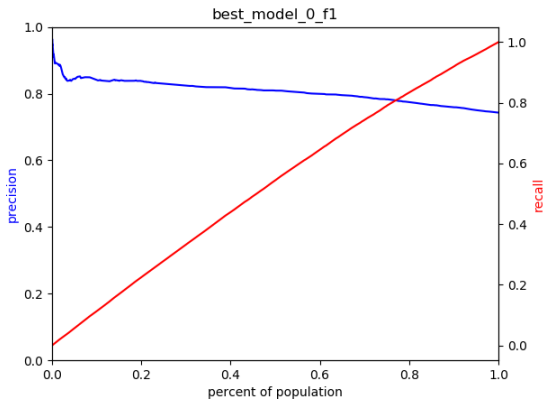
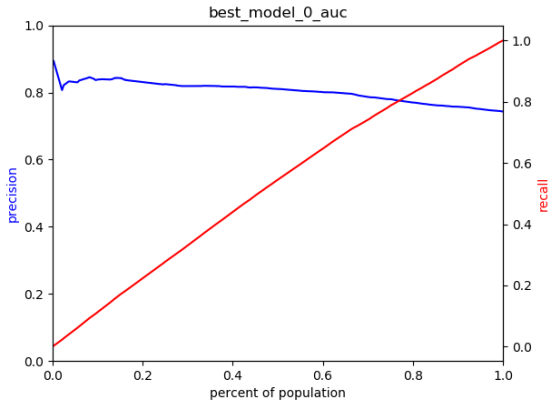
The goal of this project is to identify the projects that need to be improved in order to get funded within 60 days. Our models could be used as a reference for the sponsors to choose projects with high risk of not getting funded with 60 days. We use the data from DonorChoose in 2012 to 2013 and define the outcome as "1" if the projects got funded with 60 days of posting, otherwise it would be labeled as "0". Based on the available data, I choose 5 categorical variables as features of interests, including 'primary_focus_subject', 'primary_focus_area', 'poverty_level', 'students_reached_category', 'total_price_including_optional_support_category' (all are converted into dummies). In order to avoid predicting "the past" using "the future", I use temporal validation with a rolling window of 6 months here. For each pair of train and training sets, I train 24 models varying from logistic regressions to boosting. The results indicate that: logistic regression outperforms other models in f1 socre and random forest can also be a good option for the sake of understanding and great auc_roc score.

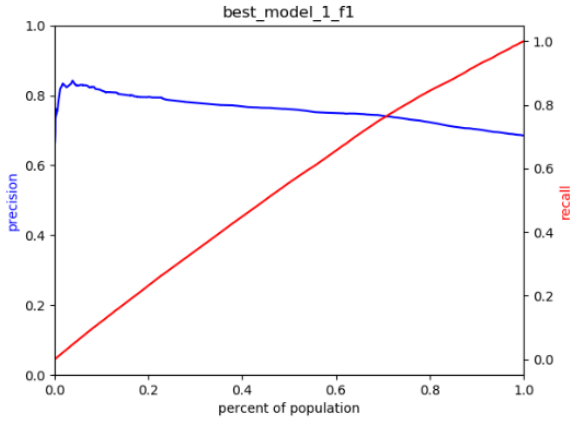
Report of Best Models

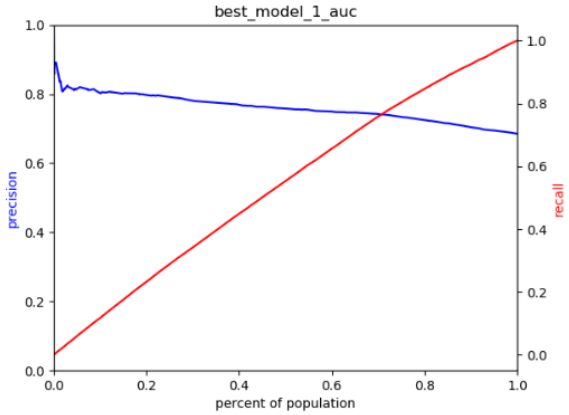
In the notebook, I report the models with highest f1 score and auc_roc for each validation set. (you can also check the models' performance on different metrics in the output data frame.) Considered that the sponsors have limited resources to intervened with only 5% of projects, I compute those metrics based on 95% thresholds (top 95% of records are predicted as positive) except for precision and recall.

