Customer Default Identification Report - Course 2, Task 3

CreditOne Investigation

Primary Objective - address the following problem:

> CreditOne has seen an increase in customer default rates, which results in the loss of revenue, customers, and clients. Credit One needs a better way to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not.

Business Question:

> Can we use Data Science to better predict the creditworthiness of current and future customers?

Data Set

For this analysis and modeling we used a provided set of 30,000 customer accounts. Within a given account we have significant demographic data (age, gender, education, etc.) and financial data (billing history, payment history, credit limit, etc). Finally, for each account we know whether the account is currently in default.

Exploratory Data Analysis

Given the large number of data fields available with the provided data set, it is important to first explore the variable correlation to highlight which variables may provide the greatest influence on the credit limit and default rate of customers.

To screen for high correlation and covariance between the large number of variables a heatmap is generated, as shown in Figure 1. In this heatmap, box size and color darkness indicate the strength of the correlation. In terms of the credit limit (second horizontal or second vertical column) we see significant levels of correlation to most fields, with weak correlation to customer marital status, age, and gender. To gain insight between customer data and likelihood of default, the bottom two horizontal rows are inspected. As can be seen by the relatively small shapes and light colors, no singular correlation between customer default and the other variables is particularly strong. However, the strongest correlation is with the status code of the customer's most recent payment (PAY1), which provides a good place to start our exploratory data analysis.

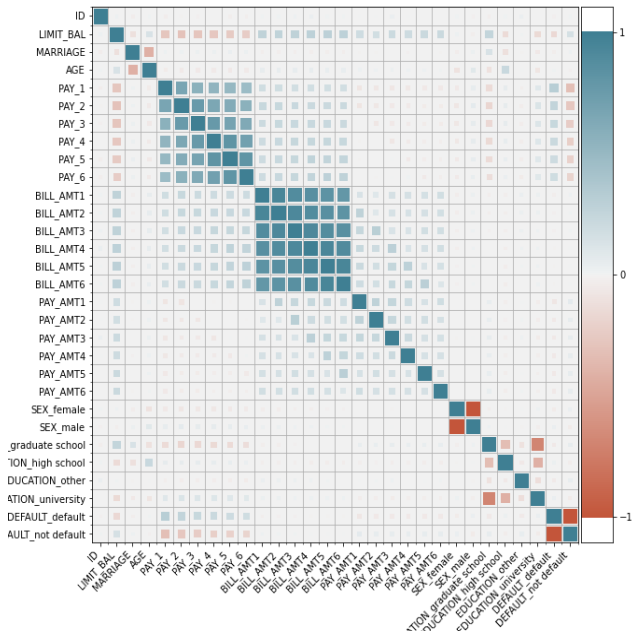

Figure 1: Data set correlation heat map

Figure 2: Latest payment status versus rate of default

As a customer account being in default is a binary event (in Default or Not in Default), the majority of the analysis will group by a factor of interest and evaluate the fraction of customers from the set that are in default, resulting in a fraction between 0 and 1. This fraction represents the rate of default, for example 0.5 indicates that 50% of those specific customers are in default.

As shown in Figure 2, the rate of default is highly dependent upon the payment classification code established by Credit One. In Figure 2 we analyze for the latest payment cycle (September 2005 in this case) and see a significant increase of default rate for customers with 1 or 2 month delays in payment. With just this one data parameter we can differentiate between customers with 15% and 75% rates of defaulting their account.

Analyzing the customer credit limit also yields insight, as shown in Figure 3. Here we can see that customers with credit limits greater than $321K are roughly 3 times less likely to default than customers with credit limits less than $40K, with a consistent trend in-between.

