Customer Credit Risk Report - Course 2, Task 2

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.

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 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.

Data Analysis

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. As can be seen by the bottom two horizontal rows, 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, which provides a good place to start.

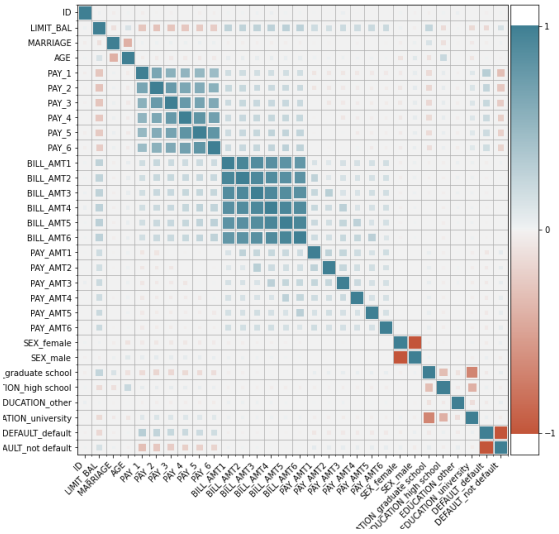

Figure 1: Data set correlation heat map

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.

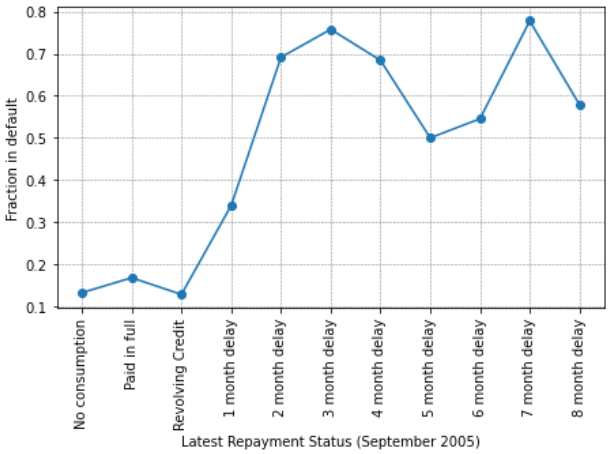

Figure 2: Latest payment status versus rate of default

As shown in Figure 2, the rate of default is highly dependent upon the payment classification code established by CreditOne. 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.

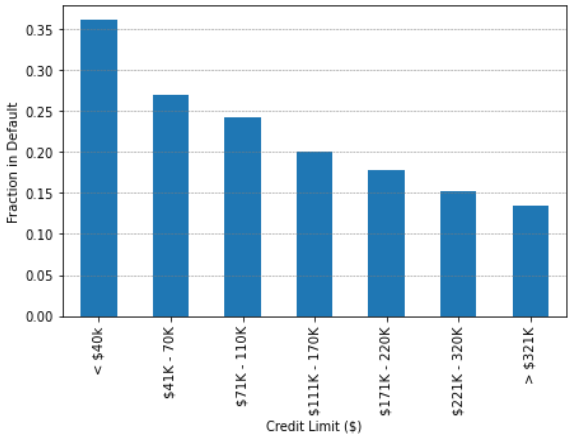

Figure 3: Customer credit limit versus rate of default

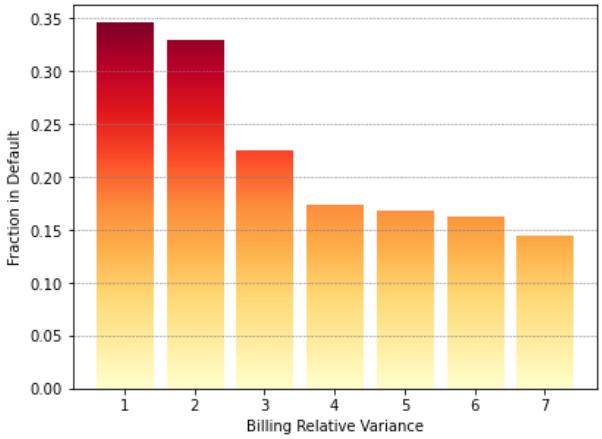

Figure 4: Relative swing in monthly bill versus rate of default

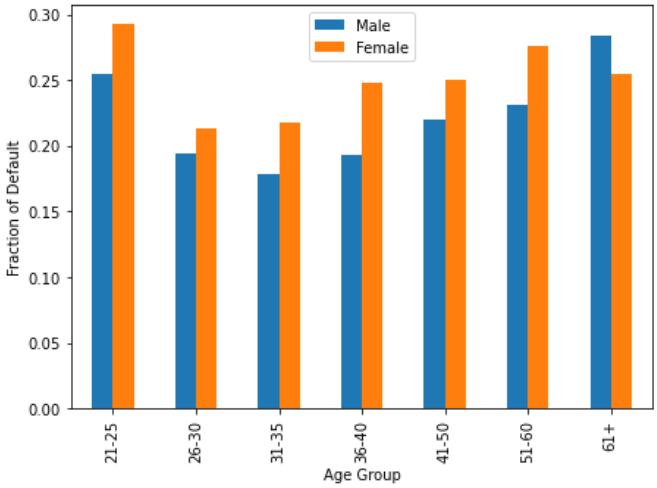

Figure 5: Customer age and gender versus rate of default

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).

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.

Lessons Learned

One of the challenges of this data set was that the result variable (default) was not a continuous variable. This caused significantly more data manipulation with pandas to group the data together in different ways prior to most plot generation.