Donor's Choose - Predicting Full Funding

May 21, 2018

1 Donor's Choose: Predicting Full Funding

1.1 Summary of the data and problem

1.1.1 Project goals

Donor's Choose is an online platform that helps teachers obtain funding for school projects. We sought to predict which Donor's Choose projects would be fully-funded at time of posting. In layman's terms, given what we know about a project when it is posted (e.g., what subject area it covers, how much it will cost), we wanted to predict whether it would not be fully funded. Donor's Choose projects tend to be successful at raising full funding. In fact, over 70% of all Donor's Choose projects are fully-funded. We thus want our predictions to correctly assign the probability of unsuccessful funding better than 30% of the time (that is, better than if we just believed from the outset that every project would get full funding). Predicting which projects are unlikely to receive full funding at time of posting will accomplish two aims: first, Donor's Choose will be able to share data-validated tips for teachers to help them achieve full funding based on what our model suggests is important, and second, Donor's Choose will be able to pre-emptively identify which projects are unlikely to get funding, and thus prioritize and improve those project proposals

1.1.2 Explanation of the data

Data provided for this project included demographic data for each project submitted to Donor's Choose as well as all project outcomes. We were specifically interested in whether or not a given project would be funded and wanted to consider which factors could impact full funding. As is the case with most public data, inevitably, some data was missing and some data was not in the right format for our analysis. We filled in missing data with median values and transformed potential variables in several ways in order to best predict our outcome variable.

1.2 Summary of what is most important from a prediction standpoint

1.2.1 Feature selection

Feature selection is a fancy term for determining which variables are most important to our model from a prediction standpoint. We need to limit the number of predictors in our model in order to ensure that the model is consistent over time (that is, that some combination of variables is extremely predictive in a six-month window, but not in the six following months).

1.2.2 Features used in this model

In this model, we included several important features. The descriptors below indicate which categories and groups within each category were most relevant from a prediction standpoint. Note, the number of predictors included could be significantly decreased with limited impact on the model; having fewer variables to keep track of and prioritize from a policy standpoint does show promise.

- School location: metrosuburban, metrorural, longitude, latitude
- Total price of the project
- Number of students reached
- Fulfillment labor and materials
- Resource type is technology or books
- Student grade: 3-5, preK-2, 6-8
- Eligible for double your impact match
- Teacher is from Teach For America
- Teacher has prefix Mrs.
- School is high poverty
- Project focuses on Literacy and Language

1.3 Summary of prediction models

1.3.1 What makes a model good?

When we consider model "goodness", we consider a few key metrics: - **Precision** tells us, of projects we have identified as likely to get funded, how many will actually be funded? - **Recall** tells us, of the projects that are fully funded, how many have we predicted will get full funding? - **ROC - AUC** tells us, how well did we rank our space? (baseline is 0.50) - **Accuracy** tells us, how many labels did we get right overall? (because about 71% of projects do get funded, accuracy of labelling all projects as "funded" would in fact be 71%)

1.3.2 Which models are best?

Ultimately, we care most about precision and recall, and we are most interested in models that improve over time, rather than models that get worse. This is because over time, Donor's Choose will have more and more data, and so models that do worse the models do with increased data, the worse they will do over time. 1. First, we eliminate models with wildly different outcomes over time. This means we eliminate aggregated models (bagging and boosting), which show markedly different performance over time. 2. Next, we prioritize models that identify which projects will be funded *best*, e.g., have the highest precision. This is because at Donor's Choose, there is a finite amount of resources to spend, and we want to ensure that these resources are prioritized best. We also want to be able to identify what about a project makes it likely to succeed.

We then can look more closely at our metrics

model	baseline	roc-auc	recall at 0.5	precision at 0.5
logistic regression	0.29	0.56	0.22	0.45
decision tree	0.29	0.56	0.27	0.42
random_forest	0.29	0.58	0.34	0.43

We can see here that, depending on what we prioritize, different models have different strengths. Logistic regression is overall best - it has low recall, high precision, and medium roc-auc. The decision tree is overall worst: it has medium recall, medium precision, and medium roc-auc. The random forest has the highest recall and roc-auc, but the low precision.

The logistic regression identifies the highest share of funded projects in the model. The random forest orders the space best (prioritizes project likelihood to funding the best) and has the highest rate of identifying projects as being funded if they truly were funded.

1.3.3 Conclusions

The sad reality, however, is that none of these models are overwhelmingly strong. Because the base rate of funding is so high, these models are performing only slightly above assuming that all projects were funded.

1.4 Outcomes and recommendations

1.4.1 Recommendations

- **Use models as decision aids, not as stand ins for decisions:** These models are not a replacement for human judgement and should be used alongside experts in the field.
- **Given the limitations of these models, use them to help predict which projects to offer additional support above baseline to**
- **Given the limitations of these models, do not use them to withdraw support from projects**
- **Continue to develop and refine these models over time so that they become increasingly powerful:** Consider including additional variables based on expert opinions as you refine these models.

1.4.2 Outcomes

Goal: Donor's Choose will be able to share data-validated tips for teachers to help them achieve full funding based on what our model suggests is important The outcome for this goal is that Donor's Choose can identify which variables are the absolute most important to securing full funding. Donor's Choose can help teachers position their projects to best align with the features our models showed were more important. For example, our models show that funding is higher for Teach For America teachers, so Donor's Choose can leverage the power of those classrooms and that branding for other teachers as well. One salient example: in many high-need schools, often Teach For America teachers serve as co-teachers for general education classrooms, and so leveraging both teachers in a Donor's Choose profile might be high impact. #### Goal: Donor's Choose will be able to pre-emptively identify which projects are unlikely to get funding, and thus prioritize and improve those project proposals The outcome for this goal is that Donor's Choose will indeed be able to identify projects to prioritize support for. The recall for this project was quite good (in other words, most of the projects that were funded were identified). This suggests that the left-out projects (those labeled as "not funded") are likely to have not been funded. This model will enable Donor's Choose to identify when to prioritize support for those projects.