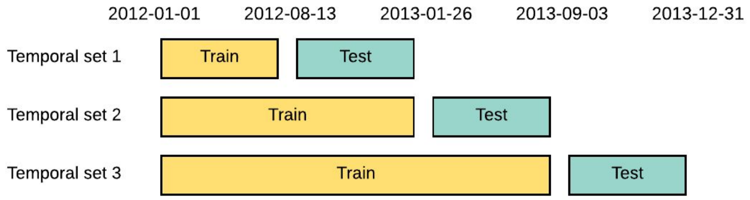
Predicting funding status of educational projects

Aya Liu 5/2/2019

In this pipeline, I ran various models to predict whether an educational project is not fully funded within 60 days. Predictions from this model can be used to target educational re

The predictors included in this model are: *school charter, school magnet, primary focus area, resource type, secondary focus area, poverty level, grade level, eligible double your impact match,* and *total price including optional support.* They are selected because they had different distributions for projects funded and not funded within 60 days. See *hw-pipeline-notebook.ipynb* for the code/output of exploratory analysis and models built.

I used temporal validation method to split data in to 7-month blocks for training and testing models. See the chart right for the time period each training and testing set spans.



Metrics comparison across classifier types

I ran decision trees, k-nearest neighbors, logistic regression, random forest, boosting and bagging models with different parameters for each test set. We will compare each classifier's precision, recall and ROC curves, both within each test set and across test sets with different time periods.

Precision

- The classifier with highest precision at less than 30% population is the boosting model for all three test sets. The model has similar precision level for 50, 100, 200 estimators as well as over time.
- At more than 30% population, 10-NN uniform has the highest precision. For the first test set (trained on 7 months of data), all models have minimal differences after 50% population. For the third (trained on 14 and 21 months), 10-NN have the best precision until 70% population, after which all models have minimal differences.
- Compared to the models from the first temporal set, models trained on data of longer time period have similar precision before 30% population but lower precision after that threshold.

Using 1) Boosting model with 50 estimators at <30% population and 2) 10-NN at >30% population using the most recent 7 months data allows us to get the highest %predictions right out of all the projects we predict to be not fully funded within 60 days.

Recall

- The classifiers with highest recall at any k population is the boosting model (50, 100, 200 estimators) and logistic regression (L1, L2). These are stable across test sets.
- Models built on temporal set 1 and 2 (14 and 21 months of data) reaches high recall faster. Boosting and logistic classifiers reach 100% recall at <50% population, as compared to the 60% population level in test set 1.

Using Boosting model with 50 estimators or logistic regression models with L1 or L2 allows us to capture the highest proportion of all projects not fully funded within 60 days with our prediction. The more data we train these models on, the the lowest %population constraint.

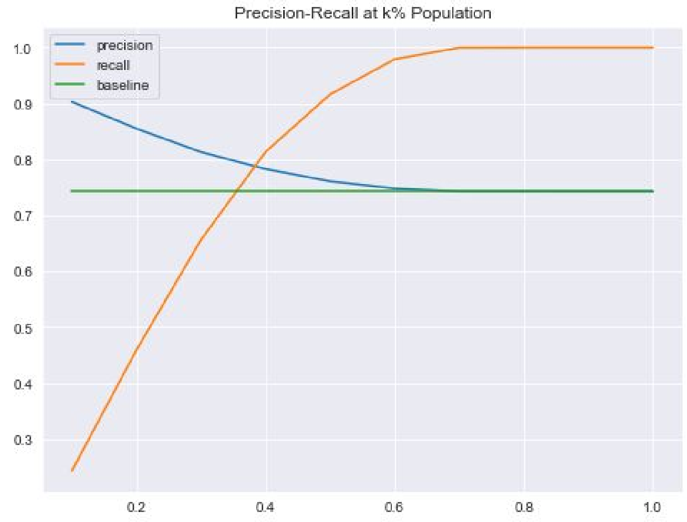
ROC

- Boosting (50, 100, 200 estimators) and logistic regression (L1, L2) has the highest AUC score, showing a better overall performance at any k% population threshold. The model immediately following these is the Decision Tree with max_depth=10.

