Predicting "Financial Distress"

Pedro Armengol

The idea of this exercise was to generate high performance predictions about "Financial Distress" suffering in a sample of almost 300k individuals. In order to do so, we estimated several models[1] with a set of different parameters for each model. After those estimations, the results where gathers in a table (results_df.csv) such that comparisons between models can be made.

Basically, we generate the predictors of the models through the discretization of the 3 variables that had a higher correlation with "Financial Distress": "Number of times borrower has been 30-59 days past due but no worse in the last 2 years", "Number of times borrower has been 90 days or more past due" and "Number of times borrower has been 60-89 days past due but no worse in the last 2 years". "Age" was also included in the estimations but was not discretized.

After selecting the predictors, 8 different models where ran with different sets of parameters. From those estimations, we got three efficiency metrics: the area under the curve of a true positive rate/false positive rate plot (AUC-ROC), the precision of the predictions under different thresholds and the graph of the precision-recall trade-off.

The following where the results of the exercise:

1) Area under the curve: Gradient Boosting is the model that ranked first in this metric with an area of 0.81. This metric is a signaling of a good trade off between true positive rate and false positive rate: comparatively the model can afford to gain precision without excessively recall sacrifice.
2) Precision: Gradient Boosting classifier was also the model with the better precision estimations in the sample. Using the average of the three thresholds: 5%,10%,20%[2], the model gives a 67.9% of precision in almost all of its configurations. However, using a threshold of just 5%, the model predicts with 100% of precision.
3) Other models, that also perform relatively well on this exercise where Random Forest and Logistic regression.

In this exercise, we generate predictors, ran models and compared accuracy metrics. Several models performed very well, however, Gradient Boosting was the most effective for predicting "Financial Distress".

---

[1] Random Forest , ABA Boost, Logistic Regression, Supporting Vector Machine, Gradient Boosting, Decision Tree, K Neighbors and Bagging.

[2] The threshold mean imputing a "Financial Distress" = 1 the first n% individuals with bigger scores (the model outcomes).