|  |  |  |
| --- | --- | --- |
|  |  |  |
| Ramses Meza  April 2021  <https://github.com/lordtable/UTDSGC_C2T2> |  | Data Analytics Course 2: Data Science with Python: Lessons Learned Report |

Lesson learned # 1: do not under-estimate the time spent on data preparation

Given my current level of Data Science expertise, this dataset was particularly challenging in terms of cleaning and preparation. Some data editing tasks can be much more time consuming than anticipated, so ample time needs to be allocated during project planning.

Lesson learned # 2: EDA Triage: early rule-out of variables

We are interested on being able to predict customer clients’ default likelihood, so considering the proxy Default status variable as the target or dependent variable, it is possible, via Exploratory Data Analysis (EDA), to make an early identification of those variables that are unlikely to be major drivers on any future predicting model. These variables are: Education, Sex, Marriage status, and Bill statements (April-Sept 2005) amount. I would expect that these variables would not be selected by a Machine Learning algorithm during a model calibration.

Lesson learned # 3: Inter-dependency (collinearity) amongst some predictors

There are several independent variables or predictors that are inter-dependent to each other. If many of these are highly correlated, then data feature reduction schemes (such as PCA) may need to be applied in order to reduce data dimensionality and improve the performance of any Machine Learning predicting model. Some of the likely inter-dependent variables include the PAY variable group, the BILL\_AMT variable group, the PAY\_AMT variable group, and also a sizeable inter-dependency between variables on the BILL\_AMT variable group and the PAY\_AMT variable group.

Lesson learned # 4: Missing variables

I wonder whether additional variables exist, but were not loaded to the SQL and made available for this project. For instance, thinking about the common information lenders ask borrowers, an *Income* variable may have come handy.

Lesson learned # 5: Feasibility of a predicting model(s)

Since we want to assess customers likelihood of defaulting on their credit lines, we want to create a Machine learning model that predicts such variable based on other multiple variables. The fact that there are some variables that moderately correlate with the Default Status (such as LIMIT\_BAL, the PAY variable group, and the PAY\_AMT variable group), then expectations are high that we should be able to build a predicting model capable of performing with a reasonably good accuracy. This is of the utmost importance for CreditOne in order to correctly triage default-prone customers and timely address the negative impact of the current defaulting customers on the company revenues.