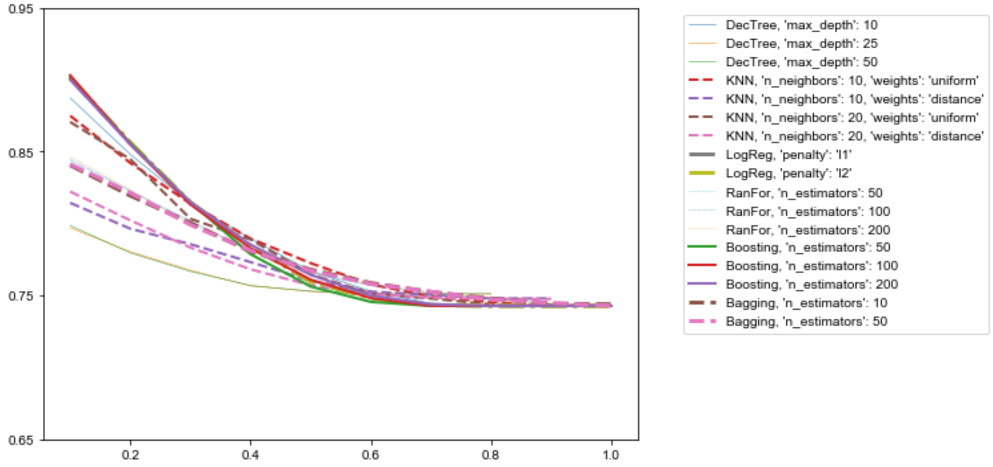
Goal and Recommendation:

Given the constraint that we can only intervene with 5% of posted projects, we would want the model with the highest precision at 5% population. In other words, we would want as many of the projects we predict to be unfunded to be actually unfunded, to maximize the effect of using our resources for intervention.

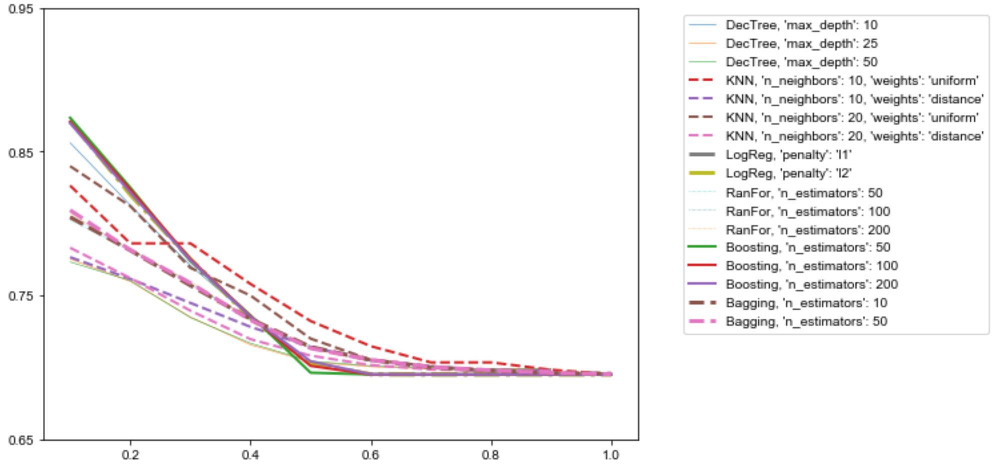
Based on the analysis above, the best model that satisfies that criteria is a Boosting model with 100 estimators. The precision/recall curve for the model is:



