To: Credit One’s Senior Data Scientist, Guido Rossum

From: Alma P. Martinez, Data Analyst/Scientist

Date: September 27, 2021

Subject: Data Science Process Report

Body:

In the last course, we learned to work with data and how to prepare it to be useful for us. Significantly, knowing the data science process is very important to state our issue/question and prepare the data we have to get an answer. In addition, knowing how to prepare your data is very important to obtain the desire results or answers we need. The first step in the Data Science Process (DSP) will be to state our question, then collect or obtain the date, and then prepare it. The data is prepared by eliminating duplicates or non-numeric data, not doing this will cause errors in your results.

After we implement the first essential steps, we can give some insights about it, for example, if there is any pattern in the data? Then, we model and evaluate the data, for example, which technique we should use? Is the model accurate enough? Finally, we can provide a recommendation to stakeholders, however, is essential to be able to present our results in a manner that the general audience can understand and help in the decision-making process.

Nevertheless, we need to mention that the most important part of this task was to learn the Data Science Process and why is so important. Without guiding us with the DSP our work wont is efficient to provide results or the necessary answer. By analyzing the data we learned that not all the variables are necessary to provide a result, this can be solved by deleting some independent variables or by creating another set of data from the main one.

In addition, another important part is to learn to identify the independent and dependent variables. Without doing this our visualizations wont is useful for stakeholders or we won't even be able to use the data. Also, sometimes our data sets are very large and we will have to discretize our data to be able to use it.

In summary, we learned the steps from the DSP and why is important to know it. Also, we were able to apply what we learned in the previous curse and applied on a different data set.

In addition, on task #2 we were able to drop some features by following the provided rules. Bill 1-Bill 6 have a correlation of 0.90 or more between each other. By following rule #2 we dropped Bill 4,5,and 6 for having the lowest correlation with the dependent variable, DEFAULT. On task #3 we worked with the OOB data set and with dfcor, which is the the previously created data set on task #2. Moreover, the Random Forest Classifier was chosen as the selected model to tune and validate with the test based on the scores obtained from the 3 models tested.

Finally, after reviewing our jupyter notebook and based on the data obtained, we cannot ensure that customers will pay their loans since we cannot control the customer spending habits. However, the data obtained can provide us some useful information, even if we cannot specifically answer our problem question.