C5T3 – Build and Evaluate Models

**TO:** Guido Rossum

**FROM:** Joni Johnson-Powe

**DATE:** December 28, 2019

**RE:** CreditOne Customer Data & Model Evaluation

**Overview**

CreditOne has seen an increase in the number of customers who have defaulted on loans recently. We were requested our Data Science team to design a solution to addressing this business issue. CreditOne provided historical data that included demographic information such as sex, education and age as well as account information such as bill and payment history and credit limits. The data included 30,000 lines of customer information.

**Approach**

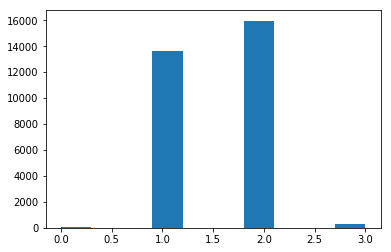
We utilized Python and Sci-Kit Learns in our evaluation of the customer data. As part of our analytic process, we engaged in the following steps:

1. Data cleansing and preprocessing
2. Covariance Estimation
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Selection/Classification
6. Model Tuning
7. Model Evaluation

**Observations**

1. As part of our initial preprocessing review of the data file, we found no missing values or unusual characters for the 30K records. A majority of customers were women and single with a college education under the age of 35 years old. The median credit limit of customers is $167K with a low balance of $10K and high of $1M. See graphs below:

**Marital Status of Customers**

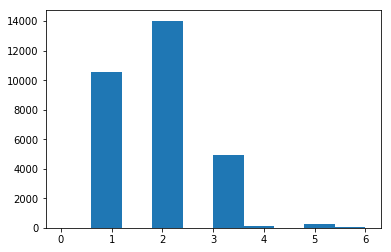


1 = Married

2= Single

3= Divorced

**Education Level of Customers**

****

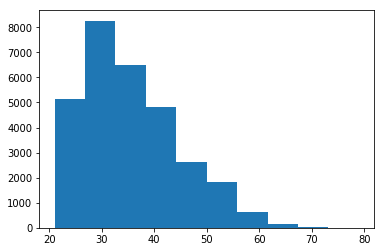
1 = Grad School

2= College Grad

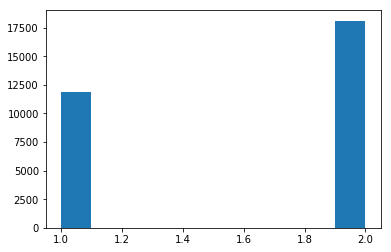
3= High School

4,5,6= Other

**Age of Customers**

****

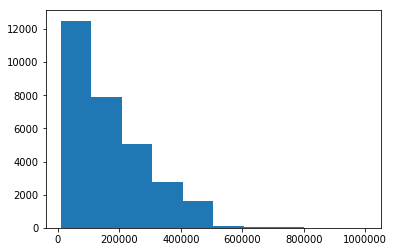
**Sex of Customers**

****

1 = Male

2= Female

**Credit Limit Balance**

****

1. **Correlation & Covariance Matrix**

The CreditOne dataset included 26 demographic, credit, bill amounts, payment history and default status variables for each customer. We executed a Correlation matrix as well as a Covariance matrix to help us understand the relationship between the various data elements.

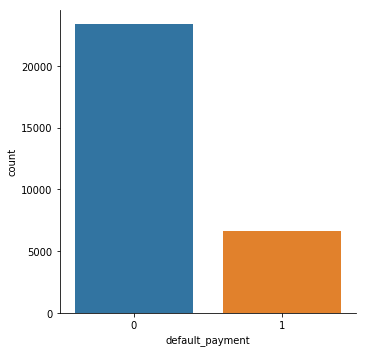
For Correlation, the closer the result is to 1.0 between two elements, the more aligned the elements are. Under Covariance, if two variables increase together the co-efficient is positive. On the other hand, the coefficient is negative if as one variable increases as another decreases.

Overall our Correlation matrix found very few highly correlated data elements in the CreditOne data. The History Payment Status had the closet relation to “default payment status” with the most recent payment have the highest correlation of .324794. Bill Amt 6 and Bill Amt 5 had the highest correlated value of .946197. Bill Amts variables were highly correlated with other Bill Amts from other periods as well as with Payment history variables.

