Chi Nguyen
HW3 Improving the pipeline (Revised)

Using Machine Learning to Predict Funding Prospect of Educational Projects

Securing funds for small school projects could be hard given the limited amount of budget allocated to public education. Donors Choose, a non-profit crowd-funding platform education platform, was founded to solve this challenge. The platform allows teachers across the country to raise funds for classroom projects and allow kids in low-income neighborhoods to get access to better education environment. Due to the crowd-sourcing nature of the platform, not every project posted will get fully funded, or funded in the desired time window. Our analysis aims to use machine learning techniques to identify these projects based on a variety of explanatory variables collected by Donors Choose, such as school type, subject area, poverty level of the school's neighborhood, number of students reached, and the total cost of the project. An accurate identification will allow Donors Choose to intervene on projects with the highest risk of failure by building product features to boost these projects in view list, or creating a special sections to feature them to donors.

The dataset I am using to build predictive model contains projects that were posted between January 1st 2012 to December 31st 2013. Because the length of a semester is 4 months, I will define a project as in danger of not receiving funding as not being funded more than 60 days after posting date. I created an outcome variable that takes the value of 1 if the project was not funded within 60 days of posting date, and 0 otherwise.

Since this dataset contains time series data points, I will use a method called temporal holding to split the datasets into three different sets of training and testing data using three validation dates that are each 6 months apart, June 30th 2012, December 30th 2012 and June 30th 2013. In order to allow for outcome to happen, I will also leave a gap of 60 days in between my training and testing sets. The variety of training and testing sets will allow us to compare the models' performance across different time periods while aiming at increasing robustness.

In order to find the best model, I used seven classification methods with varying strengths: knn-neighbors, decision tree, random forest, logistic regression, support vector machine, gradient boosting and bagging. Among these models, decision tree and random forest will be very sensitive to changes in data inputs, making the comparison between model

performance between different training and testing sets interesting. Meanwhile, logistic regression and support vector machine methods are both known to be robust to outliers.

For all models, I will compare their performance to the baseline which is 30%. In evaluating models, I will rely on two metrics: precision, recall. In order to calculate these metrics, I use a threshold, which determines the number of projects that will be included in intervention. Lower threshold translates to higher intervention coverage, which will use more resources. Since we are using many variations of each classification to build models, to evaluate on the performance of each classification type, I will use the mean value of the relevant metrics.

Precision is the portion of projects that truly did not get funded within 60 days of posting date out of all projects predicted to not get funded by our models. This metric is of interest if we are constrained financially, and want to make sure that the resources are going to the right projects. with higher threshold meaning higher coverage. If we want to intervene at the top 5% most at risk projects, the support vector machine classifier, trained on the datasets split on December 30th 2012, has the highest average precision of 42.5%, which is higher than the baseline. Using this model, about 1 in every 2 projects that we intervene on is the right use of resources.

Recall is the portion of projects of projects not funded within 60 days of posting date that the model predicts correctly. The higher recall is, the more projects in danger of failure will receive the intervention. We often care more about this metric when we want to help as many projects as possible. For the top 5% most at risk projects, support vector machine also has the highest recall rates out of all classifiers. The highest recall rate was 7.2% for the support vector machine classifier that was trained on the dataset split on June 30th 2012. Logistic Regression classifier has the second highest recall rate at the 5% threshold of 6.1%.

The two metrics I am using to identify the best models are f1 score and area under curve. A higher F1 score represents a more optimal blend between precision and recall metrics, meaning that the model will balance coverage and correct utilization of resources. When intervening at the top 5% most at risk projects, the model with the highest f1 score uses the linear support vector machine classifier taking into account a penalty for overfitting. The second metric is area under curve. A higher value for area under curve means better classification performance. The best model in terms of area under curve uses random forest classifier.

Chi Nguyen
HW3 Improving the pipeline (Revised)

Overall, it is hard to say which classifier consistently outperforms the others on a certain metrics because the results vary based on the percent of projects we want to intervene with, and the time frames for training and testing sets. For comparison purpose, we will need to know at least what percentage of the projects our budget can cover to determine the right performance metric to optimize for. Further, I also graphed precision and recall metrics for each classifier against the split date to determine how the models perform over time. For all classifiers, the precision score is consistently the highest when the model was trained on the dataset that is split on December 30th 2012. The recall score is mostly flat over time for all classifiers. This suggests that feeding more information into training the model past this date does not necessarily help improve model performance. Finally, to intervene with the top 5% most at risk projects, I will choose the linear support vector machine classifier trained on the data split on December 30th 2012 since it has the highest precision and f1 score. Since our goal is to maximize the use of our budget, this model will be the most appropriate.