Donor’s Choose: Improving Pipeline

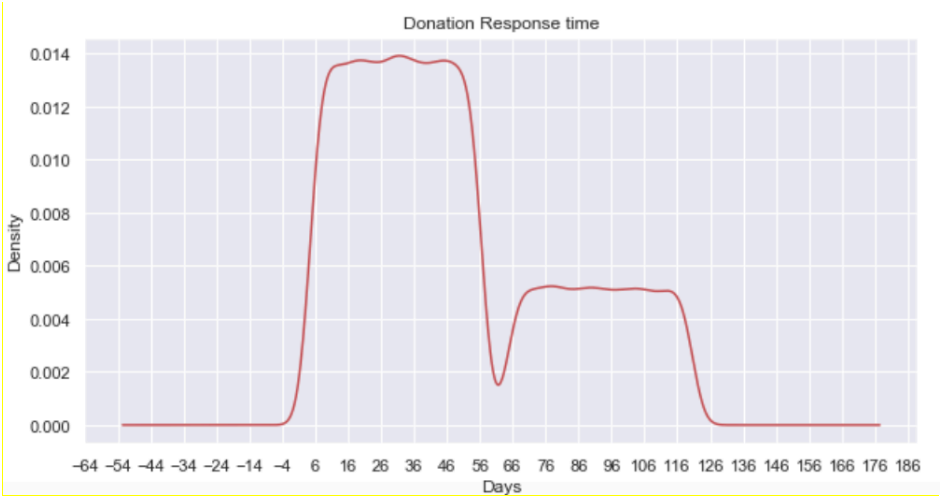
Summary: Based on the capacity constraint of reaching out to 5% within 60 days period, precision is most useful metric to consider. In other words, the model focuses on precisely identified the actually underfunded projects. L2 regularized Logistic Regression Classifier, with C =1 with performs best in this regard. The model gives sturdier prediction when trained on 1-year data and tested on 6 months window.

Background: Donors Choose is an online charity that helps schools with supplementary funding to help students with educational tools/materials. 124976 such projects identified 29947 schools between 2012-2013. First cut analysis of data reveals that, with average monthly funding ~$500 these projects were largely focused in a.) Urban areas; providing majorly b.) Supplies & Technology; concentrated largely on c.) Pre-K2 grades with d.) higher poverty. Despite the directed efforts there might be schools in direr conditions (/relatively more deserving) which might not be able to get to access to this opportunity. In an effort to further this process, the goal of this study is if a project will not get fully funded within 60 days of posting.

Data: Data available for the analysis cover characteristics about the school. Including, funding amount, school location, class grade, type of subjects, poverty level, type of resource provided etc. 59 missing values for students reached attributed and was treated by mean replacement. Ideally, given the right skewed distribution mean is not recommended, but the missing % is less than 0.001% of data Rest missing in categorical variable. Outliers in donation are plausible and have not been removed/normalised.

Features:

It is interesting to observe that most projects in sample data are funded within 60 days period. Inferring from the unique bi-modal distribution below, which matches with the baseline later in the model.



Apriori a teacher and school name can expected to influence donations but given the size of differing categories they have been ignored in the present analysis. Gender of the teacher is however taken into account. Furthermore, there are over 5900 school districts and city, and there reasonable 1-1 mapping between them. Given co-currency school districts have been chosen instead of city assuming interpretability for Donor’ s Choose selection criteria.

Analysis

Model Specification: To identify the potentially vulnerable projects a multiple machine learning classifier were used. The need for machine learning classifiers is justified by conventional limitations of the standard models which are relatively not as good in gleaning information from over 1 million records across as many features. Seven separate machine learning models were used, including Logistic Regression, K-Nearest Neighbour, Decision Trees, SVM, Random Forests, Boosting, and Bagging.

Model Implementation: These models are ensemble classifier models that are essentially a composition of simple classifiers, each having their respective parameters (levers) values to optimise on. The classifiers models were trained on separate periods of time (6 months window) and tested on subsequent period to evaluate their within-sample performance. In order to extract the best model the testing-training simulation was run across multiple permutations of model parameters.

