Credit One EDA

In order to perform an exploratory data analysis of Credit One data several tasks needed to be performed. Multiple extraneous data columns were removed, followed by removing duplicate rows, then replacing the existing generic header. After this, the data types were converted to numerical and then discretized to make them more usable for machine learning purposes.

After this, I was able to determine that we have 24 variables to work with, 30,000 rows of data, and zero missing cells. The mean limit balance was $16,784.32 overall, with an average of $13,0101 for those who defaulted. The average age for borrowers was 35, with the average age for those who defaulted being 35.72. There was very little evidence of a correlation between any of the variables of a slight negative correlation with the LDEFAULT and PAY_1 (-0.32.) Additionally, some signs of covariance exist around LDEFAULT and PAY_AMT1(5.013746e+02), and LDEFAULT and BILL_AMT1 (6.003941e+02.) Outside of this, I was not able to find very much data that would indicate a relationship between variables.

In order to perform machine learning tasks, I will be using LDEFAULT as my Y-variable, and assigning the rest of the features to X. Initially, I will use a logistic regression model to try to determine an amount at which to approve people for credit. In the case that the data is not sufficient to model this I will attempt to build a classifier model to determine simply whether or not someone should be approved for credit.