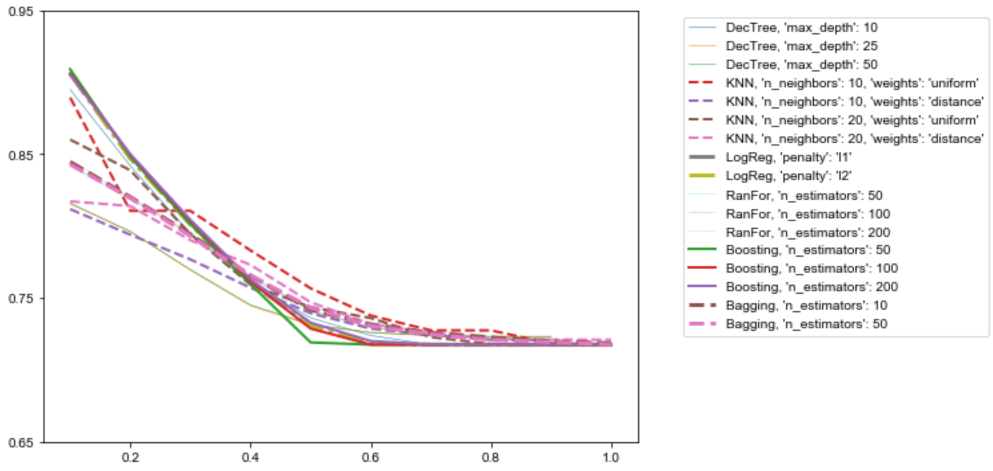
precision - Temporal set 1



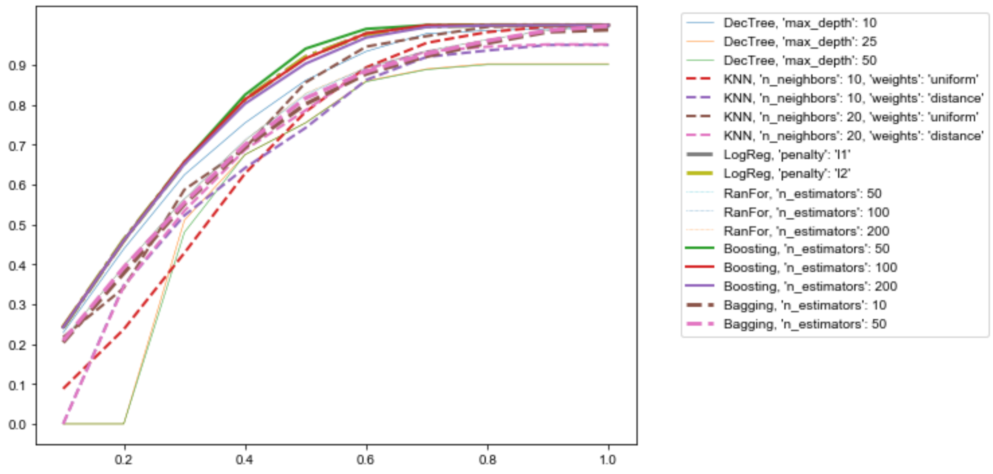
precision - Temporal set 2



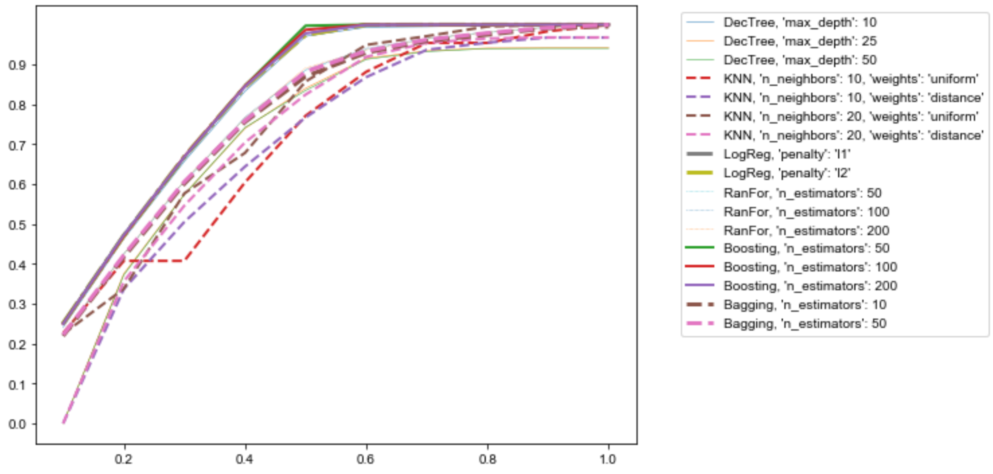
precision - Temporal set 3

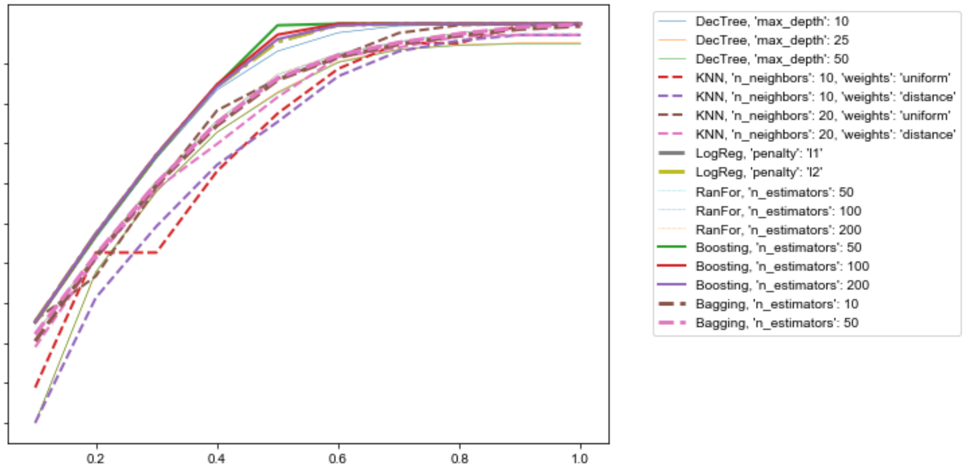


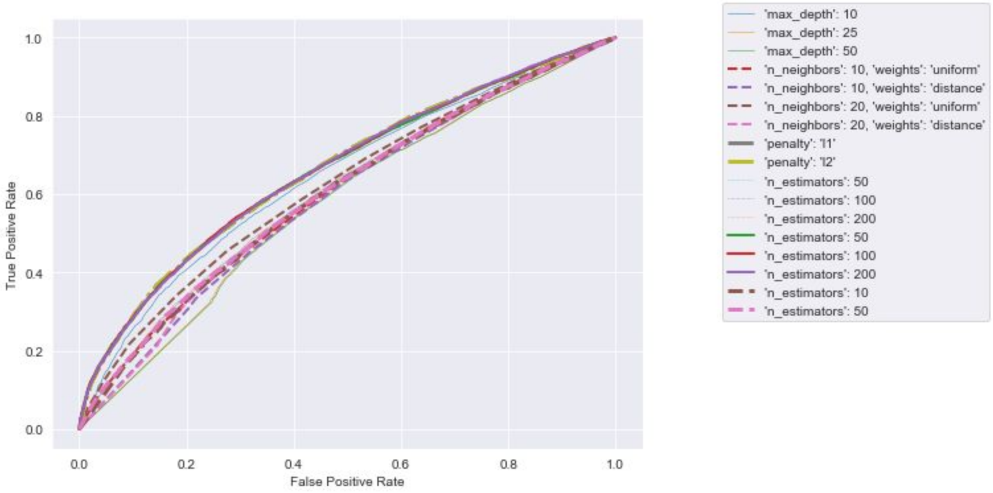
