Credit One helps companies signed up as partners decide whether to offer a loan to prospective clients based on the probability that client will repay back their loan. If Credit One judges a person to not be credit worthy, then the partner will not offer the prospective client a loan of a certain amount. For the past year, Credit One’s partners have reported an increase in clients defaulting on their loans, and if this trend continues, Credit One is at risk of losing its business partners. Therefore, Credit One’s data team has been enlisted to build a better model that predicts if a prospective client will default on their requested loan or not. We will be using historical data of current customers who have either payed back their loans or defaulted. At the end of the data science process, we hope to discover useful insights of a client’s ability to pay back their loans based on their historical and demographical characteristics. We will also have a working classifying model that will be able to predict with higher certainty the predicted client’s outcome behavior.

The historical data provided included a client’s balance limit, sex, education, marriage status, age, 6-month repayment status, 6-month bill and payment amounts and whether or not they defaulted on their loans. After importing the data and understanding its contents, I began exploring the data, and gained useful initial information. Out of the 30,000 clients observed, only 6,636 defaulted, or about 22%; the bar plot below illustrates this observation.



Figure : Count of Customers that Defaulted (1) & Repaid their Loans (0)

This shows that most of the customers do not default, however, our job as Credit One is to make sure that the few clients that do default are caught before they do so; this measurement will be a metric we will use to evaluate our models later on, it will be the classifier’s recall score. However, we do not want to classify clients that would pay back their loans as defaulters, because this would mean a loss of potential clients; the ability of the classifier to correctly distinguish between clients that will default from those that won’t will be its precision score. In order to take into account both of those scores equally, we will use the F1 scoring metric to determine which classifier best performs the job of predicting what customers will default and which ones will pay back their loans. Additional insights were gained from the initial exploration of the data. On average males defaulted 24% of the time compared to their counter female clients. Graduate school educated clients defaulted the least based on a client’s education level by only defaulting 19% of the time. On the other hand, university and high school educated clients defaulted 24% and 25% of the time, respectively. Based on a client’s marriage status, 21% of single clients, 23% of married clients and 26% of divorced clients defaulted on their loans. Half of the clients analyzed in the data were under 34, with the youngest and oldest clients being 21 and 79 respectively. There was no clear relationship between a client’s age and their ability to repay back their loans. After running a correlation matrix on the dataset, there was a clear correlation between the values of the past 6 months for the repayment status, bill amount and payment amount of the clients; therefore, I created three new columns, each corresponding to the summation of the 6-month values of the clients and named them ‘pay’, ‘bill\_amt’ and ‘pay\_amt’. The pay column had a strong positive relationship with a client’s ability to repay. As the value of the pay column increased so did the client’s probability of default. Below is the linear modeling plot visualizing the relationship.

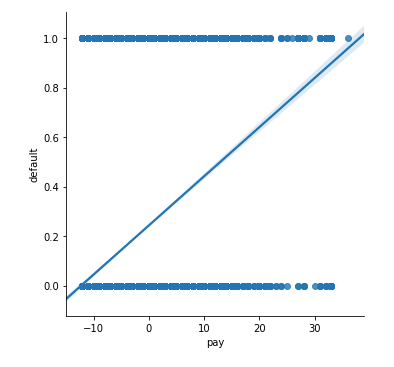


Figure : Repayment Status vs. Defaulting Probability

After visualizing this relationship, I created a new column named the ‘limit\_balance\_ratio’. This column was created by subtracting the ‘pay\_amt’ from the ‘bill\_amt’ column and dividing it by the limit balance column. This effectively described what fraction of their total balance limit they had outstanding. In order to check if this new feature helped predict the probability of default, I plotted it on a linear modeling plot like the one above and it showed the same relationship; as the client reached their limit balance, the probability that they would default also increased. The limit balance feature was analyzed, once again, by plotting it on a linear modeling plot against the default feature. The resulting relationship was opposite to the previous two observed; as the balance limit allowed to a client increased, the client’s probability of default decreased. This makes sense, since clients that are allowed a higher balance usually are more responsible and credit worthy than their low balance limit counterpart. Based on this exploratory analysis of the data, a classifier model was created in order to aid Credit One with assessing a prospective client’s ability to repay their loans.

In order to choose a classifier that best fit Credit One’s problem, we tested and evaluated three models with the given data; we built a Decision Tree, Random Forest and a Gaussian Naïve Bayes classifier to predict whether a client would default or not. We optimized each classifier’s performance by tuning each classifier’s hyper parameters. After doing so, each model was evaluated using F1 as the performance metric; the higher the model’s F1 score is, the better the model was at classifying customers as defaulting clients or not. The best F1 scores of the decision tree, random forest and gaussian naïve Bayes classifier were 0.411, 0.415 and 0.065, respectively. The highest scoring classifier was concluded to be the Random Forest Classifier with a ‘max\_depth’ of 125, ‘min\_samples\_split’ of 200 and ‘n\_estimators’ of 100. We then ran this classifier on the testing data, and it was able to achieve precision, recall and F1 scores of 0.62, 0.296 and 0.4, respectively; i.e. this RF classifier was able to prevent about 30% of clients that would have defaulted on their loans. The recall score tells us that out of all the clients the classifier predicted would default, about 60% actually did end up defaulting. Finally, the F1 score took both of those scores and combined them into a single score of F1.

In conclusion, there are recommendations that can be made based on the insights gained from the exploration of the data. While there were a few correlations between the demographics of a client and their ability to repay back their loans, the strongest relationships came from analyzing their historical behavior with previous loans. The closer a person got to their balance limit, the higher their probability of them defaulting on their loan became. In order to discourage future clients from leaving too small of a margin between their balance and limit, it can be recommended to our clients that a higher interest rate be applied to any amount spent by a client after the client reaches 75% of their balance limit. Also, the repayment status of a client proved to indicate whether a client would default on their loan; the more late and missing payments a client has, the higher the probability of that client defaulting is. Based on these strong relationships, before a client is approved for a loan, they will have to provide their past loan behavior; if they do not have previous loans, then that client will be offered a lower loan amount with higher interest and be placed under stricter supervision of their repayment status. The classifier did a good job of helping Credit One identify customers that would’ve otherwise defaulted on their loans. In the future, the data will be analyzed and the classifier finer tuned with the goal of increasing its recall, precision and ultimately F1 scores.