**Course Two Task3 Report**

**by**

**Shifeng (Steve) Li**

**01/22/2022**

In this project, the CreditOne customers repayment data has been analyzed using Python and Data Science techniques to go through the data science process steps to gain the good accuracy to identify co-related patterns and to develop prediction model for Credit One. In the following, will describe the different phases of the process.

1. Cleaning and Pre-processing data

In order to support the data analysis process to produce the analytical results with higher accuracy to realistically reflect the information, the to be analyzed business operation data must be cleaned, and pre-processed.

During the data cleaning and pre-processing phase, the following steps have been done for the CreditOne customers repayment information raw data set:

1. Removing duplicates
2. Replaced the missing data with the values that best fit the data pattern,
3. Change/adjust the data set column names
4. Sort the data
5. Change/convert the data type so that the data can be fed into the data analysis models and algorithms.
6. Exploratory Data Analysis (EDA)

After the data set is cleaned and pre-processed, will go through the EDA phase to analyze the data set to summarize their main characteristics. By using statistical graphics and data visualization methods, the EDA process will explore the data, formulate the hypotheses about the causes of observed phenomena, helps the selection of appropriate statistical tools, and techniques, assess assumptions on which statistical inference will be based. Also, the EDA process can provide a basis for further data collection to support the iterative/looping in the data analysis life cycle.

During the EDA phase to analyze the CreditOne customers repayment information data set, the following analysis have been conducted:

1. Review all the columns in the data set, to get best understanding of their value types, value meanings, size/range, and meta data background information.
2. Applied statistics to analyze the data set to list out the statistical summary of the core data attributes.
3. Applied the Phyton plot tools to develop visualizations for the CreditOne data attributes to help identify the data patterns.
4. Histograms
5. Line plots
6. Scatter plots
7. Box plots
8. Factor plots

From the EDA process on CreditOne data set, observed the following information:

1. Divorce group has the highest chance to be default, married is the second, single the third.

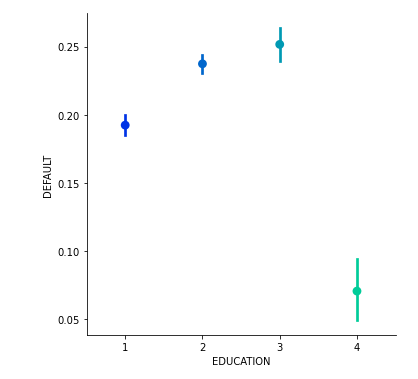
married\_default\_ratio is: 0.23462581543648758

single\_default\_ratio is: 0.2094700533082471

divorce\_default\_ratio is: 0.26006191950464397

others\_default\_ratio is: 0.09259259259259259

1. High school education level has the highest chance to be Default, university level is the second, graduate school the third, the others is the lowest



'graduate school': 1, 'university': 2, 'high school': 3, 'other': 4

#### By reviewing the September 2005 repayment situation, the customers who have No consumption: -2; Paid in full: -1; The use of revolving credit: 0 have the least chance to be default. While the customers who have payment delay for two months, three months, and four months have the highest rate to be default.

#### Customers with lower limit balance are more likely to be default

#### 

#### For the default customers, their age range are mainly between mid-twenties to mid-thirties.

#### 

1. Calculated the correlation and covariance for the CreditOne data set attributes to have quantified information for dependency analysis.
2. One-Hot Encoding

In the CreditOne Data pre-processing, the independent variable like EDUCATION has been converted to dummy variables by using get\_dummies tool in Python Panadas,

Also, the dependent variable DEFAULT has been converted to dummy variables using the sklearn.preprocessing LabelEncoder.

1. Build three models and choose the best

In this phase, built three different models using the three different classification algorithms: DecisionTreeClassifier, RandomForestClassifier, and KNeighborsClassifier. Applied each model on the CerditOne training data with the default algorithm parameters, then collected the accuracy score for each model.

Next, applied parameter tuning for all the three models with the cross-validation runs to collect a new set of score values. By comparing the score results among the tuned three models, and comparing the before and after scores, identified the best model, which is based on the DecisionTreeClassifier algorithm, with its optimized parameters discovered by the parameter tuning.

1. Model Evaluation

After identified the best model with its adjusted parameter values from parameter tuning, applied the model to run against the test data, and be able to get the accuracy score of 0.82, which means the model can be used for prediction, as its accuracy level exceeds the common threshold 75%.

From running through the Data Analysis on the CreditOne customers repayment data set, the quantified data tell us, that the customers’ education level, marital status, age range, and the granted Credit Limit have impact to the default rate. The discoveries are illustrated in the above EDA section. These features can be used for the machine learning models to support the prediction and decision making for tuning/adjusting the credit issuing process criteria. Also, after the credit issuing criteria experimental/pilot adjustment, further data can be collected to support the new iteration of the data analysis to prove the model and to have further enhancement.