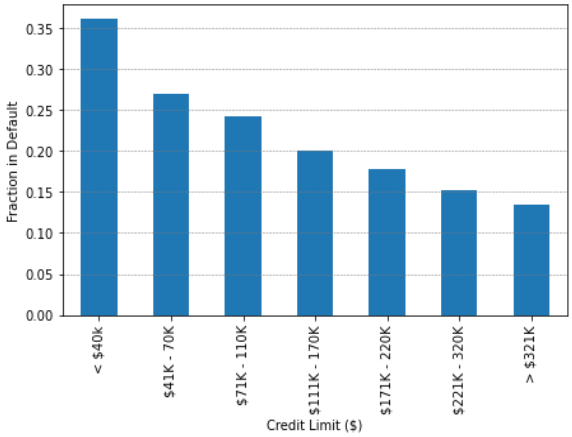

Figure 3: Customer credit limit versus rate of default

Besides directly comparing the provided elements directly to the rate of default, it also can be useful to extract more complex metrics to look for deeper insights and relationships. One hypothesis formed was that the rate of default could be linked to the swing of monthly spending behavior. However, due to the wide range of spending habits and credit limits, it becomes useful to normalize the customer data by the total amount spent over the 6 month data set. In Figure 4, the swing in billing is extracted as a relative variance (variance / mean) of the April 2005 – September 2005 bill amounts. The X-axis values represent the equally populated bins of the data set, with the higher values indicating a greater relative variance (higher swing in billing values). Surprisingly, the customers with greater relative swings in bill amounts are **less** likely to default, by as much as 20% (0.35 to 0.15).

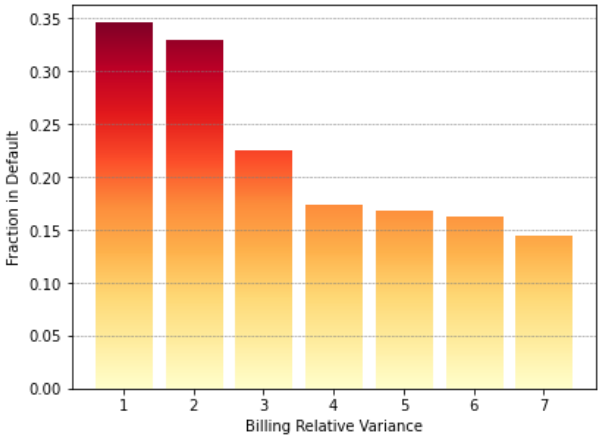

Figure 4: Relative swing in monthly bill versus rate of default

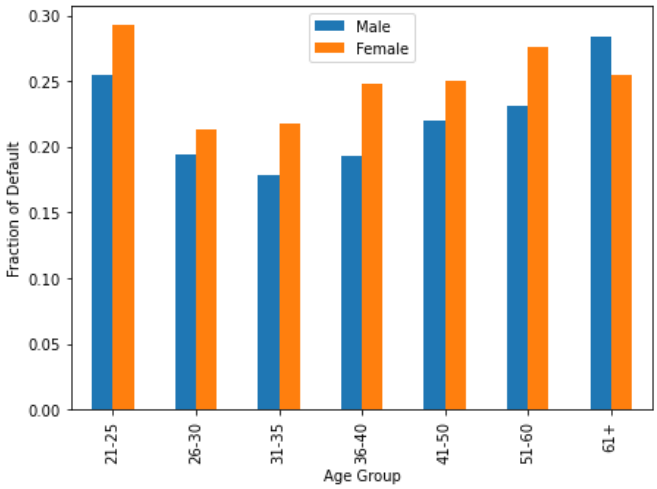

Figure 5: Customer age and gender versus rate of default

The demographic data was also investigated to look for relationships to the rates of default. While the magnitudes of impacts are lower than the financial data, there is a relationship with age and gender. For instance, the data shows that as a whole, female customers are 4% less likely to be in default. Secondly, if we look at the age profile for both genders, as shown in Figure 5, we find that the 26-35 age groups have lower rates of default that the other age groups, for both genders.

Predictive Modelling

To predict credit limits for future customers, regression models were built using the provided data set. Regression models are used to predict continuous numerical values, like we have with credit limit values. For model selection, Random Forest Regression, Linear Regression, and Support Vector Regression models were evaluated approximate accuracy. For this analysis, Random Forest Regression provided the highest performing model. In this case, the model was able to explain 48.5% of the variability of the response data around its mean. To help explain and visualize this concept, Figure 6 plots the predicted customer credit limits versus their actual credit limits in the data set. The accuracy of the model is visualized as the vertical difference between the individual points and the red line – in other words, if the predictive model were perfect, the points would all fall upon the red line. We see from Figure 6 that the model holds relatively well for mid-range values, whereas the % error is much higher at the extremes (low and high credit limits). The 48.5% value is respectable, but likely a bit low as a full predictive model to implement on new customer accounts.