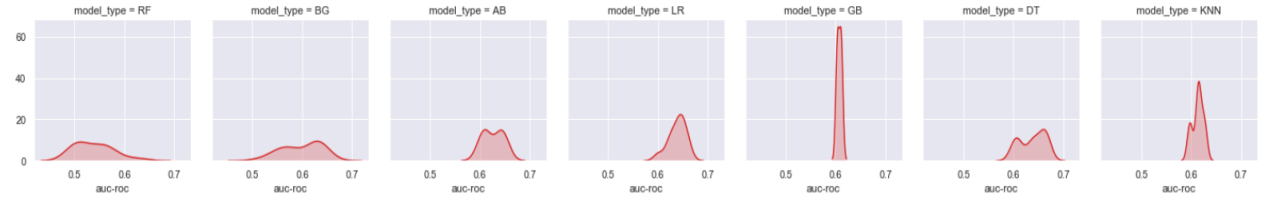
Evaluation

On Metrics: *Baseline ->Accuracy -> Precision*

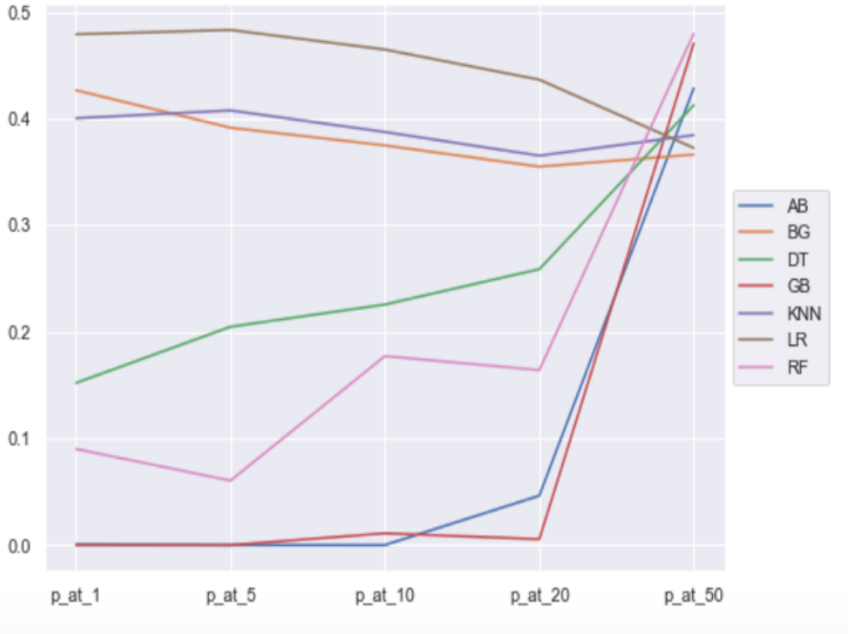
First step to access a classifier model is establishing a baseline performance metric. It represents proportion of projects that were unable to get full funding and was hovering around 30% on the given data. Progressive metrics capture accuracy and further true and false positive rates via ROC-AUC, F-1, Precision and Recall.

Classifier viz Metric: The present data has imbalanced classes in the test dataset(s). In other words, there are relatively fewer schools receive donations for different time periods as well as cumulatively. Since it more likely to estimate lesser number of schools with higher accuracy, standard model performance metric from ROC-AUC, will not suffice. Therefore, with the intent to study a.) Of all the schools that were predicted as recipients of donations, what fraction were actually underfunded ? (Precision). And, of all the schools that actually received the full funding, what fraction was correctly predicted as underfunded? (Recall) and F-1 statistic were accessed.

Choosing Optimal Model(s) & Metric: Average AUC-ROC score for all models was more than 60% which indicates that model(s) perform marginally better than a random guess. On an aggregate, the figure below compares model accuracy and it can be seen that Decision Tree (DT) had a wider spread of accuracy from 60% onwards.

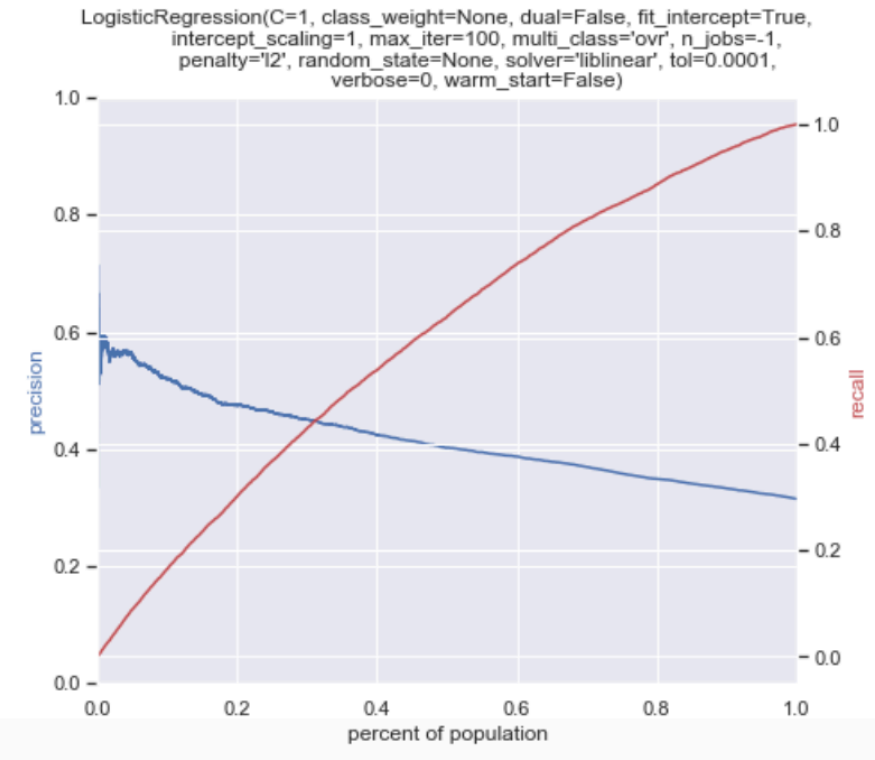


Further on selecting the best model based on optimal value of different metrics (auc-roc, f_at_50, r_at_50, p_at_50), a relatively common trend was observed, where the recall performed well and precision was low. Given the context of the current problem, it is imperative to identify the vulnerable projects and therefore recall (measure of how many truly relevant results are returned) is chosen over precision (measure of result relevancy).



The optimal model identified in this specification is a L2 regularized Logistic Regression Classifier, with C =1. The model is fairly consistent across different time windows. As can be seen the precision is not particularly promising (< 55%) but the model does a performs resonably better than picking randomly

It can be seen (optimal model chart below) we witness a higher recall value but precision is low and hence the reason for selecting recall as the metric. This Recall metric and (ergo the selected model), is also complaint with identifying projects that are at highest risk of not getting fully funded to intervene with.



Model Caveats & Extensions:

- The model performance is limited by ability to run fine-tuned simulations, therefore precision can be improved further with large grid size
- Exploiting more intuitive features and their interaction might also improve the performance
- Inclusion of other attributes (e.g. demographic from census) might also affect/improve performance
- With the growth of the program and subsequently more training data one can expect a potential improvement in prediction