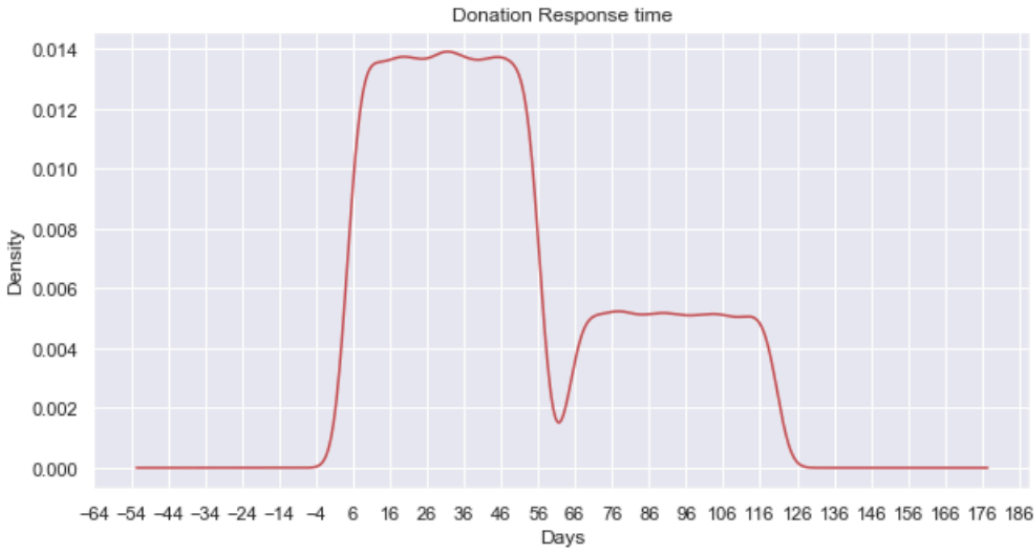
# Donor's Choose: Improving Machine Learning Pipeline

Summary: The objective of the present analysis was to prioritize which 5% of school projects should a donor prioritize their intervention on, to help with aid and assistance? Based on likely capacity constraint faced by a prospective donor (Donor's Choose) 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. Gradient boosted model with 100 estimators, learning rate: 0.5, max depth: 20, min sample split: 500 & subsample 0.8 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 (disproportionately higher frequency of values on either side of the mean) mean imputation is not recommended, but the missing % is less than 0.001% of data. Outliers in donation are plausible and have not been removed since they exist for variables like funding, where it is plausible to consider higher values. They have been normalized however to help in model prediction process.

**Features:**
It is interesting to observe that most projects in sample data are funded within 60 days period. Inferring from the unique two peaks(bi-modal) distribution below, which confirms with the baseline later in the model.



In order to predict better, few of the given variables in the data have been explored further to create additional attributes. Apriori a teacher and school name can be 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 considered. Another tested metric was teacher qualification, teacher name having prefix 'Dr.' but it was not considered as there were proportionately very low in the present data. Major cities have also been considered given proportionately large representation of schools in them. Furthermore, information on primary and secondary subject and focus area was present in the dataset. But given almost 1-1 mapping between subjects to respective focus areas, the present analysis considers subjects over focus areas.

## 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, KNearest Neighbors, SVM, Decision Trees, Bagging and ensemble (improved) methods: Random Forests, Boosting (Ada and Gradient).

**Model Implementation:** These models are ensemble classifier models that are essentially a composition of simple classifiers, each having their respective parameters (levers) values to optimize model on. The goal to build a durable or a generalized model, i.e. a model that does **not predict only** the current period/data well. In order to build generalize the model over different spans of time, the model is first built on a subset time window and then tested for subsequent time window. This time window (defined as 6 months) is progressively increased to capture the information from the dataset exhaustively and at the same time ensuring model durability. Since the project donation is assessed for a 60-day window, a buffer/gap was introduced between the period where the model was and in the prepared period was tested in. In order to extract the best model, the testing-training simulation was run across multiple permutations of model parameters.

