# Policy Memo

Subject: Improving student outcomes and wellbeing by optimizing project proposal applications

## Policy Goal

DonorsChoose.org makes it easy for the general population to help students in need through school donations. Thousands of teachers in K-12 schools submit project proposals requesting material resources to enhance the education of their students, and proposals that hit their goal receive funding. In order help the most students receive the materials they need to learn, it is imperative that teachers develop the strongest proposals possible to meet their funding goals within a reasonable timeframe and that strong proposals are highlighted on DonorsChoose.org to maximize their potential reach.

To this end, the goal of this analysis is to compare the strength and performance of a variety of machine learning classifiers that predict if a project on DonorsChoose.org *will not* get funding within 60 days of posting. Understanding which types of proposals have historically led to, and not led to, success will enable policymakers and education advocates to validate their assumptions and identify projects unlikely of receiving funding within two months. In turn, advocates can intervene on specific proposals at an early stage to improve their potential for success, thereby improving the outcomes and wellbeing of students.

## Data

The cleaned data were obtained from DonorsChoose.org and contain records for 124,977 project proposals submitted between January 1, 2012 and December 31, 2013. Each record has a posting date and funding date, enabling the creation of a new binary outcome that indicates whether the proposal was not funded within 60 days of posting (value: 1) or funded within this timeframe (value: 0). Several variables were assessed for their explanatory and predictive power, including:

- School location: urban, suburban, or rural
- Charter school: yes or no
- Magnet school: yes or no
- Teacher prefix/gender: Dr., Mr., Mrs., or Ms.
- Poverty level: highest, high, moderate, or low
- Grade level: several
- Eligible double your impact match: yes or no
- School city/state/county/district: several
- Primary focus subject/area: several
- Second focus subject/area: several
- Resource type: several
- Total price: continuous
- Number of students reached: continuous

All variables listed above were included in the analysis. For categorical variables with over 20 options (such as School District), the top 20 most common responses within each training dataset were preserved and remaining categories were grouped as “Other”. Testing data were fit into these groups. The purpose of this modification was to improve model performance, as well as to focus the analysis on specific categories that are most likely to have a significant impact on the outcome variable. Categorical variables with fewer than 20 options were not modified. To capture additional nuances in the data, two continuous fields were included in the analysis: total price and number of students reached.

## Models & Assumptions

Several classification models were trained, tested, evaluated, and compared to understand their performance under different conditions and over different timeframes. This process helps to validate underlying assumptions about the data and determine which models are best to use under which conditions. The following models were trained and tested on the dataset:

- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machine
- Random Forest
- Boosting (AdaBoost and Gradient Boosting)
- Bagging

A rolling window of 6-months was used to identify testing data, and a 60-day gap was maintained between training and testing sets. This resulted in three separate testing timeframes:

- Test Set 1: July 1, 2013 to December, 31, 2013 (most records used for training)
- Test Set 2: December 31, 2012 to June 30, 2013 (moderate number of records used for training)
- Test Set 3: July 1, 2012 to December 30, 2012 (fewest records used for training)

## Model Performance & Implications

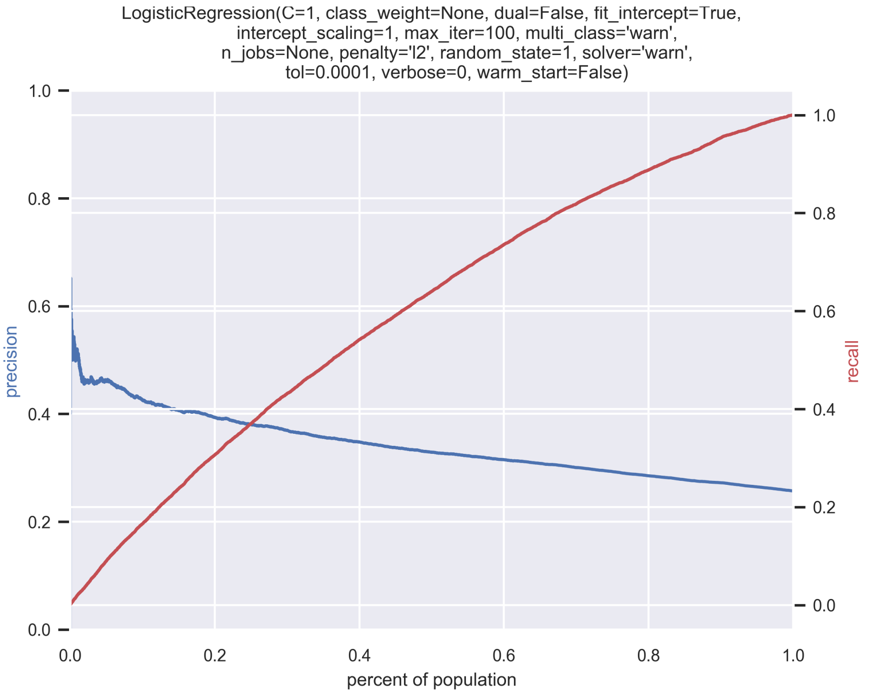
The results of the analysis indicate that the performance of classification models varies based on which data and time period they are trained on, as well as the metrics used to evaluate them. The metric of precision attempts to answer the question of: what proportion of proposals predicted to be unfunded within 60 days were actually correct? High precision can be seen as a measure of exactness or quality and indicates that an algorithm returned substantially more relevant results than irrelevant ones. When a population cutoff of 5% was applied to the models, indicating that 5% of proposals with the highest predicted probabilities of being unfunded within 60 days were classified as unfunded, the Logisitic Regression classifier consistently had the best performance with C=1 (score of 0.49 for Test Set 1). Meanwhile, the Bagging classifier performed the best with population cutoffs of 10% (score of 0.48 for Test Set 1) and 20% (score of 0.46 for Test Set 1).

