Analysis Report

This analysis report is divided into three sections. First, it compares the overall performances of seven models (Logistic Regression, Decision Trees, ExtraTrees, Random Forests, AdaBoost, GradientBoosting and Bagging) by using the datasets separated by timestamps. Then it analyzes the performances of different models in different metrics with the 20% of posted projects that are at highest risk of not getting fully funded. 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 dataset is constant. To evaluate which model does better on which metrics, I selected the standard of identifying 20% of posted projects that are at highest risk of not getting fully funded, and compared three datasets historically.

For the first dataset, I trained models with the data from 2012-01-01 to 2012-06-30, and used the data from 2012-09-01 to 2013-03-01 to validate and evaluate the models I trained. For the second dataset, I trained models with the data from 2012-01-01 to 2013-01-01, and use the data from 2013-03-01 to 2013-09-01 to validate and evaluate the models I trained. For the third dataset, I trained models with the data from 2012-01-01 to 2013-06-30, and use the data from 2012-09-01 to 2013-12-31 to validate and evaluate the models I trained. There are two things that should be noted: there is a 60 days gap between training data and testing

data; there is only 4-month data for the third testing data. Thus, I selected the performances of the models in the second dataset to analyze.

With the 20% of posted projects that are at highest risk of not getting fully funded in the second dataset, we can find that Logistic Regression (LR) performs best in the metrics of precision score (0.4853), recall score (0.3099), accuracy score (0.6810) and ROC-AUC score (0.58). Meanwhile, compared with the model with the shortest training time (AdaBoosting, 0.1101), the training time (0.4638) of LR is not disadvantageous. Thus, based on the comprehensive analysis, Logistic Regression performs best overall at the 20% of posted projects that are at highest risk of not getting fully funded.

To find the best model used to identify 5% of posted projects to intervene, I set all criteria at 5%. Also, I selected the performances of the models in the second dataset to analyze. Based on the outcomes, we can find that Logistic Regression (LR) performs best in the metrics of precision score (0.5373), recall score (0.8579), accuracy score (0.6906) and ROC-AUC score (0.5261). In addition, the training time (0.4638) of Logistic Regression model is appropriate. Therefore, Logistic Regression is the best model at the 5% of posted projects that are at highest risk of not getting fully funded.