HOMEWORK 3 – POLICY MEMO WRITE-UP

Allison Collins

BACKGROUND & GOALS:

The goal is to allocate interventions for DonorsChoose projects by predicting whether or not a given project will be funded within a 60-day period. I tested a variety of models on a dataset spanning the year 2012. A separate Jupyter Notebook houses the actual analysis and outputs for running each of the models; this write-up serves to primarily discuss policy implications rather than technical explanations.

MODEL GENERATION & EVALUATION:

As can be seen in the output charts by model type (for reference in the notebook), these models did not have exceptionally high performance among the evaluation criteria considered (accuracy, precision, AUC, F1, recall). To give one brief example, the highest score for area under the curve exceeded .6 on the fitted models, but remains in the .5-.6 range in the dummy classifier (this looks at the model's ability to properly classify true positives across different thresholds – at random, we would expect .5 from a random classifier, that is it would get it right half of the time). It follows that if we compare to the baseline using a dummy classifier (as illustrated in the Jupyter notebook), we can see that the strongest performing models among the different types will outperform this, but overall they are not performing at strong values of the evaluation metrics. Moreover, we can also see that for some cases the model was classifying all things as not getting funded for instance (where there were no predicted positive outcomes).

One exception is overall, there were more models across type that fared better on recall. Additionally, at higher levels of the threshold, most of the models performed better on accuracy. The SVM had accuracy consistently hovering around ~.8, which a number of the other models attained, but not across all iterations of date and different parameters. Across the board, the models fared poorly on precision.

When investigating this general poor performance (compared to what we have learned to be "good" scores), resources suggested that feature selection is key, but a model with the right feature selection when stacked against a different model (e.g. tree + random forest) using the same features – on the whole, recognizing models fit different problems/datasets better – will perform more similarly than the same model type (e.g. decision tree) using a bad set of features and a good set. Notably here, the project's total price is a strong feature, which has implications when considering how to allocate among projects.

One implication of the above two points, meaning given the fact that the models' performance across the different models (including over the 6-month time horizons evaluated) was relatively similar, is that as policymakers we could consider picking a more understandable model to use. In some cases, we may be faced with the tradeoff between a more understandable model and one that better predicts outcome,

but here for instance we could consider using a decision tree that can be clearly followed and explained given the best one's performance (looking at highest area under the curve, FI (combined precision/recall) and accuracy – which is not the best measure) is not very different from other models.

PRIORITIZATION OF PROJECTS FOR INTERVENTION:

The choice of which model should be selected to inform predictions given that only 5% of the projects can be intervened on greatly depends on the priorities of those who are planning the intervention. The models suggest ways to predict which projects are predicted to be fully funded in the 60-day time period, and which are not. However, deciding how to spend funding requires answering several strategic questions, which I will tee up and present ways that we could think about making choices, were we to have more information on how to answer these:

- What is the policymaker's priority – is it to provide assistance to the projects that have the least chance of receiving funding, or to "push projects over the edge" which may get close?
- How is equity being incorporated? Is there consideration being given to whether there are certain features of a school (for instance, being low resource or high poverty) that will additionally be considered
- How will resources be split and what is the funding amount for these interventions?
- Are we more concerned with missing projects (e.g. we are afraid of false positives, where here a positive 1 means getting funding) or with avoiding funding projects which actually would be likely to be successful on its own (here, a false negative)?

With a deeper understanding of the specific goals at hand, we could take specific steps to identify the optimal 5% of projects to support

- If it relates to projects at the margin vs. lowest probability, we could use the specific probability scores being output vs. the binary classification to allocate funding to projects at the margin (this could apply across models)
- If we are concerned with maximizing the correct prediction of true positives (here, correctly predicting the highest amount of positive, projects being funded), to make sure we are not missing out, then we should pick the model with the highest recall
- If we are concerned with the highest amount of positives right, we should pick the model with highest precision

In the case of this dataset in particular, we must consider that the project's total price is a key feature, and whether while that is a strong predictor, it is the way in which we want to structure our intervention (again returning to the question of priorities and equity).