, and the best predictor features are age, number of times an individual was 30-59 days late, the number of real estate loans or lines, and the number of dependents.