Capstone 1 Inferential Statistics

In this section, inferential statistics were applied to my dataset of credit card default records, keeping in mind our overall goal of predicting credit card default. Three steps were taken in this stage.

Step 1: Identifying Predictive Variables

First, I examined the correlations between our target variable, default, and each of the other variables. I found that there were no strong correlations or even moderate correlations (defined as having a correlation magnitude greater than 0.7 and 0.5, respectively: https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/). Notwithstanding that, the PAY_X variables which measured the status of repayment of a customer were found to be somewhat correlated with default (a correlation value of greater than 0.1 for all 6 variables). This fact confirmed my intuition that being further behind on repayments should be associated with greater default. Additionally, it was found that the LIMIT_BAL variable which measures the amount of given credit was negatively correlated with a value of about -0.15. My conclusion was that large credit was only extended to customers that the credit agencies believed to be reliable to repay.

Step 2: Correlations Between Independent Variables

Next, correlations were examined between pairs of independent variables. The results of this examination can be found in both the Jupyter notebook on EDA and Inferential statistics, and are reprinted here. Most of the results are very intuitive:

- Age and Marriage are negatively correlated. In the context of our dataset where a value of 1 for Marriage indicates marriage, it means that age and marriage are correlated, as we would assume.
- LIMIT_BAL and Education are negatively correlated. In the context our dataset where lower values for education indicate more education, it means that more credit is given to individuals with more education, as we would assume.
- LIMIT_BAL and the repayment status variables (PAY_X) are negatively correlated. This means that individuals with more credit have better repayment records. This was a finding from our EDA and explains why they are able to take on so much credit.
- Education and Marriage are negatively correlated. Since lower values for Education mean more education and a value of 1 for Marriage indicates marriage, this negative correlation holds. A working hypothesis for this is that individuals pursuing more education need to push off marriage.
- LIMIT_BAL and Marriage are negatively correlated. This means that married individuals take on more credit than single individuals. This makes sense since married individuals tend to have more expenses than single individuals.
- The bill statement variables (BILL_AMTX), previous payment variables (PAY_AMTX), and repayment status variables (PAY_X) are correlated with each other and themselves. This makes sense: they are all in one sense measuring the same thing. Greater spending one month should be related to greater spending the next or previous month; it should also be related to the amount paid back, and whether or not you are behind on payments.
- Age and LIMIT_BAL are positively correlated. This means that older individuals have more credit taken on. This makes sense since they have had more time to take on credit.

- Education and Age are correlated, meaning that they are negatively correlated because greater values for education indicate less education. A possible hypothesis is that credit lenders are willing to lend to individuals with more education at a younger age because they have greater confidence in their ability to repay due to their higher education.
- LIMIT_BAL is correlated with repayment totals and bill amounts. This is intuitive: taking on more credit should be associated with greater amounts spent and greater amounts paying back that credit.

Step 3: Hypothesis Tests on Relationships Among Variables

Many relationships among variables were found from the visual Exploratory Data Analysis stage, but only 5 were chosen for testing:

1. The correlation between default and final repayment status is about 0.4.
2. Men default at a higher rate than women
3. People with only a high school education have higher rates of marriage than people with graduate school education
4. The mean age of people with only a high school education (about 40.3) is considerably higher than people with a graduate school education (about 34.2)
5. Married individuals are more likely to default than singles (23.5% vs. 20.9%)

These relationships were tested with a Hypothesis test on the Pearson correlation coefficient (claim 1) and two-sample bootstrap hypothesis tests for difference of means (claims 2 – 5).

It was found that for all the relationships, the Null Hypothesis that there was in fact no real difference between the groups could be rejected.