Machine Learning for Public Policy

Assignment 3: Improving the Pipeline

Kyle Schindl

In order to predict if a project on DonorsChoose will get fully funded within 60 days we investigate a number of machine learning models. Because different models will have different trade-offs it is important to get a holistic view of the tools available to us. For the purposes of this project we will consider Random Forests, Logistic Regression, Ada/Gradient Boosting, Decision Trees, K-Nearest Neighbors, and Bagging (implemented with Decision Trees) - notably, we will be iterating over every combination of different parameter values for each model in order to determine what is most effective for prediction.

Determining which models are most effective can be a difficult task in itself. Thus, it is important to establish a set of evaluation metrics to judge the models by. For this project we will focus on accuracy, precision, recall, and the F1-score. While there are other metrics that can be used, they become increasingly abstract and difficult to communicate. To the uninitiated, we may think of accuracy as the percentage correct overall, precision as the percentage of correct positive predictions (or, how many predictions were relevant), and recall as the percentage of relevant results.

While different in many respects, we find that overall our models deliver many similar results. Models such as Decision Trees score the highest precision over time (and different thresholds), although it is often only a few decimal points over others (.05 or less). Overall, we find that models perform the best on our evaluation metrics in the third and final period. This is likely attributable to the larger amount of training data that they have to work with. We find more variation in scores such as accuracy, where differences in magnitude grow quite large (>

0.20) - however, this seems largely attributable to the threshold in place. On different thresholds we are clearly able to see a precision-recall tradeoff. Over time we see small changes to metrics such a precision, however it is not large. Part of the lack of changes may be attributable to feature selection.

If we are looking to identify the 5% of posted projects at the highest risk of not getting fully funded, there are a few metrics we may consider. However, before any analysis we should sort our predicted probabilities in order to find those most at risk. This will not affect our evaluation metrics, but it will allows us to identify those most in need at a given threshold. One metric that we are interested in is a model with a high degree of precision - this will guarantee our interventions are not being wasted on false positives. At the 5% level we find that there are many models with precision of 1.0, so we further sort by the AUC-ROC to determine an appropriate model. We find a Decision Tree in the third and final time period is most accurate at the 5% level.

Interestingly, we find that while Decision Trees score the highest both for precision and recall (and therefore the F1-score), Gradient Boosting and Random Forests score highest for the AUC-ROC. Notably, both of these classifiers rely on Decision Trees for their implementation. It is possible that this type of classifier is especially useful for the data we have, given that there are a large number of binary variables.