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 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 was obtained from DonorsChoose.org and contains 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 (but not limited to):

- School location: urban, suburban, or rural
- Charter school: yes or no
- Teacher prefix/gender: Dr., Mr., Mrs., or Ms.
- Primary focus subject: several
- Resource type: several
- Poverty level: highest, high, moderate, or low
- Number of students reached
- Total price

In order to optimize model performance and focus the analysis on characteristics of a proposal that provide the most information about the outcome, a feature selection methodology was employed to identify the top 10 most explanatory characteristics for predicting funding (eg. primary focus subject of technology or school location: urban). The analysis then proceeded using these narrowed down features.

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
- Bagging

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 threshold of 0.01 was set, meaning that if a proposal was predicted to have a 1% or higher probability of being unfunded within 60 days it was classified as unfunded, the K-Nearest Neighbors classifier consistently performed the best, with a maximum score of 0.33 when the final 6 months were predicted. However, the Random Forest model demonstrated a stronger performance than the K-Nearest Neighbors model when this threshold was increased to 30%, reaching a maximum score of 0.43 when the model was tested on a timeframe spanning from 1/1/203 to 6/30/2013.

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. It is interesting to note that at the highest threshold (50%), the Logistic Regression model consistently had the highest recall, reaching a maximum score of 0.63 while most other models had a score of zero or close to zero at this level. 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 Logistic Regression model received a higher score of 0.58, demonstrating a higher capacity to distinguish between outcomes.

Overall, it is interesting to note that across most models, precision appears increased up until the 30% level, after which it dramatically decreased. This may indicate a general tendency toward proposals being funded within 60 days across observations. Additionally, precision, recall, and AUC all demonstrate fluctuations over time. In general, 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 improved. However, there was a slight drop off during the final testing period, July 1, 2013 to December 31, 2013, indicating the strength of utilizing a one-year period to train the models with a 6-month testing period. The lowest performance metrics were achieved when the classification models were only trained on six months of data, or 23,740 proposals, as compared to a maximum of 71,379 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=1e-05 (strong regularization), class weight=balanced (using the outcome to automatically adjust weights), and penalty=L2. This specific model often outperformed the others across metrics, receiving an AUC score of 0.58, an F1 score of 0.44, an accuracy score of 0.58, and a particularly high precision score of 0.35 at the 50% threshold. 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, given fluctuations in scores over time, I recommend using no less than one year of data to train a given model.