**Project 2, Task 2: Report to Credit One**

Lessons Learned through EDA of Credit One Customer Payment Records

Anders Dowd - Spring 2022 Data Analytics Certification Program

**Lesson 1: Not All Features Are Necessary**

When examining the correlation matrix of the raw data set, it becomes immediately apparent that there is high correlation amongst monthly bill amounts and amongst monthly delinquency statuses. This tells us a few things. One, those with high outstanding balances one month are likely to carry a high balance the subsequent month. The pay down rate is, in most cases, a gradual process, so we typically won’t see wild fluctuation between the BILL\_AMT features. Two, if a customer has trouble paying their bill one month, they are likely to have trouble paying their bill in subsequent months. But most importantly for us as data scientists, it means there will be collinearity in our model if we incorporate all of these variables. We will need to calculate a mean amount for the BILL\_AMT features as well as for the PAY features to use instead. The PAY\_AMT columns also demonstrate slight correlation to each other and may necessitate a mean as well. However, these correlations are not nearly as strong and may not create an issue. It may even be possible to use only some of the values and disregard the rest.

**Lesson 2: Bigger Borrowers Are Not More Likely Defaulters**

Higher credit limit does not correlate to a higher rate of default. In fact, there is a slight inverse relationship. Across nearly every demographic and financial backdrop, customers with higher credit limits were less likely to default. The only exception to this rule was in the case of dormant accounts (those with no payments or balances over the past 6 months), where higher credit limits were associated with more defaults. This could suggest that the prior method of granting loans had some successful mechanism for determining who was worthy of higher credit limits. However, because we do not have any data available to us regarding those who had their loan applications denied, we can’t make assumptions about why those who borrowed less were more likely to default.

**Lesson 3: Demographics Don’t Tell the Whole Story**

We can’t determine who will be a good credit customer based solely on demographics like sex, age, educational background, and marital status, but they can clue us in on some likely habits. While there are no clear cut predictors, there are some tendencies associated with different characteristics. The lowest risk characteristics are: single, female, highly educated (grad school), in their 30s. The highest risk are: divorced, male, high school educated only, over 70 years old. A customer checking just one or two of these boxes shouldn’t qualify or disqualify them from a loan, but these features may prove instrumental in the overall success of our algorithm to determine whether a customer is a good loan candidate.

**Lesson 4: Slow and Steady Doesn’t Win the Race**

Lower payment amounts are associated with eventual default. Those who remained in good standing paid on average 75% more per month than those who defaulted. The average outstanding balance of non-defaulters was only about 4% higher than defaulters, but non-defaulters paid nearly 13% of their balance each month versus less than 8% by their defaulting counterparts. While we can’t change a customer’s spending habits, if we have any way of encouraging higher payments we could drive the default rates down. Are customers only paying the minimum payment? If so, can minimums be raised?

**Lesson 5: Delinquency Breeds Default**

There is a clear correlation between delinquent payments and eventual default. Each month, customers are assigned a number based on their delinquency status: -2 = no consumption; -1 = paid in full; 0 = the use of revolving credit; and 1 through 9 = payment delay for the corresponding number of months. Since -2 through 0 all indicate an account in good standing, we can group those into a single value (0). We can then use the sum of these values over the past six months to calculate a mean value for delinquency status.

Knowing how many months delinquent in payment a customer has been in the last 6 months is a very strong predictor of default likelihood. If the customer has 0 recorded delinquent payments, they have only a 12% probability of defaulting. If they have a single delayed payment in the past 6 months, they have a 25% chance of defaulting, which is slightly higher than the current overall rate of default in the dataset (22%). Probabilities only continue to rise from there. If the customer has more than 6 months of delinquency, they have almost a 2 out of 3 chance of defaulting.

**Variables to incorporate into the model:**

Credit Limit

Age

Sex

Education Level

Marital Status

Average Delinquency Status (past 6 months)

Average Bill Amount (past 6 months)

Payment Amounts (past 6 months) - We will try building the model with both an average as well as up to 6 individual payments

Default?