Report

The objective of the data analysis was to develop a model that could predict the delinquency of people that apply for a credit, based on data previously collected. First, the database was cleaned, by removing the observations with no complete information, and eliminating the outliers. Additionally, the number of features was reduced in order to avoid redundancy and a phenomenon called overfitting, that occurs when we use more information than the information needed. Finally, the data was divided in two groups: 70% of the data to train the models and make predictions and 30% to test the performance of our models and chose the best option.

In this case we experiment with 6 different types of methods and several parameters that produced 60 different options, and for every model I computed 18 different metrics of performance. Based on these results we can chose the best predictor and take some other important decisions, such as the size of the sample for future predictions or the kind of recommendations and interventions that we can make with the information produced.

The summary of the results is the following:

- Accuracy is a metric that shows the percentage of correct predictions (positive or negative) of a model with respect to the total of predictions. In this case a positive prediction means delinquency and a negative prediction means not delinquency.
- The model that present the highest accuracy is Boosting method with 75 estimators.
- Precision reports the rate of correct positive predictions with respect to the total of positive predictions. In this case the model with the best performance according to the precision metric was Random Forest method with 75 trees, gini criterion and max_features parameter using log2.
- Recall measure percentage of correct positive predictions with respect to the sum of correct positive prediction and incorrect negative predictions (which is generally known as the “relevant cases”). In our analysis the model with the highest recall is Bagging method with 75 estimators and max-features = 2.
- F1 score is the harmonic mean of precision and recall, that reach its highest value when precision and recall are both 1. In this case the model with the best F1 score is Random Forest with 75 estimators, gini criterion and max_features parameter using log2.

On addition, our analysis includes precision and recall metrics using different proportions of the data: 1%, 2%, 5%, 10%, 20%, 30% and 50%.

Our goal is to predict credit delinquency but at the same time as policy makers our goal is to benefit people, we want to minimize the false positive cases, that are the cases when delinquency is predicted incorrectly. This mistake implies to deny a credit to a person that would pay the credit. Because of that, we are more interested in chosen the model with the highest Precision.

Particularly, the highest precision is reached at the top 1%, using the model Random Forest with n_stimators = 75 and entropy criterion and max_features = auto.

To conclude, I would recommend using the latter model in order to minimize false positive cases and at the same time get an accuracy over 80%. In general, random forest method shows a good performance in this case.