***From:*** Gary Geng

T2C2

***To***: Guido Rossum

***Report 0003 – 12/2020***

Thank you for the opportunity to help on this project, please consider the following suggestions as a potential solution to the current problem:

1. One way to solve the increase in customer default rates would be to reevaluate if the credit should be issued and then, the available credit limit for these customers, these two steps can be accomplished by running two separate models and having the decision for each one of the questions for every current customer and potential ones.
2. It is important to notice that since the goal tacitly involves avoiding loss of customers and revenue for both the client and Credit One, some of the conditions may be softened but keeping in mind high risk customers should only represent a small portion of the total credits issued.

Regarding the investigative questions:

1. There is an opportunity to investigate if customers can and will pay their loans, since there is no direct way to influence in customer’s habits, it is needed to gather more information that allows to foresee if a customer can fulfill their obligations, such as income, employment situation, household income, debt to income ratio, liabilities, etc.

At the same time these are some insights from the data analysis performed:

1. After cleaning and transforming the data a total of 82 variables are being used
2. Not all the variables are used for each one of the models
3. Some of the variables (such as pay amount and bill amount, Education and pay history) are slightly correlated
4. Most of the variables (such as Age and pay amount, Age and Bill amount) have no correlation
5. Most of the features present a slight level of skewness
6. Most of the features show spikes in the histograms
7. The variables in need of transformation to “dummy variables” are: Sex, Education, Marital status, Age (Discretized), Payment History, Default status
8. Most of the variables related to amounts paid, billed or in statements present one or two data far off from the normal values causing heteroskedasticity
9. Some of the features might explain other variables causing autocorrelation, must use caution choosing which features will be part of the model.

For the lessons learned part, after a detailed exploratory analysis of the data available for Credit One, these are some valuable lessons learned from it

1. Not all the variables are useful to build a model for specific goals
2. There are more tools around the web to help speed up the process of building the EDA, but must be cautious when choosing which are the best ones for the project
3. There is more than one way to do things in data science
4. There are more rules to consider when cleaning and processing the data
5. Python can connect to almost any database
6. The importance of cleaning and preparing the data should never be underestimated