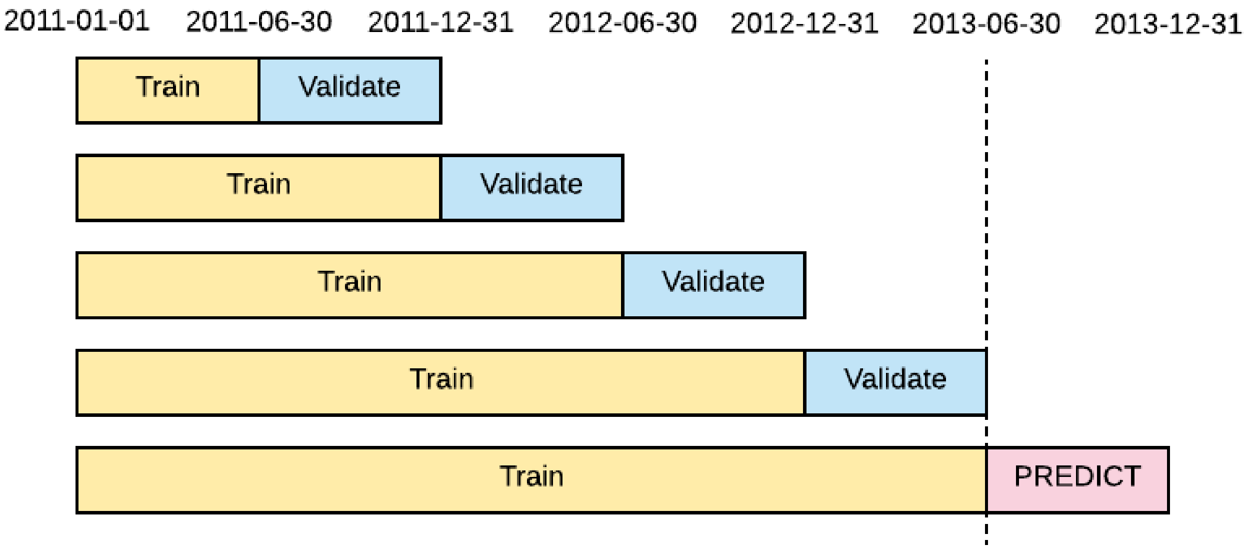
Evaluation of Models on Predicting Fundraising Project Success

Background and Objectives

DonorsChoose.org is an online charity and crowdfunding platform that allows individuals to donate directly to public school classroom projects. The project postings come from teachers in K-12 schools requesting school materials ranging from basic school supplies like paper and pencil to technology requests like computers. The goal of this challenge is to predict, based on what we know about the project at the time of its posting, whether a project will be fully funded. Data we have include geographical information about schools, grade level, teachers, projected cost of project, and potential impact in number of students reached.

Temporal Holdout Framework

Given data on 2011 to 2013, we reserve the final six month time period (2013-07-01 to 2013-12-31) to make a final prediction. We train models with various machine learning models using successively increasing time chunks, validating the models with the six months of data immediately following the training period. The details of the timeline used for model validation is shown below. This framework will be helpful in discerning whether models perform differently over time or there is an anomaly during a given period.



Results of Models

Our analysis seek to assess different models and parameters' efficiency and accuracy in predicting whether a project listed on DonorsChoose.org will receive full funding. Regarding efficiency, K-Nearest Neighbors and Support Vector Machines are extremely time-intensive to run, and, therefore have been deemed unsuitable and excluded from our comparison. A preliminary evaluation of KNN on the full training data took approximately 1,900 seconds.

We fit six different learning algorithms (Random Forest, Decision Trees, Logistic Regression, Naive Bayes, Bagged Trees, and Boosted Trees) to the training set 1) using all features and 2) using only the top 20 "important" features computed using a random forest of trees for each training set and find the following:

- In the case where we consider only important features, the model with the highest auc-roc is a random forest with a max depth of 5 at 0.66. In the case where we consider all features, a random forest with max depth of 50 has the highest auc-roc of 0.68. Given the baseline accuracy for the model is 0.71, the result is unimpressive.
- However, all models perform better on precision than recall, maintaining a precision above 80% for those projects with the top 10% highest probability of being classified "fully funded." While our models do not make many false positive predictions, they tend to misclassify those projects that might actually have a chance of being fully funded as having a low probability.

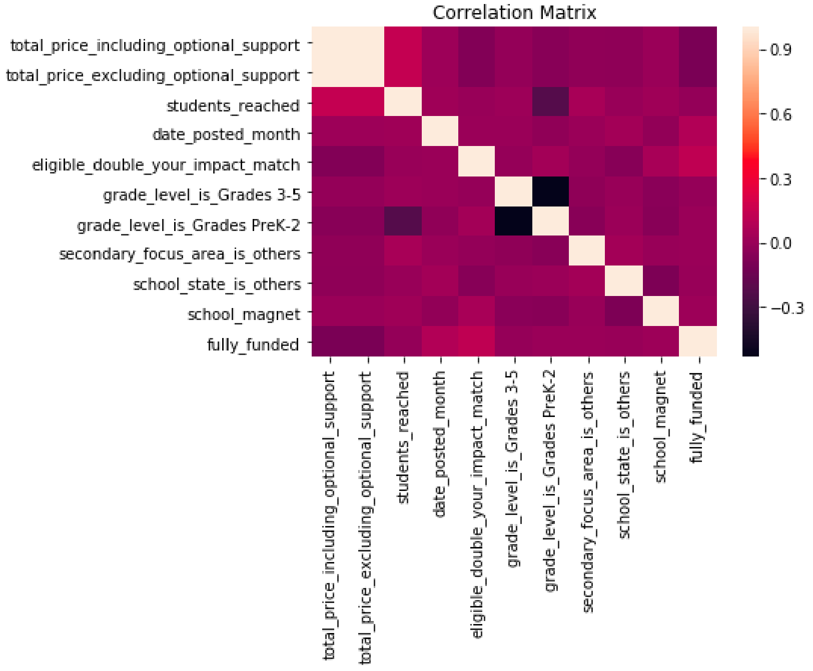
Recommendation Based on Feature Importance

A natural thing to wonder and perhaps is most salient for teachers who seek to receive full funding for their projects is which features have the most predictive power or are most correlated with the outcome. Based on our results, we conclude that while our models perform better when we include everything we know about the project at the time of prediction, there are few variables that tend to drive the prediction.

Important Features

	feature	importance
1	total_price_including_optional_support	0.143728
2	total_price_excluding_optional_support	0.143607
3	students_reached	0.119425
4	date_posted_month	0.105699
5	eligible_double_your_impact_match	0.020855
6	grade_level_is_Grades 3-5	0.017929
7	grade_level_is_Grades PreK-2	0.016612
8	secondary_focus_area_is_others	0.015934
9	school_state_is_others	0.015385
10	school_magnet	0.015365

Correlation



The above are the top 10 important features computed using a random forest of trees on the full training set and their correlation with the outcome variable and one another. The importance of each feature drops after the top four. The total price of the project emerges as the most important determinant in funding. Number of students reached is also an important feature. Therefore, an interesting aspect to explore may be the tradeoff between the cost of a project vs. its potential impact.