**Course 2, Task 3: Report to Credit One**

Using Machine Learning to Predict Credit Default Likelihood and Determine Credit Limit

Anders Dowd - Spring 2022 Data Analytics Certification Program

***(A) Overview***

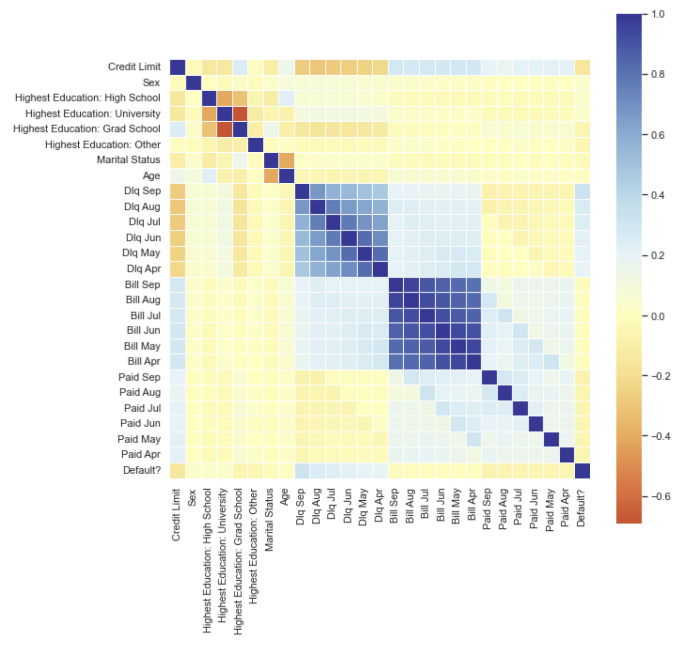
Over the past year Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved. We are tasked with finding a better solution to the question of who should be granted loans and to what limit.

We know that we cannot control customer spending habits, so we must find a way to identify and deny high risk borrowers before determining what level of credit to issue to those who can be approved.

***(B) Data Cleaning and Pre-processing***

Credit One has provided a dataset of over 30000 customer records. The data includes demographic identifiers (age, sex, marital status, education level) as well as financial ones (current credit limit, default status, and six months worth of billing, payment, and delinquency records). The data cleaning process consisted of removing duplicate entries, reformatting variables to useable data types, renaming columns, one-hot encoding of certain variables (namely education, sex, and default status), and dropping unnecessary columns and duplicate headers. The cleaned data left us with 29965 unique entries.

**(*C) Exploratory Data Analysis***



**Figure 2.1: Correlation Heatmap of Credit One Data Set**

Analysis began with basic visualizations of the data in the form of histograms, scatterplots, and correlation heatmaps. The clean data frame was also broken down into various pivot tables to analyze individual variables and how they relate to the other features. This preliminary analysis highlighted which features had the greatest influence on the target variables of credit limit and default likelihood. Highest and lowest risk classifiers were also identified.

Strongest correlations (see Figure 2.1):

Month to month bills (0.80-0.95) -- higher bills one month tend to indicate higher bills across all months.

Month to month delinquency (0.47-0.82) -- delinquency in payments indicates a higher chance of a pattern of delinquent payments.

Delinquency to default (0.19-0.32) -- higher delinquency rates indicate a higher risk of default.

Monthly bill to delinquency (0.18-0.29) -- higher monthly bills indicate slightly higher risk of delinquent payments.

Credit limit to default (-0.15) -- slight inverse correlation between credit limit and default risk.

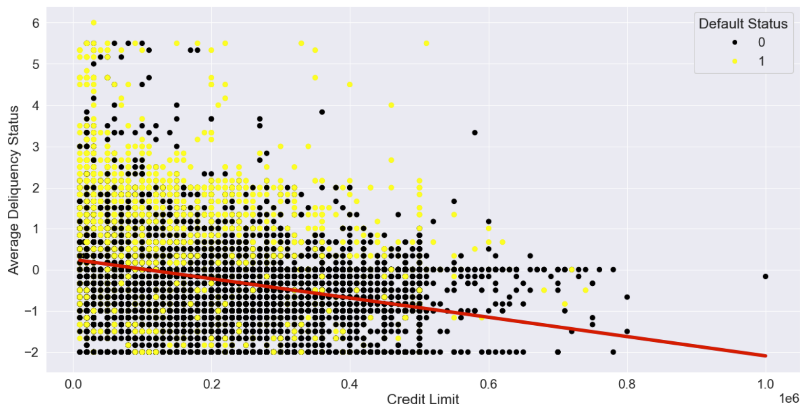
|  |  |
| --- | --- |
| **Highest Risk Customers (Default Rate)** | **Lowest Risk Customers (Default Rate)** |
| Male (24%) | Female (21%) |
| Divorced (26%) | Single (21%) |
| High School only education (25%) | Grad School education (19%) |
| Over 70 years old (33%) | 30something years old (20%) |
| Dormant 6 months or more (38%) |  |

**Table 2.1: Comparison of Highest Risk and Lowest Risk Customers by Demographics**

By analyzing the default rates of individual classes of the demographic variables, it became evident that there were identifiable risk factors that could prove helpful in developing a method for identifying high risk customers (see Table 2.1). Furthermore, delinquency in payment is clearly a predictor to some degree of default, while customers with higher credit limits tend to be more reliable in paying their bills in a timely fashion. In Figure 2.2, The measurement scale for the delinquency status is:

-2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1-9 = payment delay in months

Credit limit is in USD; yellow markers indicate defaulters, while black indicate non-defaulters.

