## A. Introduction: problem, goal and summary

Donor’s Choose is an online platform that helps teachers to obtain funding for school projects. The goal of this project is to improve the probability that a school project gets funded within 60 days. Based on this goal, I want to predict which school projects have high risk of not getting funded within 60 days of posting. The prediction could be beneficial in two ways: first, the Donor’s Choose are capable of intervening with high-risk projects based on our prediction; second, the teachers who post projects on this platform can learn something to improve the probability of getting funded earlier.

Therefore, I define projects which did not receive funding within 60 days after posting as outcome of 1, otherwise 0. In this dataset, there are over 80% projects getting funded within 60 days.

Overall, albeit none of those models are particularly strong, they provide a significant amount of lift compared with the baseline precision. The unsatisfactory performance might be attributed to the imbalanced distribution of two labels in dataset. Considered the performance in different evaluation metric, I would recommend Donor’s Choose uses Logistic Regression to identify the project that are less likely to get funding with 60 days.

## B. Feature Selection and Results

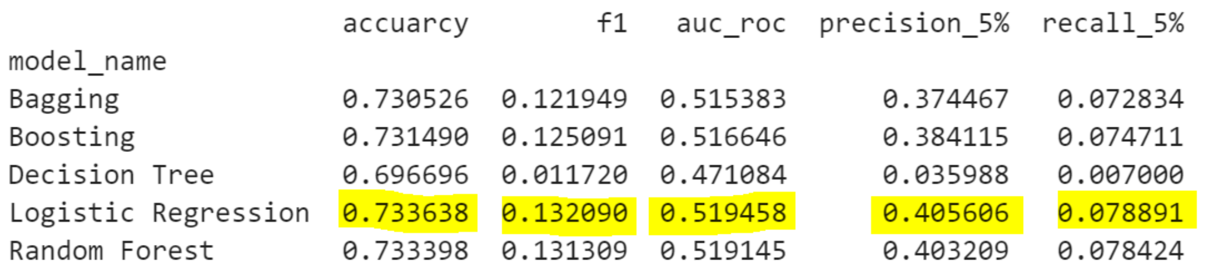
In order to provide as many information as possible but avoid overfitting, I only drop the columns of identification and time information. For categorical variables, I imputed the missing value as ‘unknown’. For continuous variables, I imputed the missing value of corresponding mean of training dataset. In addition, since the distribution plots of ‘students reached’ and ‘total price including optional support’ indicate that there exist outliers, I discretize these two variables based on the percentile statistics of training data set.

For all the Decision Tree and Random Forest models, I print out the top 5 important features. It turns out that the geographic information and price information could be the best features to predict outcome. The latitude, longitude and “eligible_double_your_impact_match” can provide relevant information, but this needs further study. And the projects with low cost or not is a good separable standard. One salient and relevant fact: in the help center of Donor’s Choose, their tips suggest the teachers to charge as low as possible since low cost project are more likely to get funded in 3 months and if the project is over 1000 dollars, it would be better to split it. So, our model could verify that their tips are useful.

