Customer Default Identification Report

This document will set out to answer the two primary questions involved with this project. Those questions are:

How do you ensure that customers can/will pay their loans?

Can we approve customers with high certainty?

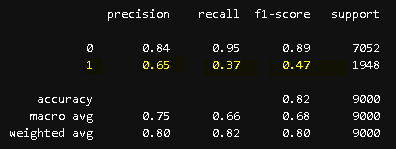
The answers to these questions will be informed by utilizing visualizations, commenting on important performance metrics for the chosen ML tools, and restating problems and areas of improvement for the dataset.

The answer to the first question is a resounding no. There is absolutely nothing about the data, nor are there any ML-related tools (algorithms or models) with which would result in the control of customer behavior. Unfortunately, given the dataset, there are no clear signs, trends or exploitable patterns present with which to determine, with any reasonable amount of certainty or confidence, that customers can be made or encouraged to pay off their loans in a timely manner. The question of enforcement cannot be answered by Data Analytics or Data Science principles.

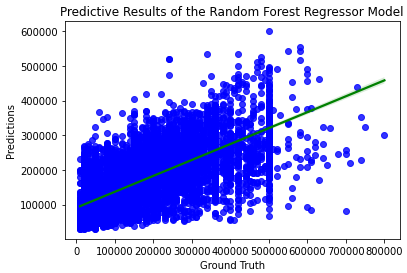
The answer to the second question is also a resounding no, however there are potential workarounds for this issue. Such workarounds mostly consist with the quality of the data and the scope of the data. I will speak about these two factors later in the document.

To substantiate my conclusions above, I will now provide commentary on the ML portion of this project. That is, I will be outlining why the current state of the data, and its scope, is currently ill-suited to be utilized with ML-related tools.

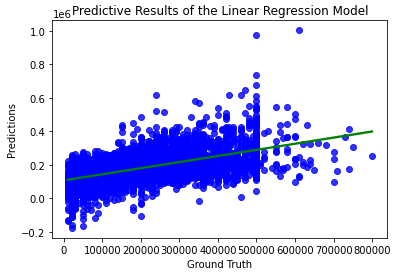
The three chosen ML-related tools for this project did not perform in a way that would reasonably result in a “yes” for the two aforementioned questions. For this, I will provide commentaries about some visualizations below:



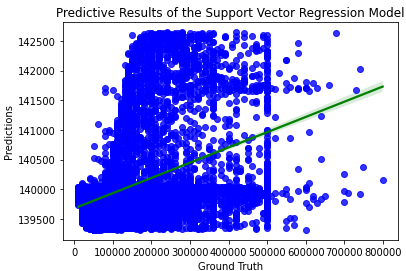
Classification report for predicting whether or not a customer will default on their loan. Unsurprisingly the “0” value is high, which refers to non-default loans. As good as that is, Credit One’s primary task is understanding and mitigating the patterns of loan defaults (which refers to “1”). Unfortunately, accuracy is low for predicting whether or not a customer will default on a loan.



This graph displays the predictive performance of one of the ML tools used in determining the amount of credit a customer should be given. The green line is used to help visualize how well Random Forest Regressor Model performed. Given how much variation there is, this model did not perform well.



This second graph displays the predictive performance of one of the ML tools used in determining the amount of credit a customer should be given. The green line is used to help visualize how well Linear Regression Model performed. While this model had considerably less variation, there still is enough variation that warrants questioning this model’s efficacy in predicting (determining) how much a customer should be given in credit.



This third graph displays the predictive performance of one of the ML tools used in determining the amount of credit a customer should be given. The green line is used to help visualize how well Support Vector Regression Model performed. In great contrast to the other models, this model performed the worst. It has a massive discrepancy between its predictions and the actual credit limits featured in the dataset (“Ground Truth”). This model clearly can be disregarded for further use in the scope of this project.

While the tools did not perform reasonably well and did not generate outcomes consistent with the desires of Credit One’s primary goal --- these outcomes should be considered a “learning moment” for the company and how it conducts its data management processes. The quality (performance and functionality) of ML tools is wholly dependent on the quality of the data with which is given to the models. If Credit One wants to eventually be able to have more reassuring answers to the aforementioned questions, its data must be expanded and improved upon. Some areas of improvement for data quality:

1. Relevant data (better approximations of credit history + longer period of time of past credit payments (April to September, or 6 months, is not enough time to ascertain clear patterns or make generalizations about patterns) + more representation of defaults in the data (or, have less of a overwhelming skew towards non-defaults, as any major skewedness in the data may present challenges for ML tools)
2. Understanding the limits and scope of the data. In this case, none of the variables had a strong correlation with the variable we were trying to predict (“loan\_result\_default”). What this means is that, in its current form, the data is incomplete or unable to lead to the insights and conclusions that Credit One was hoping for.

Keeping in mind the two aforementioned considerations, and improving the quality of the data and better understanding the limits and scope of the data and the core business problem, Credit One can develop better solutions to resolving its challenges. Whether or not ML would play a part in that is contingent on the implications of this report. I strongly recommended additional data collection, data cleaning and further EDA so as to inform the necessary groundwork for better ML outcomes.