**Papyss Ntoukap**

**Exploratory Data Analysis Lessons Learned Report**

The objective of Task 2 was to perform an Exploratory Data Analysis (EDA) from Credit One Scoring historical data previously pulled from its database.

Recall that the EDA enables Data Scientists to do a few exercises and manipulation of the dataset and to visualize them using python visualization libraries such as Matplotlib or Seaborn as it was the case in Task 2.

We should note that EDA is a fundamental and vital aspect of data analysis as it helps to decipher significant insights regarding the data patterns and trends likely to be used to build, through machine learning, a model useful to address the business problem we are trying to solve. However, the "Garbage In, Garbage Out" concept recommends the dataset to be thoroughly cleaned and making sure we are using the correct data types to prevent unexpected results or errors. Here comes the step of data preprocessing where the quality of data should be obtained to prepare for data exploratory analysis and visualization.

Before that, the Credit One dataset was imported, and a quick check of the first five rows was performed using the appropriate Pandas command function. The structure of the dataset was performed to have an overview of the Credit One dataset, such as the total number of rows and columns, the datatypes (this can be obtained separately by running the appropriate command function), and the number of null values if any. We looked at the statistical summary of the dataset (this can be done only for numerical value) using the .describe() method, and alternatively, a Pandas Profiling function to have a comprehensive overview of the statistical description of the data.

We were looking to build a regression model to answer the business question related to Credit One. The categorical features in some of our dataset needed to be converted in numerical value because the regression model does not handle any categorical value. To do this, after knowing our datatypes object, we used one of Pandas command functions (a built-in pandas functions pd.to\_numeric()). Here, it should be mentioned that we have to use a "coerce" extension to be able to convert our categorical data to numerical values (part of the Data Transformation of the Preprocessing phase) after removing all of the duplicated data and viewed the entire datatypes of the dataset.

The "coerce" extension added to Pandas data conversion function (using a built-in pandas functions pd.to\_numeric()) to convert the object datatypes gave me a "float64" object instead of an "int64" object, though both are numerical values. This figure means Credit One's dataset contains decimal values.

We used Pandas get\_dummies() function to convert the non-numerical data (the categorical data) of our dataset to numerical data (binary values). This must be done to be able to build the regression model used for prediction. All these activities are part of the preprocessing of the dataset. At this stage, and after verifying that the entire datatypes are numerical, the EDA portion of the data analysis can then be performed, following the importation of Pandas require modules. Here, the column name is checked to verify their correctness (of course, these column names may have been renamed, if needed, during the preprocessing phase of the data).

The visualizing phase of the EDA has involved:

* Visualization of the relationship between variable using the Scatter plots and verifying the range and mode of each variable using the Line plots. This helps to measure the fitness of the model, to prevent bias, and even to check for variable collinearities
* Checking for collinearity (correlation and covariance) between variables
* Building histograms for different variables of the dataset to see how each data is distributed
* Building the Box Plots to observe variances in the dataset and identify any data outliers
* Correlation intended for addressing collinearity or checking for relationships between features
* Covariance. It brings some emphasis on the correlation by helping to measure the degree of change between variables, and in the case of Credit One, to understand the impact of any variables on default rates. This is also important because it will help to eliminate any variables that unnecessary to answer the business question, therefore restricting the regression model to be built around variables that are significant concerning the business problem and the derived business question.

The EDA must be made under the sight of addressing the business question that underlies the business problem. This means any visualization performed at this stage should be intended for answering the business question and to build the regression model that will be used for future predictions.

**Summary of Our Findings**

First, we noticed that Credit One's dataset has a weird look at the first observation. The correlation, along with the covariance analysis of all the features, shows that there is a relationship between the predictor variables (independent variables) and the predicted variable (the default loan), but at various degree and range, either negatively or positively.

The dataset shows some strong correlation (at different ranges) between variables such as EDUCATION, SEX, MARRIAGE, AGE, and LIMIT\_BAL with the Default Loan.

Likewise, the correlation between variables PAY and the Default Loan is somehow low or very low.

The dataset did not present any missing value. However, some data within the dataset are duplicated, which we have removed during the preprocessing phase (data cleaning).

Upon checking the datatypes of Credit One's dataset, all the data appears as Object datatypes whereas Credit One's data file as extracted from its database show only THREE COLUMNS with non-numerical datatypes, which in the case of predicting the Default Loan using a Regression model, is an issue to be solved.

Solving the object datatypes issue was done by forcing all the numerical columns to be changed to numerical datatypes and converting the non-numerical column in binary values in order to be able to perform the EDA.

The conversion of the non-numerical datatypes to binary values for the EEDUCATION and Default Loan columns variable with the help of built-in get\_dummies() function of Pandas has created news columns for each specific variable of those columns respectively, increasing therefore the number of columns of the entire dataframe.

I discovered that a built-in Pandas data conversion function is useful to convert the object datatypes object instead to numerical values, and the get\_dummies() function of Pandas is essential to deal with non-numerical data of our dataset to binary values that will permit to develop the regression model make predictions.

One limitation found in the dataset during the analysis is the absence of the customers' income status, housing status (homeowner or renter?), and debt status and any other expenses driving factors (extra alimonies) which, as a recommendation, should be used in conjunction to their credit score in order to approve the loan request and determine the amount of loan.