Analysis Report

This analysis report is divided into three sections. First, it compares the overall performances of seven models (Logistic Regression, K-Nearest Neighbor, Decision Trees, SVM, Random Forests, Boosting, and Bagging) by using the datasets separated by three timestamps (2012-07-01, 2013-01-01, 2013-07-01). Then it analyzes the performances of different models in different metrics with a threshold of 50%. Lastly, it compares the models so as to find the best one to recommend to someone who's working on this model to identify 5% of posted projects to intervene with.

The overall performances of different models across different datasets is constant. To evaluate which model does better on which metrics, we select the dataset 2, that is, we trained models with the data from 2012-01-01 to 2013-01-01, and use the data from 2013-01-01 to 2013-07-01 to validate and evaluate the models we trained. Meanwhile, we set the threshold of 50%. At the threshold of 50%, Support Vector Machine (SVM) performs best in the metrics of precision score (0.853) and ROC-AUC score (0.616), and the models of the highest recall score (1.000) are AdaBoosting (AB) and Random Forest (RF). Although SVM performs best in both precision score and ROC-AUC score, it has the lowest recall score (only 0.395). AB and RF have the same score in the three criteria (highest score in recall score). Despite the fact that the precision scores (0.705) and ROC-AUC scores (0.5) of AB and RF are smaller than SVM, but their performances are not as unstable as SVM. In addition, AB costs the least training time (0.037). Thus, based on the comprehensive analysis, AdaBoosting performs best overall at the threshold of 50%.

To find the best model used to identify 5% of posted projects to intervene, we set all criteria at the threshold of 5%. Through the rank of different dataset separated by timestamps, we can find the rank is constant across three datasets. The model of the highest precision score and the highest ROC-AUC score is Support Vector Machine (SVM). Although SVM dominates both precision score and ROC-AUC score, it has the lowest recall score. On the other hand, other models perform the same in all three criteria at the threshold of 5%. In addition, if we evaluate the models by training time, we can find that Decision Tree (DT) performs the best. Thus, DT is the best model at the threshold of 5%.