Another important model performance metric is recall, which is a measure of completeness or quantity and demonstrates whether an algorithm returns most of the relevant results. In other words, high recall and low precision indicates that many proposals are predicted to be unfunded, whether or not they truly end up being unfunded. As with precision, the Logisitic Regression classifier consistently performed the best with a 5% population cutoff when C=1, achieving a score of 0.09, while the Bagging classifier with 10 estimators consistently performed the best at 10% and 20% population cutoffs, achieving scores of 0.17 and 0.32, respectively (Test Set 1). This trend is consistent for F1 score, which is a function of precision and recall.

Meanwhile, a metric known as Area Under the Curve (AUC) is one of the most important metrics for checking a classification model's performance, as it provides an aggregate measure of performance across all possible classification thresholds and tells us how much a model is capable of distinguishing between two outcomes. While the baseline model used had an AUC score of 0.5, which means that the model has no class separation capacity whatsoever, the Gradient Boosting model received the highest score for this metric across holdout sets. For instance, for Test Set 1, with parameters set to learning rate: 0.5, max_depth: 5, n_estimators: 10, and subsample: 0.5, and for Test Set 3, with parameters set to learning rate: 0.001, max_depth: 5, n_estimators: 100, and subsample: 0.5, Gradient Boosting received a score of 0.68, demonstrating a higher capability to distinguish between outcomes than the baseline.

Throughout the analysis, precision-recall graphs were generated to assess the relationship between these two metrics at different population cutoffs (k-values). These plots demonstrate that typically, precision decreases and recall increases as the percent of population cutoff increases. The plot below demonstrates this relationship:

**Figure 1: Logisitic Regression Precision-Recall Curves**



Overall, while precision, recall, and AUC all demonstrate fluctuations over time, generally, as the training dataset increased in size and a longer period of time was used to train the model to predict outcomes further into the future, performance scores across metrics improved. In other words, Test Set 1 typically outperformed Test Set 2 and Test Set 3. Specifically, precision for 10% of the population averaged around 0.4 for Test Set 1 (longest training, most recent testing), 0.3 for Test Set 2, and 0.2 for Test Set 3. This demonstrates the strength of utilizing a one-year period to train the models, followed by a 60-day gap and a 6-month testing period. The lowest performance metrics were achieved when the classification models were trained on fewer than six months of data, or 21,423 proposals, as compared to a maximum of 74,359 proposals.

## Recommendations

Given that an education advocacy nonprofit organization has resource availability to support the improvement of five percent of proposals, these models can be used to direct attention to those with the highest predicted likelihood of remaining unfunded for greater than two months, which can be considered to be at high risk. I therefore recommend that a Logistic Regression model be deployed using the following parameters: C=1 (default regularization), class weight=None, and penalty=L1. This specific model often outperformed the others across metrics, and specifically had the highest precision scores when the population cutoff was set to 5% during all three time periods (score of 0.49 for Test Set 1). Given the specific intervention proposed, I prioritize the evaluation metric of precision because it is important to minimize false positives over false negatives. In other words, since there are only resources available to support a small number of proposals, it is important that funds not be misdirected to proposals that are not actually at high risk. This specific model also had one of the highest AUC scores among models, 0.65, for Test Set 1, 0.63 for Test Set 2, and 0.64 for Test Set 3, as well as an accuracy score of 0.71 and F1 of 0.15.

A Logistic Regression model can be used to estimate the relationship between a binary dependent variable and independent variables and calculates the probability that a proposal is not funded within 60 days. Specifically, it can help to better understand if there are strong relationships between proposal characteristics and between these characteristics and the outcome. Finally, I propose further comparing this model to the Bagging classifier, with number of estimators set to 10. This model received the second highest precision score among models but the highest AUC, which may have greater implications if the percent of proposals that the nonprofit can support shifts away from 5% in the future. Finally, given fluctuations in scores over time, I recommend using no less than one year of data to train a model.