Credit Default Identification

Megan Ball

Objective

Credit One is at risk for losing revenue and customers due to increase in customers defaulting on loans. There are currently no active methods to quantify or predict risk of customer defaults. Our main objective is to answer the following questions:

1. How does Credit One ensure that customers can/will pay their loans?
2. Can Credit One approve customers with high certainty?

Data

The data used in this analysis is historic customer data over a six month time period. Included in the data is age, gender, education, limit balance, payment status, payment value, and bill amounts for each customer. The data was complete with no missing values.

One primary modification to the data for ease of analysis was compacting the levels of education. The original data set included values from 0 to 6. Since, as per the metadata, levels 4, 5, and 6 were listed as "others," they were all grouped together to a single level as 4. Level 0 was also labeled as "other," but in this analysis it was assumed there was an inherent hierarchy to the levels and 0 was kept separate.

Some key observations noted in the data include the following:

- There are nearly 1.5x as many females as males in this data set
- Most of the customers are highly educated, with the top three education levels being graduate school, university and high school
- The vast majority of the customers are either married or single
- The overall default rate for this data set is 22%, so most of the customers did not actually default on their payments

In preparing for the analysis, a total of four data sets were created based on feature engineering. The first set included the original data with only minor modifications to education level. The second set discretized the data for age, bill payments, and payment amounts. The third set also discretized age, but additionally removed highly correlated (> 90%) features. The last set discretized age, but only kept key features as identified by recursive feature elimination.

Analysis

Despite the differences in all of the data sets, there were not many significant performance differences between a single type of model and the data used. The key metrics targeted for modelling were both accuracy and recall, as the objective is for Credit One to be able to determine with high confidence which customers are likely to default. Based on feature importance, the primary indicators found to determine risk for default all related to payment status; in particular, the most important feature was the payment status in regards to the first month of tracking the customers (see figures 1 & 2).

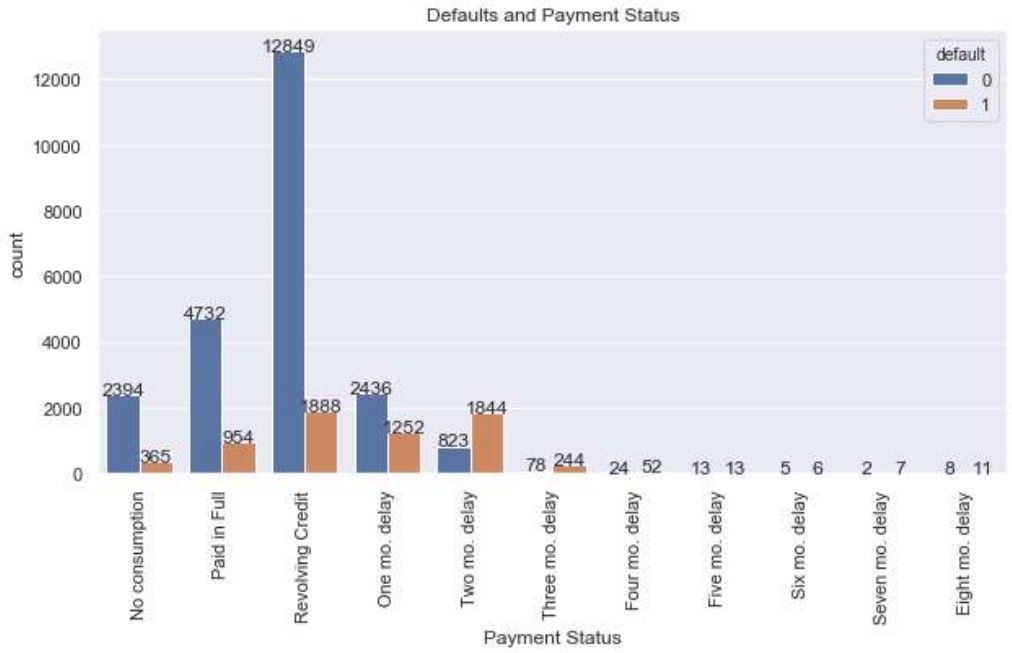

Fig 1. Chart of payment status (if payment is required, and if so, how long is it past due) in relation to default status