recall - Temporal set 1

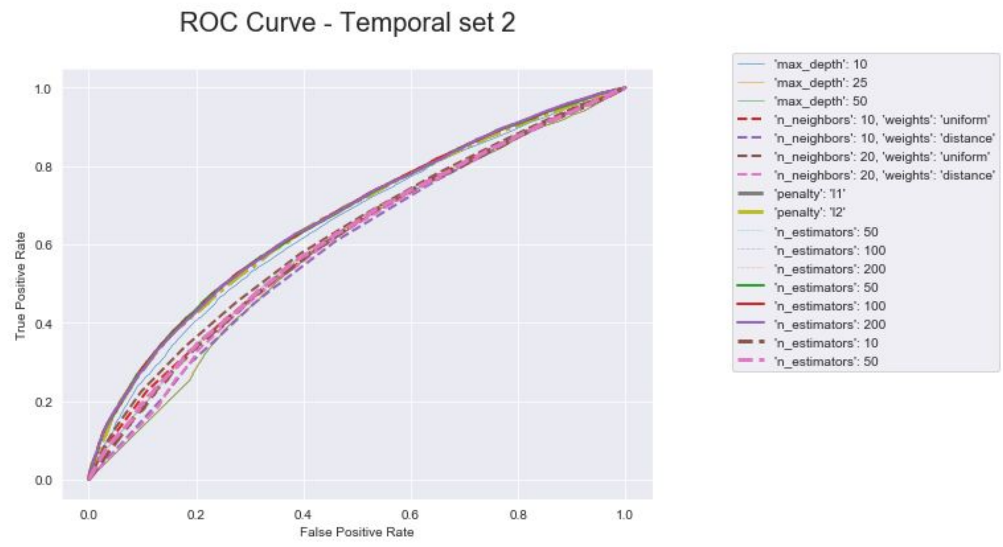


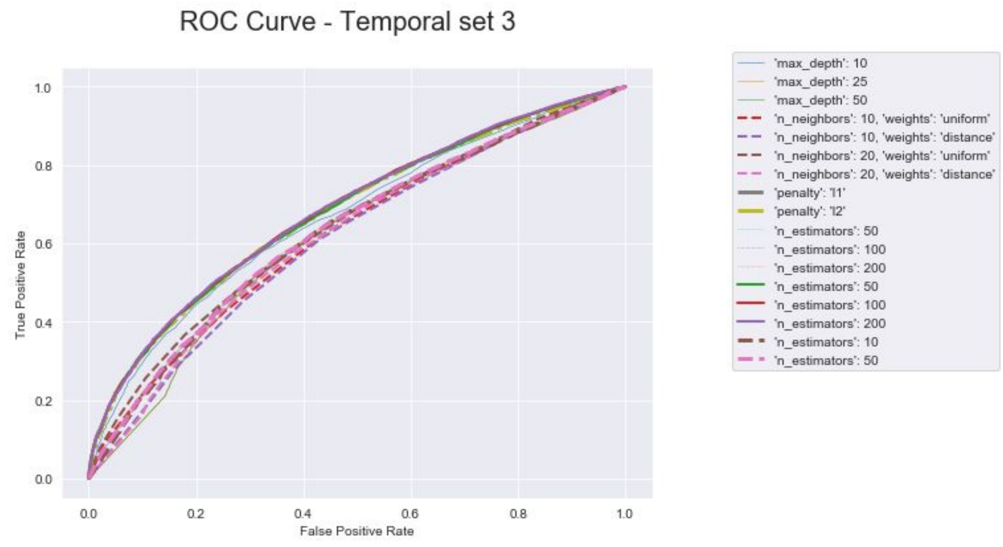
recall - Temporal set 2




recall - Temporal set 3


ROC Curve - Temporal set 1





What would be your recommendation to someone who's working on this model to identify 5% of posted projects to intervene with, which model should they decide to go forward with and deploy?

The report should not be a list of graphs and numbers. It needs to explain to a policy audience the implications of your analysis and your recommendations as a memo you would send to a policy audience.