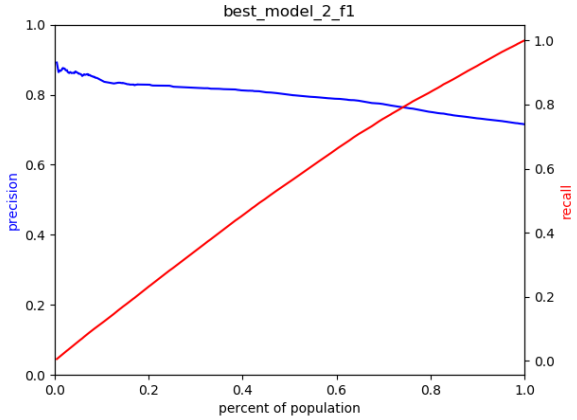
For the training set starting from January 1st, 2012 to June 30th, 2012 and testing set starting from July 1st, 2012 to December 31st, 2012, Logistic Regression with penalty of l1 and C of 0.1 has th e best f1 score (0.84235947) and Boosting with Logistic Regression as base estimator, number o f estimators of 10 and learning rate of 1 has the best auc_roc score (0.58824384).For the training set starting from January 1st, 2012 to December 31st, 2012 and testing set starting from January 1s t, 2013 to June 31st, 2013, Logistic Regression with penalty of l1 and C of 0.1 has the best f1 scor e (0.80624504) and Random Forest with criterion of 'entropy', number of estimators of 3 and m ax depth of 9 has the best auc_roc score (0.58931038). For the training set starting from January 1st, 2012 to June 30th, 2013 and testing set starting from July 1st, 2013 to December 31st, 2013, L ogistic Regression with penalty of l1 and C of 0.1 has the best f1 score (0.8268974) and Random Forest with criterion of 'gini', number of estimators of 10 and max depth of 9 has the best auc_ro c score (0.6052695).

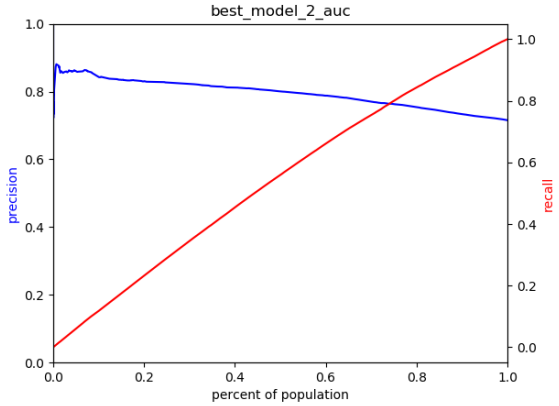












From the six pictures above, we can see that, all the best models perform well at both precision and recall at roughly 0.8 percent of population, which is close to the ideal size of sponsors' intention. In addition, since we pick a high threshold, the precision scores are always close to the baselines which are close to 0.7.

Recommendations

Since logistic regression performs best in f1 score for all the three periods, I would recommend logistic regression as the best model. Albeit it does not perform the best in all the metrics, it does not fall far behind, either. And it is also easy to implement.

Furthermore, considering about easy understanding, I would also recommend random forest as the best option. As an adjustment technique for decision tree, it can also help us to identify the importance of different features. I check the feature importance of the two random forests and both of them report categories of 'total_price_including_optional_support' as the most important feature. Second comes the category of 'poverty level'.

Therefore, I would suggest the sponsors to post tips about how to design reasonable budgets for those projects and then pay attention to the poverty level of the schools. The donors may prefer to helping with poor schools.