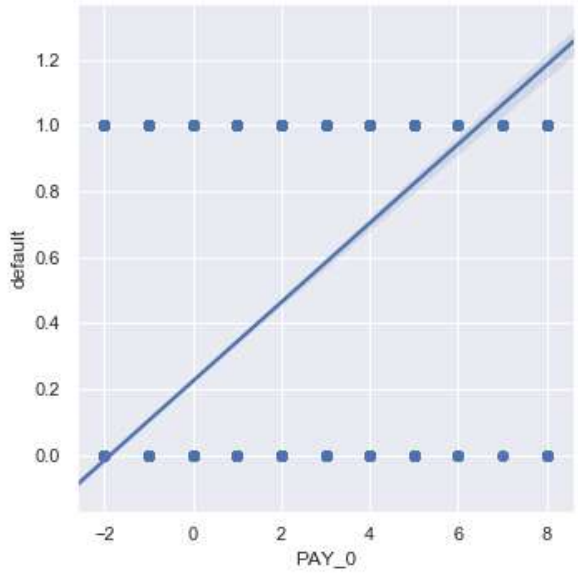

Fig 2. The slope of the line indicates a positive relationship between payment status and default.

The best performing model in regards to both accuracy and recall was the random forest model using the data set with the highly correlated features removed. Based on this model, we can predict with 83% accuracy which customers will default.

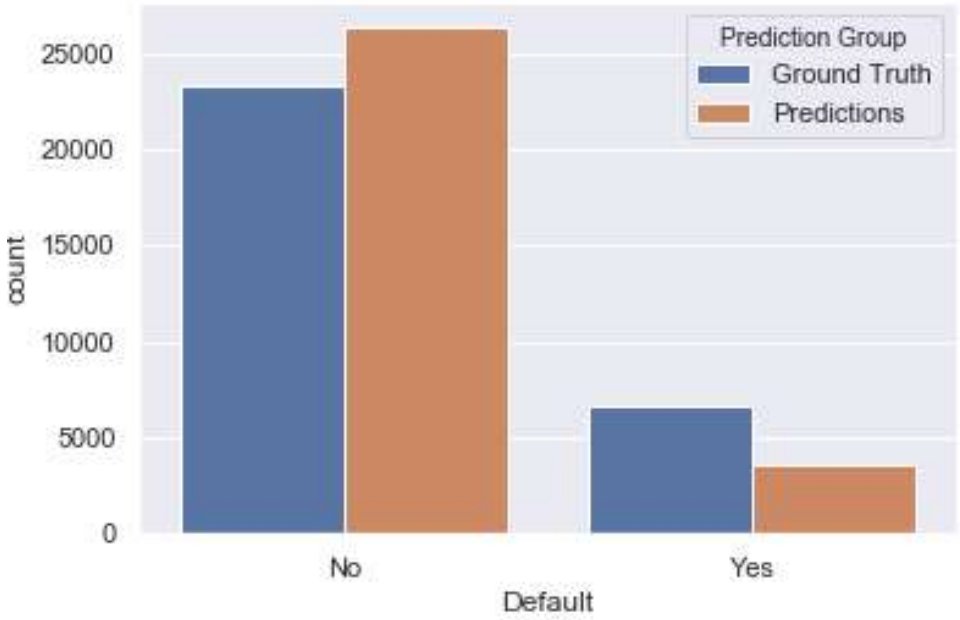

Fig 3. Results of the model predictions compared to actual default values.

Recommendations & Conclusions

1. How does Credit One ensure that customers can/will pay their loans?

There is no singular way to ensure customers will pay their loans as it is difficult to predict human behavior with 100% accuracy. Key factors to review when assessing this risk would be to look at a customer's payment history. If the customer frequently is behind on payments, there is a higher chance that he or she will default. However, there is still a risk of default even with customers that have historically paid on time or used revolving credit. This brings us to our second question:

2. Can Credit One approve customers with high certainty?

Despite the 20% difference between the model's predictions and the actual default values, there is still a fairly high level of certainty (83%) in predicting which customers will default. There is still improvement to be made on this accuracy which could be accomplished with some of the recommendations below.

We can conclude with a reasonable amount of certainty that age, sex, marriage status, and education level do not significantly influence default rates. However, this is dependent on how the data is presented, and we may gain additional insight if the education levels were more accurately defined. For future analysis, it would be beneficial to understand what levels 0, 4, 5, and 6 actually indicate to more accurately represent education in the customer base. One option also could be to lump in college-educated and above as one level, high school as a second, and all levels below high school as a third.

Then, the modeling would be repeated to determine whether or not this feature is significant for the model. If it still proves insignificant, there would be no need to further capture this information.

Additionally, more data is always better in trying to improve accuracy in our models. One downfall for this data set is that the default rates were imbalanced, with far more customers in the group that did not default. If more data is collected over a longer period of time, we may be able to capture more customers that default (with the goal being to get to as close as a 50/50 rate as possible) in addition to being able to see additional payment history. Since the key factor in determining default is payment status, it stands to reason that it could only benefit analysis by collecting this information for a longer time frame.