# DonosChoose.org Analysis

Rose Gao

The goal is to “help DonorsChoose.org identify projects that are exceptionally exciting to the business, at the time of posting. While all projects on the site fulfill some kind of need, certain projects have a quality above and beyond what is typical. By identifying and recommending such projects early, they will improve funding outcomes, better the user experience, and help more students receive the materials they need to learn.” (Kaggle)

To achieve this goal, I decided to decompose the goal into two questions – what is the most predictive model overall, and which model is appropriate at each intervention level?

If there are no limits on intervention, then a simple decision tree classifier with a max depth of 50, a minimum samples split of 5, and calculating splits with a weighted Gini score is the most predictive model, with an accuracy of 1.00 and a precision of 0.91 at 100% of the population. While this model is very predictive at 100% of the population, it is unrealistic to assume limitless resources, and we must focus the resources on the projects that are most likely to succeed. In fact, this model only gives a result when considering 60% of population or more.

Furthermore, while a recall of 1.00 and a precision of 0.91 seems high, I need to compare my results with a baseline. In the 2011-2013 dataset, there are 247,698 projects that were fully funded, and 105,453 that were not fully funded out of a total of 353,151 projects. This yields a success rate, and a baseline of 70.14%. The precision of my decision tree classifier is indeed much higher than the baseline at 91%.

If we suspect that there is a shifting trend in projects being funded in the years 2011-2013, then it is possible that some models outperform the baseline during some time periods, but not during others. In my analysis, I created one model for every 12 months, and to compare my models’ performance, I needed to recalculate the baseline for the same 12 months. Since the data ranges from January 1st, 2011 to December 31st, 2013, I can simply look at the baselines for each calendar year. In 2011, there were 70,749 fully funded projects and 33,447 projects not fully funded out of 104,106 projects total, yielding a baseline of 67.80%. In 2012, there were 84,550 fully funded projects and 33,076 not funded projects out of 117,626 projects, yielding a baseline of 71.88%. Similarly, in 2013, there were 92,399 fully funded projects and 38,930 not funded projects out of 131,329 projects, yielding a baseline of 70.36%.

It’s interesting to note that while the success rates per year have increased 4% from 2011 to 2012 to 71.88%, and then decreased slightly to the 70% mark in 2013, the number of projects overall has grown steadily.

In comparison to these baselines, several of the classifier models outperformed. Simple logistic regression classifiers performed very well with an L1 penalty and C = 10 for both years, at the 1%, 2%, 5%, and 10%. In fact, at 1%, the logistic regression model achieves a precision of nearly 100%. In comparison, tree models like the decision tree and the random forest models did not perform as well at 1%, and quickly decreased in precision to ~85% at the 20% population mark, even when increasing max depth from 5 to 25. The K nearest neighbors model also yielded high precision of ~90% at 1% of the population when using 100 neighbors and the KD tree algorithm, but also declined to ~80% at the 20% population mark.

Due to simplicity and interpretability, I would recommend using a logistic regression model with an L1 penalty and C = 10. It yields the highest precision at every point from 1% to 50%, and a low recall (although most models had very similar recall curves).