TO: Supervisor
FROM: Lauren Li, Policy Analyst
DATE: June 4, 2019
RE: Understanding differentiating characteristics of DonorsChoose.org projects

SUMMARY:

After identifying a model to predict the top 5% of projects that are at risk of not receiving funding, we distinguish the characteristics of these projects by clustering. We also examine characteristics of the full dataset covering 2012 to 2013. We found that the focus area and grade level were important characteristics in dividing groups, while indicators for charter schools, location, or type of area (suburban, urban, etc) were not as differentiating. The clusters were not meaningfully different for the overall dataset and the top 5% predictions.

ANALYSIS:

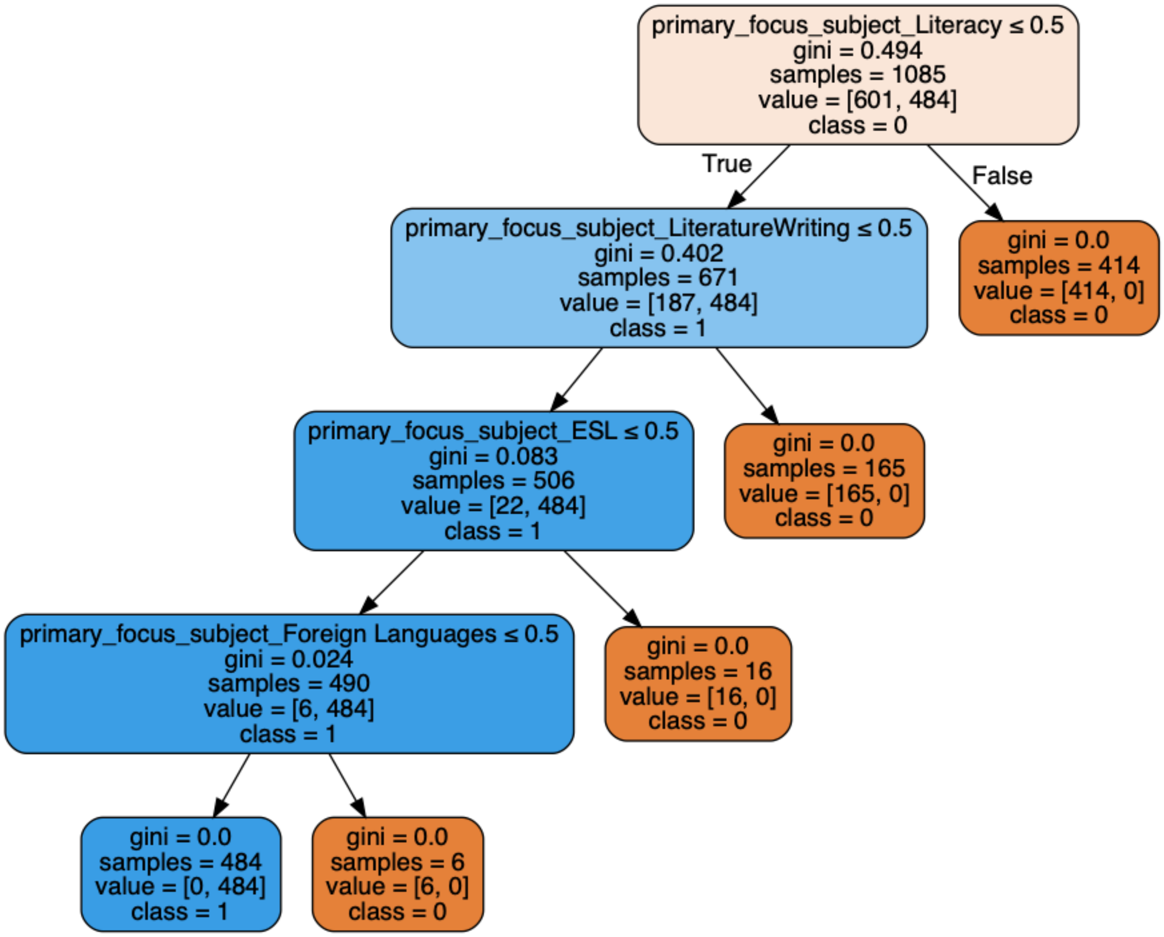
After a comparison of numerous supervised learning models, two key findings emerge for the overall dataset and the top 5% predictions:

- Location, type of area, and whether or not a school is a charter school are not differentiating characteristics.
- Primary focus area and grade level tend to be a differentiating characteristics among projects.

In this analysis, I used 2 clusters to distinguish the projects on characteristics such as school type (charter, magnet, or neither), school location (city, state, district, urban or suburban), primary focus, resource type, grade level, students reached, total amount, poverty level, and eligibility for double your impact match. Clusters were generated on the full dataset as well as the 5% of total projects generated previously.

Location, type of area, and charter school indicators were not distinguishing features. We saw similar frequencies of each across the two clusters for the overall dataset as well as the top 5% predictions. The most distinguishing feature for both datasets was the primary focus area of

Literacy and Language. Furthermore, other writing and language-related focused projects were also clustered together for both data sets. When removing Literacy and Language, the main features were still language-related. Below visualizes a decision tree on the top 5% predictions to illustrate the cluster features.



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.