|  |  |
| --- | --- |
| To: | Guido Rossum; Senior Data Scientist; Credit One |
| From: | John Enrietto |
| Date: | 4/26/22 |
| Subject: | Request for analysis of Credit one increase of default creditors |

This summary is to review data generated by reviewing data submitted by Guido Rossum regarding an increase in default accounts seen by Credit One of their customers. For analysis, Mr Rossum supplied an SQL database at deepanalytics:Sqltask1234!@34.73.222.197/deepanalytics. The information was downloaded to a local machine on 4/5/2022.

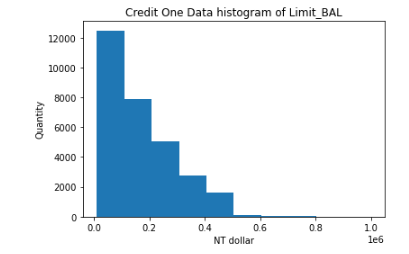
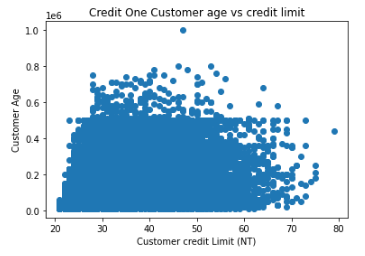
After download, data was loaded into a datafield and cleaned to remove duplicate and erroneous data. The datafield was then reduced to 29965 fields (discrete customer accounts), by 24 features for analysis. The 24 features are defined as follows:

|  |  |
| --- | --- |
| X1 | X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. |
| X2 | X2: Gender (1 = male; 2 = female). |
| X3 | X3: Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others). |
| X4 | X4: Marital status (1 = married; 2 = single; 3 = divorce; 0=others). |
| X5 | X5: Age (year). |
| X6-X11 | X6 - X11: History of past payment. We tracked the past monthly payment records (from  April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7  = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005.  The measurement scale for the repayment status is:  -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1 = payment delay  for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight  months; 9 = payment delay for nine months and above. |
| X12-X17 | X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in  September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of  bill statement in April, 2005. |
| X18-X23 | X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005. |
| X19 | Y: client's behavior; Y=0 then not default, Y=1 then default" |

A pandas function (.corr) which runs a Pearson Correlation function on each combination of the selected data features was run. None of the single feature to feature vectors showed any meaningful correlation. Neither the defaultcode column or the Limit\_Bal feature show a correlation to any of the input data: (data the bank will have before any credit history). See table below.

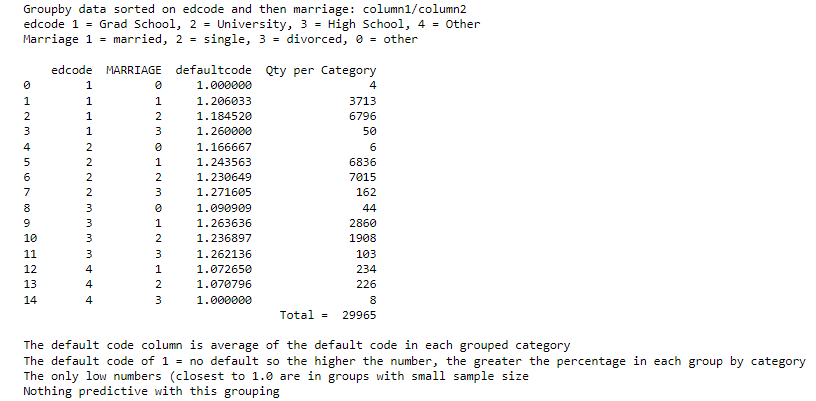


Data in the datafield had a mix of integer and object data. To allow for other types of analysis, the three object features (sex, education, and default payment next month) were converted to integer data. There are several ways to accomplish. A pre-written function called dummie was reviewed. It works but the author doesn’t like to have to add multiple binary variables. Instead using a simple for loop with imbedded if/elif was used to create 3 replacement variables respectively. (gendercode, edcode, defaultcode)

Multiple x vs y and histogram data plots were developed to review the data and review ay possible insight to further analysis. None of these plots showed anything worthwhile to pursue further analysis. Some of these plots are shown below.

Profiling is another Pandas subroutine used to perform additional

EDA. Profiling did not appear to show much correlation that would give insight why there is an increase in customer defaults. Full profile report is attached in the github account for reference.

Groupby functions were used as well to try to find a predictive correlation.

