C5T3 - CreditOne Build and Evaluate

**CreditOne Report**

horizontal line

CreditOne’s primary goal is to obtain a better understanding of customer profiles in order to improve their credit approval process and reduce defaulting clients.

Following are the important features of the dataset provided for analysis and description of their impact on the defaulter finding process.

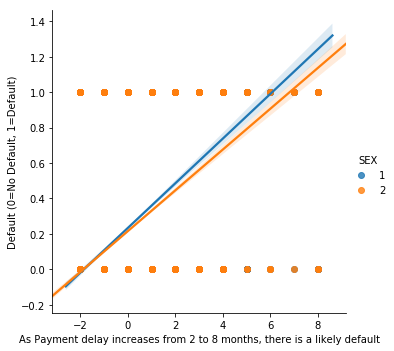
* **A clearly defined business problem**

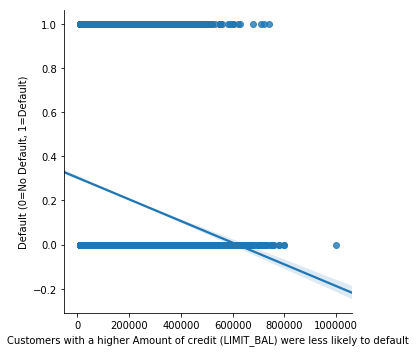
The first priority of this project was to understand if the distribution of the customer’s dataset supported a clear understanding of the business objective. The distribution of the dataset was uneven with only about 20% of the dataset containing information on defaulting customers and 80% of the dataset containing non-defaulters. The 80-20 distribution needed to be adjusted to make the dataset balanced for determining if there was a trend in the defaulting customer patterns. This introduced the first level for feature adjustment. Going forward having an equal distribution of the defaulters and non-defaulters would help in anticipating a pattern better.

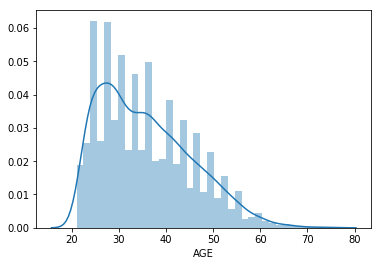


* **Identification of performance features to serve as proper metrics for default detection**

Some features in the provided dataset proved to be of higher significance to the end result of whether the customer will default, than other features. Past payment behavior, credit limit and age group of the customers were more significant than the actual bill and payment amounts. Other demographics such as education and marriage only marginally affected the results. Going forward it would be helpful to find out if there are any other relevant customer features that CreditOne may have that can be included in the dataset.







* **Behavior of features – to - results of Interest**

The defaulters and non-defaulters were coded as 1 and 0 respectively. This did not correlate with the payment amounts for those customers consistently. Customers with complete payments in the past 6 months also were found coded as defaulting. This could be an indicator of incomplete data, and that the behavior of the features doesn’t always support the result of interest. A deeper dive into learning the feature definitions is needed to identify features of interest. For example -

o Some clients who Paid in Full in the most recent month, were coded as default.

o Some clients with negative Bill Amount were coded as default.

o Some clients have Bill Amounts higher than their Credit limit balance.

* **Sufficient Data**

We used the "supervised learning” approach for solving this problem. This means that, we utilized a set of historical data to complete our analysis. Each record in this data set consisted of attributes of the individuals under analysis, and a 'desired output' attribute that corresponded to the behavior supporting our business objective.

The desired output variable was a "1" or "0" -- a binary representation distinguishing between individuals who display the behavior and those who do not.

The algorithms utilized to develop our models searched the available candidate attributes and developed a mathematical formula. Our models effectively are a scoring system indicating the likelihood of that individual to default or not.

Without sufficient data, it is impossible for us to make use of these techniques. With sufficient data quantity and quality, we can improve our “mathematical formulas” to do a better job to solve the business problem.

From the insights that we gained here, we are able to propose to management a few things:

* To lower the risk of default, we need to take action on delayed payments sooner than later.
* Clients in the 25 - 45 age group should be closely monitored for payment delays.
* Improve client screening among clients with lower credit limits.
* Communicate with data collection team that we need a well distributed dataset of defaulting and non-defaulting clients to improve our analysis.
* Ensure proper documentation of attribute description and possibly maintain a data dictionary. Having clearly defined business rules to classify defaulters and non-defaulters will help identify the key attributes and help develop an accurate dataset.
* Have a process in place to check data errors, caused during data collection.