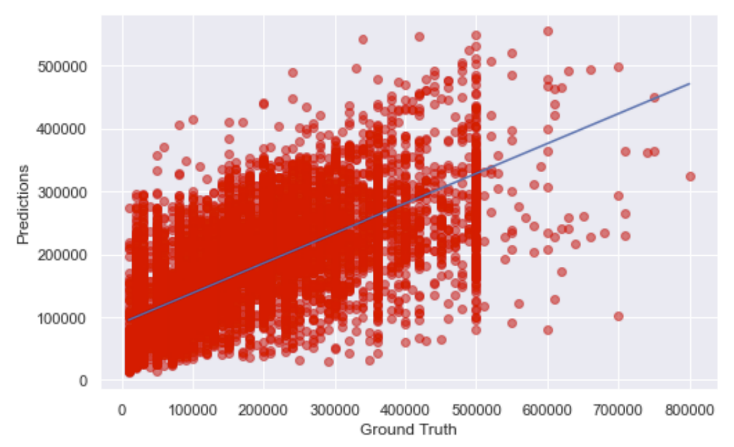


**Figure 2.2: Relationship of Payment Delinquency to Credit Limit and Default Likelihood**

***(D) Determining Credit Limit through Regression Models***

In attempting to build a model to determine credit limit, I began by deciding which features to use. The first set of features included all demographic info (sex, education level, marital status, and age); 6 month averages of delinquent payment status, outstanding balance, and payment amount; and default status. I ran cross validation and compared the R2 scores of three algorithms: Random Forest Regression, Linear Regression, and Support Vector Regression. I used this same cross validation strategy on various combinations of feature variables to see if one combination would be more successful than others. Unsatisfied with the efficacy of those three algorithms, I also tested Logistic Regression and AdaBoost Regression.

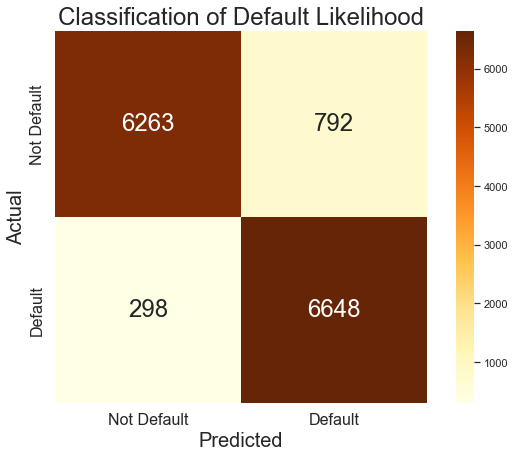
In this regression exercise, the best results come when we use all of the available features rather than whittling them down to just a handful, even when that handful consists of only the features most highly correlated to the target. The average figures also are less effective than the individual monthly figures. Random Forest Regressor has the best return of any of the algorithms, but none of them are very reliable. With RFR, the root mean squared error is close to $100000, which is a huge margin of error in terms of individual loan qualification and an ineffective range for dialing in a credit limit. With an R2 score of .465 it left something to be desired in its reliability.



**Figure 2.3: Ground Truth Credit Limit vs Predictions with Regression**

Given the nature of the data available to us, creating a regression model to predict appropriate credit limits appears to be outside the realm of possibility.

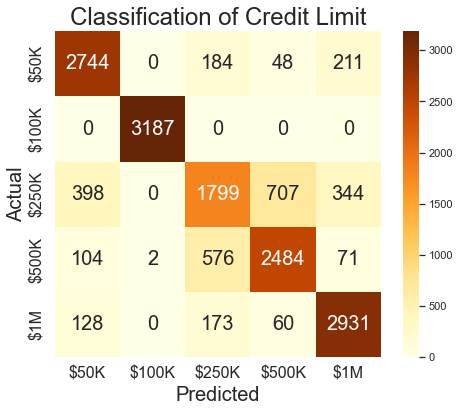
***(E) Determining Likelihood of Default through Classification Models***



**Figure 2.3: Confusion Matrix of Default Predictions**

Unable to return a successful regression model for credit limit, we turn to answering the question of whether or not we can build a better, more reliable process for approving loan applicants on an up or down basis. For this task, three classification algorithms were employed before the best was chosen. Of Decision Tree, Random Forest, and Gradient Boosting Classifiers, Random Forest achieved the best balance of accuracy, precision, and recall. However, due to the highly imbalanced nature of the target variable, I was only able to achieve this success by training the model using a similar number of defaults and non-defaults. I balanced the classes in the target variable (default likelihood) using the RandomOverSampler function in the imblearn library of Python. After doing so, I was able to attain a model that achieved 92% overall accuracy with 89% precision and 96% recall in the minor class. This was critical to building a reliable model for our purposes since it is absolutely essential that we avoid false negatives when predicting a customer will not default.

***(F) Determining Credit Limit through Classification Models***



**Figure 2.4: Confusion Matrix of Credit Limit Predictions**

Since building a regression model for determining credit limit proved untenable, I attempted to build a classification model by discretizing credit limit amounts. I experimented with several binning strategies, including quartiles, but the best results came from binning into five bins of expanding ranges ($50K, $100K, $250K, $500K, and $1M). I employed the same oversampling strategy to balance the target and was able to achieve a Random Forest model with 81% overall accuracy, precision, and recall. However, some classes were more successfully predicted than others, so I would not recommend making decisions based on this model alone. A more prudent strategy would be to use the more consistent Default Likelihood Classification model first to determine if a customer is credible or not, then dial in a credit limit based on the recommendations of the Credit Limit Classification model.