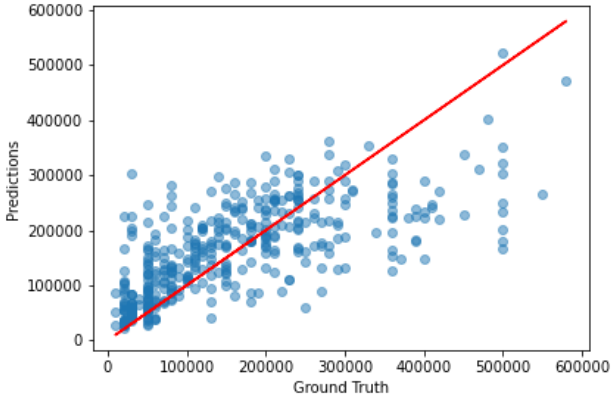

Figure 6: Error between predicted and actual credit limits

Another goal of the predictive modelling was to establish whether a new customer should be approved or not. For this prediction, we look to actively avoid customers that would be likely to default. As this state of the account is binary in nature, we turn to classification models to investigate if we can accurately predict if a customer will be in default, given the remainder of the dataset. For model selection, Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier were compared for accuracy. Upon implementation and tuning, both the Decision Tree and Gradient Boosting models were able to predict default with **82% accuracy**. As seen by the nodes Figure 7, the model is largely driven by the *monthly account repayment status codes* ("PAY_1, PAY_2, etc), meaning that both customer demographics and actual payment/billing amounts play minor or non-existent roles in the model. This provides insight into which specific customer data should be prioritized in gathering

when evaluating new customers. This supports the graphical evidence seen in Figure 2 where those with delays in payments have significantly higher rates of default, corresponding to the right side of the decision tree in Figure 7.

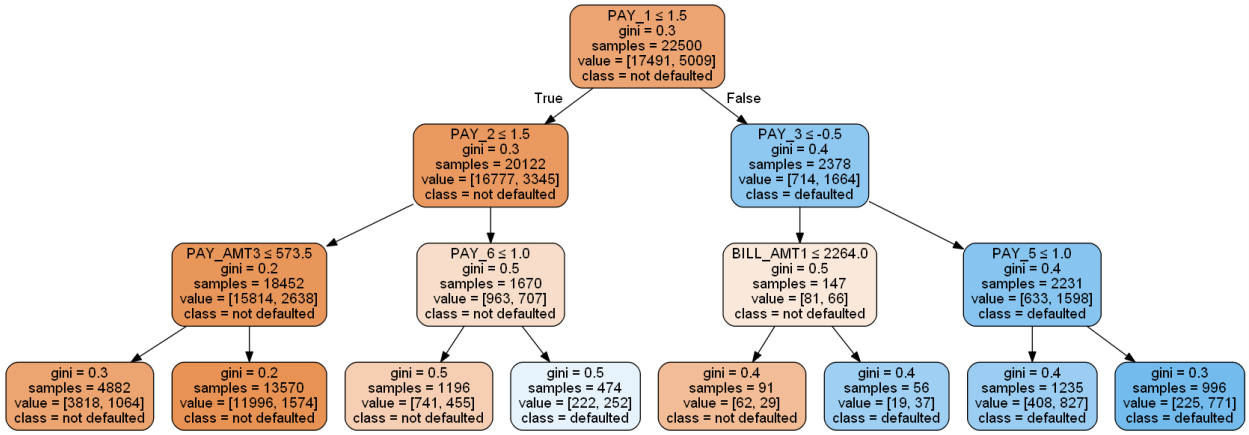

Figure 7: Decision Tree model for predicting account default

Conclusions

This data set did not enable the creation of a strong model for predicting customer credit limits over the wide spectrum of values. While ranges exist where the predictive model is more accurate, we cannot create a credible model for predicting the credit limits for new customers.

Fortunately, determining whether new potential customers will default is within our predictive capabilities. If enabled in the Credit One system, this model could potentially help in reducing the default rates with the customer base. Beyond applying the model, the exploratory data analysis also shows the customer attributes that result in the lowest default rates, which can give insight to future marketing efforts to attract more customers of this category.