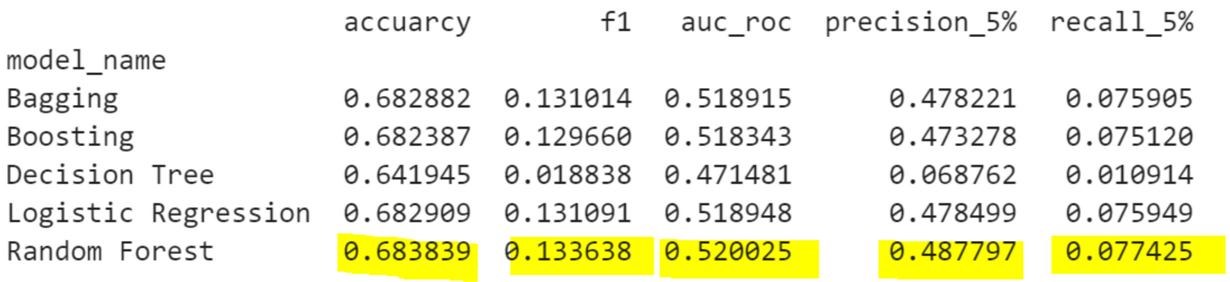
## C. Evaluation Performance

In this project, we use 5 evaluation metrics: **Accuracy** tells us for all the projects, how many of them that we predict their outcomes correctly; **Auc-Roc** tells us how well did we rank our space (the baseline is 0.5 for random guess); **Precision** tells us among all the project we identified as not likely to get funded within 60 days (we identified top k percent of projects as not likely to receive funds based on predicted scores ranking, the same as recall), how many of them actually did not get funded in 60 days; **Recall** tells us among all the projects which actually did not receive funds with 60 days, how many of them are correctly predicted by our models. **F1 score** reports the harmonic average of the precision and recall. Also, I provide the precision of baseline as a reference, which is the precision when I guess all the projects will not get funded within 60 days (all predicted as label 1). Since Donor's Choose hopes to intervene with 5% of the project, all metrics are calculated at the 5% threshold except for precision and recall.

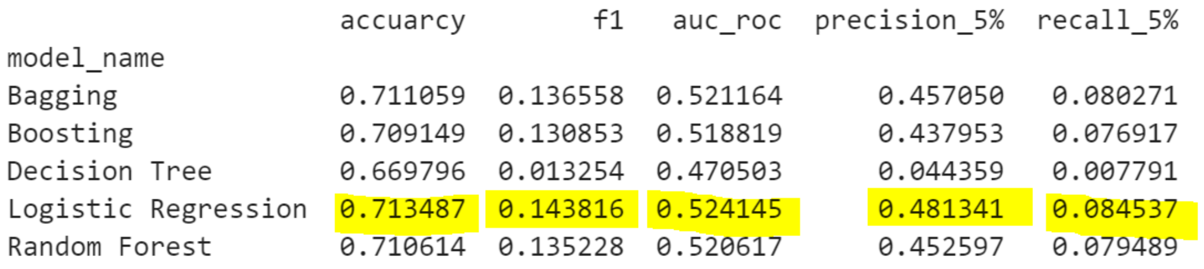
Basically, I train 5 kinds of models and adjust their hyper parameters: Decision Tree, Logistic Regression, Bagging, Boosting and Random Forest. And for each pair of training and testing data set, I report the average performance of these models (The highlighted scores are the highest average score).



```
                     accuracy        f1   auc_roc  precision_5%  recall_5%
model_name
Bagging              0.730526  0.121949  0.515383      0.374467   0.072834
Boosting             0.731490  0.125091  0.516646      0.384115   0.074711
Decision Tree        0.696696  0.011720  0.471084      0.035988   0.007000
Logistic Regression  0.733638  0.132090  0.519458      0.405606   0.078891
Random Forest        0.733398  0.131309  0.519145      0.403209   0.078424
```



```
                     accuracy        f1   auc_roc  precision_5%  recall_5%
model_name
Bagging              0.682882  0.131014  0.518915      0.478221   0.075905
Boosting             0.682387  0.129660  0.518343      0.473278   0.075120
Decision Tree        0.641945  0.018838  0.471481      0.068762   0.010914
Logistic Regression  0.682909  0.131091  0.518948      0.478499   0.075949
Random Forest        0.683839  0.133638  0.520025      0.487797   0.077425
```

