TO: Supervisor
FROM: Lauren Li, Policy Analyst
DATE: May 2, 2019
RE: Identifying projects that need support attaining full funding

SUMMARY:

With the organization's new grant, we are able to provide support to 5% of the projects posted on donorschoose.org. To help identify which projects should be given supplemental support to reach their goal, our analysis predicts which projects will not be fully funded in the first 60 days. Given our limited resource constraints, we aim to target 5% of the total projects. With this in mind, I recommend using a decision tree model with one year of training data to identify the top 5% of projects at risk and intervene on those.

ANALYSIS:

After a comparison of numerous supervised learning models, two key findings emerge:

- Logistic regression and decision tree models perform the best overall with respect to key evaluation metrics.
- There is a notable increase in precision when using one year of data to train the model.

In this analysis, the features used to predict whether or not a project would be fully funded in 60 days were school type (charter, magnet, or neither), school location (urban or suburban), primary focus, resource type, grade level, students reached, total amount, poverty level, and eligibility for double your impact match. I use temporal validation to prevent using future projects to predict projects in the past (i.e. no projects from 2013 were used to predict on projects in 2012). Given our ability to support 5% of total projects, we'd like to direct resources to the projects most in need. I use precision as a measure for this. Recall, the ability to find all the projects that wouldn't be funded in 60 days, and AUC, which provides a measure of performance without regardless of threshold, are also included for comparison purposes.

When comparing models with time series data using precision and recall, the threshold and time period of data need to be taken into account. Models trained on a year of data rather than 6 months or 18 months have higher precision. It could be that too few months or too many months insert noise into the data set or donation trends change within those periods. When only trained on 6 months of data, it's likely that the data is not well separated which can make it better suited for linear models. As more history is used to train the model, the structure of the data possibly changes to be suited for nonlinear boundaries, which can be provided by decision trees. Over all time periods and thresholds (from 1% to 50% of total projects), logistic regression models and decision trees have the highest precision and recall. AUC does not depend on threshold, but there is an increase in AUC when using one year of data to train the model rather than only six months.

CAVEATS:

The data used in this analysis includes information on past projects and when they were posted and funded from the beginning of 2012 through the end of 2013. There was no further data on neighborhood/location characteristics included in the model other than what was provided in the dataset. Missing values in the data were imputed using the most frequent occurrence for categorical data and the median value for numerical data.

CONCLUSION:

This analysis shows that logistic regression and decision tree models have the highest precision across periods and thresholds. When comparing at the 5% threshold, a decision tree model trained on a year of data has the highest precision and has the benefit of being easily visualized and explained to other department leads. I recommend using a decision tree model trained on a year of data to predict whether a project will be fully funded within 60 days.