With little initial information gleaned from one and two variable analysis, we then moved on to more complex analysis using regression techniques. Data was broken into 4 groups with both the dependent and independent variables broken out with 70% used to train the model and 30% held back as clean data for predictive trials. Two tests of three different functions were run. The different predictive analysis functions used in each test run are listed below:

1. Random Forest Regression
2. Linear Regression
3. Support Vector Regression (SVR)

Each of these returned an R2 value which will be used evaluate the predicted capability of the function. As mentioned above, each set of these 3 functions was run twice: once with LIMIT\_BAL as dependent variable, a second time with default payment next month. All 23 remaining features were included in the initial analysis The R2 output variable is listed below and shows little predictive information, with all values below 0.5. Two additional trials were run with the dependent variable Limit\_Bal varying the “cv” parameter from 3 to 5: where the cv parameter specifies the number of splits in data when running the function. This change as well showed no improvements in predictability.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regression Analysis results (Avg of R2)  23 independent variables | | |
| Dep Var = Limi\_ Bal | | Dep Var = default status |
| cv=3 | cv=5 | cv=3 |
| Random Forest Regression | 0.4673 | 0.4693 | 0.1832 |
| Linear Regression | 0.3516 | 0.3529 | 0.1206 |
| SVR | -0.05030 | -0.0402 | -0.0852 |

To try to improve predictability, LIMIT\_BAL data was binned using the qcut process in pandas. Two different methods were used to do this. First was a function called qcut which put an approximately equal quantile of accounts into a series of bins. The second method was to try setting bins with equal limits using 4 bins and 8 bins. Summary of the two concepts are listed below. Bin limits are listed at the end of the report on page titled Reference data.

|  |  |
| --- | --- |
|  | Regression Analysis results (Avg of R2)  23 independent variables |
| Dep Var = default status  LIMIT\_BAL discretized into equal quantile bins |
| cv=3, # Quantile Bins = 8 |
| Random Forest Regression | 0.1687 |
| Linear Regression | 0.1201 |
| SVR | -0.0856 |

|  |  |  |
| --- | --- | --- |
|  | Regression Analysis results (Avg of R2)  23 independent variables | |
| Dep Var = default status  LIMIT\_BAL discretized into equal sized category bins | |
| cv=3, # Bins = 4 | cv=3, # Bins = 8 |
| Random Forest Regression | 0.1687 | 0.1686 |
| Linear Regression | 0.1201 | 0.1204 |
| SVR | -0.0856 | -0.0856 |

|  |  |
| --- | --- |
|  | Classification Analysis results (Avg of R2)  23 independent variables |
| Dep Var = LIMIT\_BAL  LIMIT\_BAL discretized into equal sized category bins |
| cv=3, # Bins = 8 |
| DecisionTreeClassifier | -0.0979 |

The entire data set does not seem to give any predictable results. Next we will remove features to see if we can find a set that will give better results. The best results to date, which we will use for future testing, is RandomForestRegression (RFR), which will be used moving forward.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis type | Independent Vars | Dep Var. | Avg R from CrosVal |  |
| RFR | Marriage, Sex, Education, Age, Limit | default | 0.4669 |  |
| RFR | Marriage, Ed, Age | default | .4664 |  |
| RFR | Marriage, Ed | default | .4672 |  |
| RFR | Ed, Age | default | .4669 |  |
| RFR | Sex, Marriage | default | .4672 |  |
| RFR | History of Past Payment  (X6 – X11) | default | .4661 |  |
| RFR | History of Past Payment | Limit Bal | .4661 |  |
| RFR | Amount of bill statement  (X12 – X17) | default | .4665 |  |
| RFR | Amount of bill statement | Limit Bal | .4657 |  |
| RFR | Amount of Previous Payments  (X18 – X23) | default | .4665 |  |

**CONCLUSION:**

It is the opinion of this analyst that the information given is not sufficiently adequate to predict future performance of customers.

Additional information that may help in the future would include the individual’s credit score, their income, and their total net worth.

Reference Section:

Bin limits for 8 equal quantile bins

|  |  |  |
| --- | --- | --- |
|  | Bracket Limit | Qty |
| 1 | (9999.999, 30000.0] | 4080 |
| 2 | (190000.0, 240000.0] | 4074 |
| 3 | (90000.0, 140000.0] | 3837 |
| 4 | (50000.0, 90000.0] | 3770 |
| 5 | (30000.0, 50000.0] | 3593 |
| 6 | (340000.0, 1000000.0] | 3580 |
| 7 | (140000.0, 190000.0] | 3553 |
| 8 | (240000.0, 340000.0] | 347 |

Bin limits for 8 equal bins sizes: LIMIT\_BAL( max-min)/8

|  |  |  |
| --- | --- | --- |
|  | Bracket Limit | Qty |
| 1 | (9010.0, 133750.0] | 14532 |
| 2 | (133750.0, 257500.0] | 8725 |
| 3 | (257500.0, 381250.0] | 4460 |
| 4 | (381250.0, 505000.0] | 2042 |
| 5 | (505000.0, 628750.0] | 147 |
| 6 | (628750.0, 752500.0] | 53 |
| 7 | (752500.0, 876250.0] | 5 |
| 8 | (876250.0, 1000000.0] | 1 |

Bin limits for 4 equal bins sizes: LIMIT\_BAL( max-min)/4

|  |  |  |
| --- | --- | --- |
|  | Bracket Limit | Qty |
| 1 | (9010.0, 257500.0] | 23257 |
| 2 | (257500.0, 505000.0] | 6502 |
| 3 | (505000.0, 752500.0] | 200 |
| 4 | (752500.0, 1000000.0] | 6 |