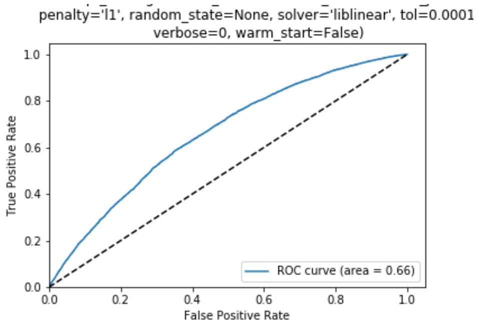


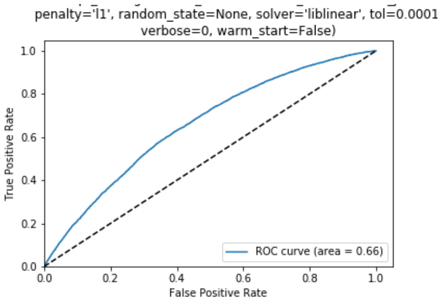
```
                     accuracy        f1   auc_roc  precision_5%  recall_5%
model_name
Bagging              0.711059  0.136558  0.521164      0.457050   0.080271
Boosting             0.709149  0.130853  0.518819      0.437953   0.076917
Decision Tree        0.669796  0.013254  0.470503      0.044359   0.007791
Logistic Regression  0.713487  0.143816  0.524145      0.481341   0.084537
Random Forest        0.710614  0.135228  0.520617      0.452597   0.079489
```

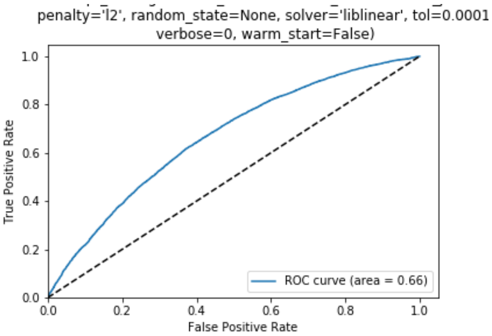
We can see that in the first and third validation, Logistic Regression earns the best performance in all metrics while Random Forest has the best performance in the second validation. Overall, the precision at 5% are around 40%, which is much higher than baseline (0.256928) and the auc_roc score is better than random guess (0.5). However, due to the bad performance in recall (roughly 0.08), those models all did badly in f1 score.

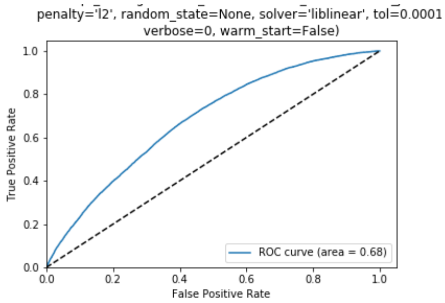
Considered that this is an imbalanced dataset, I also report the details of models who earn the best auc_roc score in each validation and their roc curves (all are Logistic Regression):

In the first validation, the best models are the ones with penalty of “l1” and C of 1 or 10. (The top two graphs) In the second validation, the best model is the one with penalty of “l2” and C of 0.1 (The left one on the downside). In the third validation, the best model is the one with penalty of “l2” and C of 10 (The right one on the downside).









Hence, I would recommend Donor’s Choose to use Logistic Regression to predict the outcome of projects for the following reasons: first, the performance of Logistic Regression in temporal validation in most metrics is growing, which is particularly conducive for this platform because they would have more data in the future and this model would be more accurate; Second, among all the models, Logistic Regression has the best or the second best average performance. Third, Logistic Regression is easy to implement. But given the imbalanced data structure, I would suggest using certain techniques to solve this problem. For example, the platform should collect more data of projects that are not get funded within 60 days or considering resampling.

## D. Policy Recommendation

Donor’s Choose could use Logistic Regression as an assistance to identify the high-risk projects and thus prioritize them and offer helps to improve their proposal. But they should not withdraw support for their old clients since the current models are limited and cannot be a perfect alternative for human rules. They had better to further refine these models especially when they collect more data about high-risk projects. The important features identified by Decision Tree and Random Forest models could also be a reference for Donor’s Choose to provide tips for teachers albeit they did not perform well in evaluation metrics.