The purpose for this machine learning project is to make predictions for individual education projects with different values of project-related features, and provide early intervention to the projects which are less likely to get fully-funded.

To achieve the goal of this project, multiple models are implemented including Logistic Regression, K-Nearest Neighbor, Naïve Bayes, Decision Trees, Ramdom Forest, Boosting, and Bagging. Different evaluating metrics give different best models, and different models rank differently under each evaluation metrics. Details are discussed below, and finally, a suggestion of model to choose with an overall evaluation is raised.

1. AUC

- When splitting the training set and test set with the split date sets at 01/01/2012, we see Logistic Regression, scoring 0.9059, ranking the best with a performance way better than the rest of the models (the second being Gaussian Naïve Bayes, scoring 0.7299). K-Nearest Neighbor and Bagging is somewhat off when we look at their AUC.
- If modeling with the splitting date set at 06/30/2012, Logistic Regression still gives us the best model with the highest AUC of 0.9003, which is slightly lower than splitting training and testing sets at 01/01/2012. Gaussian Naïve Bayes falls to 5th place with a sharp drop in the AUC—from 0.7299 to 0.5851. If looking at AUC in the second set of models splitting on 06/30/2012 compared to those on 01/01/2012, the overall models performances drops, among which the ones of Naïve Bayes and Decision Tree have seen the most significant drop, implying those two models potentially being temporal sensitive.

2. Accuracy

- Logistic Regression performs the best under the evaluation of accuracy when splitting on 01/01/2012, with a score of 0.9167, and the second and third places are taken by Boosting and Decision Tree, with a significant drop in the accuracy—both have an accuracy of 0.8397.
- Logistic Regression and Boosting still rank the top two when we split the training and testing sets on 06/30/2012, when Random Forest beats Decision Tree and come to the third place, with an accuracy of 0.8370 in a tie with Boosting.

3. F1_score

- Once again, Logistic Regression rank in the first place when under the evaluation of f1_socre (0.9408), with Boosting, Decision Tree and Random Forest in the second, third and fourth places respectively (all scoring 0.8987 in a tie). Yet we can see from the results that, the models don't show strong variances in f1_score as with accuracy and AUC.
- When shifting to split date set on 06/30/2012, with Logistic Regression still being the best-performance model to choose, Random Forest moves up to and takes the third place which is Decision Tree for 01/01/2012. Notably, Naïve Bayes get a significant low score for f1_score.

4. Precision

- Evaluating models with precision splitting data on 01/01/2012, once again, give us Logistic Regression as the best model, with Naïve Bayes and Boosting being the second and third best choice. Generally speaking most models perform well at predicting true positive against false positive.
- However when we modeling with training and testing sets split on 06/30/2012 gives us differently the best model to choose—Naïve Bayes outperforms Logistic Regression and ramp up to the first place.

5. Recall

- When looking at how many projects ending up being not sufficiently funded are identified by our models, Bagging ranks the best among all candidate models. Logistic Regression has a modest rank for recall. However, almost all models have a satisfactory enough level of recall only except for K nearest neighbor.
- Ranking by recall doesn't change much of the pictures when we split the training and testing date set on 06/30/2012.

6. Training time

- Training time is one of the crucial criteria in that it is a decisive score for the feasibility to build up the model. That's the reason why SVM is out of scope in this project in the first place. Among all the models that are selected as candidate models, Logistic Regression, while giving us the best performance almost always, also has the longest training time. Yet the training time of Logistic Regression is still reasonably acceptable. Among all the models, Decision Tree takes the shortest time to train.

7. Recommendations for Models

Based on the analysis above, Logistic Regression undoubtedly outperform the rest of the candidates as the best model to choose. Even though not always ranking the best when evaluating with recall and precision in the second temporal validation, it doesn't fall back and perform poorly under these evaluating metrics. Also with a well acceptable training time, it makes it possible to realize the prediction by implementing Logistic Regression.

Boosting is also a second best model that is worth our consideration. Yet it almost never outperform Logistic Regression whatever the evaluating metric is.

Overall speaking, Naïve Bayes is not a good choice only expect that when we emphasize precision and training and testing time more than any other metrics. However, candidate models don't show strong differentiation in performance under precision and training time. Besides, with the accuracy, recall, AUC and f1_score all at an unsatisfactory level, it won't be a wise choice to pick Naïve Bayes as the model to employ.