Benjamin Spiegel

Credit Card Default Prediction

Looking at the heatmap of the dataframe’s we can see the features that have the greatest correlation to whether or not an individual will default next month. I chose to use a heatmap because there are over 30 features and this allowed me to easily see correlations for all features at once.

Of the initial features, ‘PAY\_0’ was the most predictive. The feature shows whether or not an individual made the minimum payment on his/ her very last bill. As would be expected the most recent behavior of individual is be most predictive of the sequential action.

From the original data ‘Limit\_BAL’ was also predictive. This feature tells the credit limit that the credit card company has designated to each individual. Credit card companies give higher credit limits to individuals who are most capable of paying their debt and have a positive credit history. As would be expected people with higher limits are less likely to default in the dataset.

While these features performed well in predicting default. I also created features which combined different features. A couple of these features were shown by the heatmap and other graphs to be quite predictive of whether customers would default. Of the features that I created, the most predictive were ‘ever\_default’ and ‘aggregate\_default’. Ever default simply shows weather a consumer has ever defaulted in the past. In order create this column I used the following code:

lister = []

for row in np.array(paydf):

value = np.any(row>0, axis = 0)

lister.append(value)

I than mapped the output so True became 1 and False was 0 and turned the list into a row.

In order to create a column for the total number of months defaulted per each individual, I used a dictionary and mapped new series’ that made the only possible values for each month a 0(no default) or 1(default). I then added up the total for all months for each individual and now had a total number of months of default per individual as a row. This value had a strong predictive correlation with default as shown by the heatmap, and by a histogram.

Looking at demographic features there were some minor correlations with likelihood of default. These are helpful because they are very independent from financial history and are not repetitive with other features. The demographic information in the dataset included age, gender, education, and marital status.

Running a linear regression between the number of months that individuals had defaulted on their payment and age there was little to no useful correlation.

I also used seaborn’s regplot to create a linear regression looking at how difference in credit limits can predict likelihood of default. There were a clear correlation showing that people with higher credit limits are less likely to default. This is not surprising given that factors that determine credit limit are based on individuals ability to pay off debt and history of paying of debt.

Educational level also had a correlation with default. Looking at education you can see certain groups are much more likely to default. However, there is only a moderate difference in likelihood.

In order to analyze the correlation between categorical groups and likelihood of default I used seaborn’s Countplot. I also used groupby with normalized value\_counts . Using these tools it was shown that the probability of defaulting in the next month for individuals with Graduate Degrees, College Degrees, and Highschool Degrees was 19%, 24%, and 25%, respectively. Interestingly people classified as other while a small group had a much lower likelihood of default with a 5-6% percent likely to default. I used the same tools to look at the correlation between gender and the likelihood of default as I did with education. The variability for different gender’s was also small. Males were 24.2% likely to default in the following month, while females were 20.8 percent likely to default in the following month.

Finally using the same tools as the other categorical features I looked at how marriage effects default rate. I found that for the next month, married people had 23.5 % chance of defaulting, while sing individuals had a 20.9 % chance of defaulting. Again similarly to other demographic features marital status was not strongly predictive of default.