For Covariance, Payment History had the most positive covariance variable with “default payment status”. Bill Amt data elements had an inverse or negative relation with “default payment status”. As Bill Amts increase the default payment status was closer to 0 or no default.

1. **Default Status**

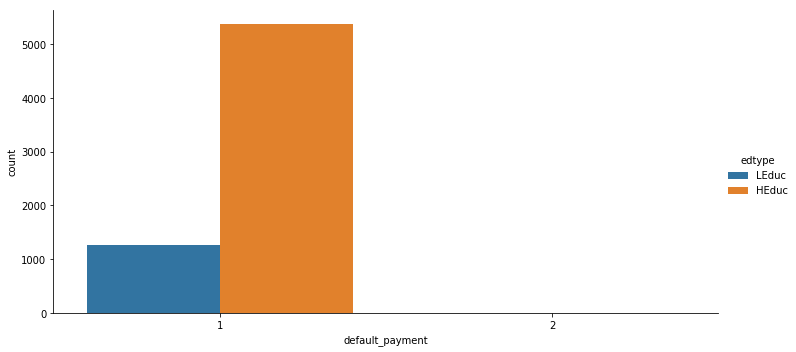
We generated additional graphs around the default payment status that presented interesting observations. Customers not in default significantly exceeded customers in default. Also, higher educated customers were in default at a much higher rate than lower educated customers. Finally, lower credit limit customers were in a much higher default status than higher credit limit customers.

****

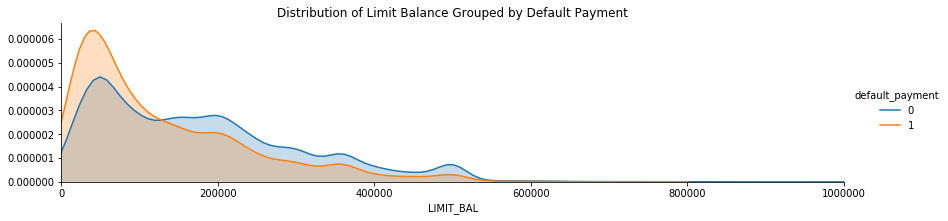
0= No Default

1= Default

**Default # Comparisons between Lower Educated vs High Educated**

****

**Limit Balance Distribution Grouped by Default**

****

1. **Models: Tuning and Evaluation**

We evaluated 4 models for purposed of the CreditOne data. Before running models, based on the Correlation Matrix and Covariance results, we narrowed our data selection to the following variables:

**EDUCATION**

**PAY\_0**

**PAY\_2**

**PAY\_3**

**PAY\_4**

**PAY\_5**

**PAY\_6**

**default payment next month**

We selected these variables over the results derived from the RFE process as it resulted in a better cross validation score but without overfitting. Accordingly, we derived the following results for each of the models:

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest Regression | Linear Regression | SVR | RFE |
| 0.99981356 | 1.0 | 0.88555109 | 1.0 |
| 0.99992297 | 1.0 | 0.89302408 | 1.0 |
| 0.99995936 | 1.0 | 0.92823418 | 1.0 |
| **0.9999866282114104** | **1.0** | **0.9684907495138042** | **1.0** |

The model selected based on the results above was SVR. The performance of the model is as follows:

**R Squared: 0.912**

**RMSE: 0.116**

****

**Conclusions**

In reviewing the results of our analysis, we found that the vast majority of data elements including demographics such as sex and age as well as bill amount and payments amounts had little impact predicting whether a customer would default on loan payments. However, the payment history had the most predictability in terms of default payments. Education level also seemed to have a relevant relationship with predicting default. However, the data surprisingly showed that higher educated customers had more defaults than lower educated customers.

The low correlation values seem to infer that perhaps CreditOne should look to gather and evaluate other factors that may impact a customer’s ability and/or willingness to pay loans timely. For instance, perhaps more insight may be gained by through salary information, debt levels, locations and type of debt. Does region or state have any impact? Would revolving credit vs. secured credit have an impact? We suggest that CreditOne consider gathering additional consumer information, validating additional data elements and remodel.