Claire Herdeman
ML Assignment 3
May 21, 2018

Donor's Choose Model Evaluation

With increasing cuts to school funding, teachers are turning to crowdfunding websites like Donors Choose to provide necessities and educational tools for their classrooms. It is important for as any of these projects as possible to be funded, especially considering that 80% of 2013 campaigns came from classrooms categorized as either "high poverty" or "highest poverty". The goal of this project is to identify campaigns that are in danger of failing to reach full funding in order to enable interventions that can improve the campaign's chance of success.

To identify projects that may not reach full funding, we used a selection of machine learning classification models trained with a variety of parameters over a series of 6 and 12 time periods. Using a variety of evaluation techniques, this strategy allows us to identify the models and time window that best predict at-risk campaigns.

For this analysis we trained 7 model types: k-nearest neighbors, decision trees, and logistic regression were simple classifiers, while random forests, bagged logistic regression, boosted decision trees, and gradient boosting were all ensemble models, or models that are a composition of simple classifiers. To evaluate these models, we first established a baseline performance metric. This baseline is simply the proportion of all campaigns in the test set that fail to reach their funding goal, which was consistently around 30% over all time periods. This helps us determine whether a given model is performing better than assigning all observations to a single category. For instance, if our accuracy score, i.e. the fraction of campaigns that our model correctly classifies, is 70%, the model may simply be classifying all campaigns as having reached their funding goal while misclassifying all of the campaigns we want to identify.

Among the model types tested, the gradient boosting model consistently performed the best. With a threshold of 50%, the gradient boosting model reached a peak accuracy of 0.807 when trained over a 30-month period and tested on a 6-month period. Precision at 50% for this model, or the proportion of all campaigns categorized as failing to meet their goal that are correctly classified, is 0.61, while recall at 50%, the proportion of all campaigns that failed to reach their goal that were correctly classified, reached 1.0. This means that no campaigns that failed to reach full funding were categorized as having reached full funding. It is important to note, however, that the AUC-ROC score is only 0.5. The AUC-ROC score is a measure of how much new information was learned by the model, and a score of 0.5 indicates that the model may not be much better than a random guess.

With all of that in mind, additional testing and modeling should be done before any of these predictions are used as a basis for intervention. While the accuracy, precision, and recall metrics are the best of the set of models trained and tested, they do not represent a sufficiently significant improvement over the baseline. For instance, while accuracy reaches 80%, baseline accuracy is already 70%. To improve these models for possible deployment, the models would first need to be trained over a larger set of parameters, then tested over smaller time increments. The best performing model was tested on the last 6 months and trained on the 30 months prior, so decreasing the final testing window to 1 or 3 months and training on all previous available data may further improve predictive results.