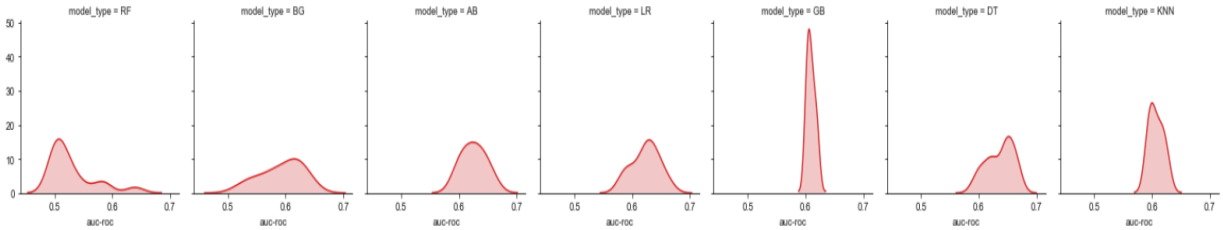
## Evaluation

Metrics: Baseline ->Accuracy -> Precision

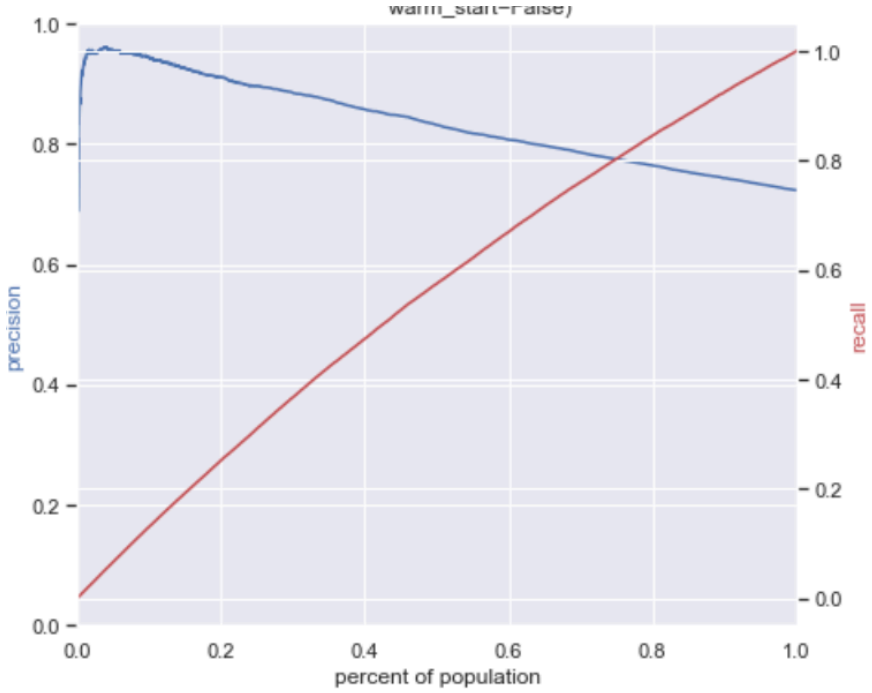
First step to access a classifier model is establishing a baseline performance metric. It represents proportion of projects that were unable to get full funding and was hovering around 30% on the given data. Progressive metrics capture accuracy and further true and false positive rates via ROC-AUC, F-1,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

Classifier & Metrics: 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 & 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 Gradient boosted model with 100 estimators, learning rate: 0.5, max depth: 20, min sample split: 500 & subsample 0.8. The model is fairly consistent across different time windows. As can be seen the obtained precision is high (~96%) and the model does a performs reasonably better than picking randomly. The Precision metric and (ergo the selected model), is also complaint with identifying projects that should be prioritized for intervention and maybe even considered at a relatively higher risk of not getting fully funded.




## Model Caveats & Extensions:

- The model is tested for 2012 and 2013, and should potentially do well if there is not much change in the variables used in the model
- Inclusion of other attributes (e.g. demographic from census) might also affect/improve performance