Homework 3

May 21, 2018

Alexandra Gates

0.0.1 1 Overview of the Project

1.1 Project Summary Education programs are seriously underfunded, undermining children's possibilities from the get go. This is especially true for children living in and just above poverty. At the moment, it is unclear which projects the school board will fully fund, hindering our ability to brainstorm and develop programs that will be effective. In order to ensure we the Parent Teacher Association (PTA) can properly allocate our time in developing programs, because we are exttremely busy, we want to develop a program that will predict which programs will be fully funded.

0.0.2 2 Project Work

2.1 Data The data are from two files. Projects: contains information about each of the projects and is provided for both the training and testing sets; and Outcomes: contains information about the outcomes of the projects in the training set.

2.2 Coding, Exploration, and Cleaning The two datasets are joined and the predictor (fully_funded) is immediately removed so as not to bias the data. All true, false entries are converted to 0, 1 boolean, and NAs are filled with the median of the column. Additionally, categorical variables that have less than 100 unique entries (for space and time purposes) are converted to binary data. This means that if the column school_metro has the options 'urban', 'suburban', and 'rural', those three options become their own options with a 1 or 0 indicating whether that project is in a school that is 'urban', 'suburban', and 'rural'.

2.3 Feature Generation To create a list of potential features to include in the model, I ran 2 models on the entire dataset: random forest classifer, LassoCV, and RidgeCV. Random forest retains all features with non-zero importance. LassoCV and RidgeCV then retain all features with non-zero coefficients. LassoCV and RidgeCV returned nothing, so the top 30 variables random forest provided were used in the modeling.

2.4 Modeling The dataset is modeled first by creating time-dependent testing and training datasets. Then, it is run with random forest, boosting, bagging, SVM, decision trees, logistic regression, and K-nearest neighbors and evaluated using accuracy, precision, recall, f1 score, and ROC-AUC.

0.0.3 3 Results

3.1 Model Data Our selection of data spanned January 1, 2011 to December 31, 2013. The models are trained on the first 6 months and tested on the remaining 1.5 years of data. I used a 6 month time period because that what was prescribed.

3.2 Model Outcomes

3.2.1 Which classifier does better on which metrics? To determine this, I would look at the accuracy, f1, recall, precision, and roc across all the classifiers and see hwo each of them did. Accuracy is the fraction of predictors, both positive and negative, that the classifier got correct. F1 is hthe weighted average of precision and recall and takes both false positives and false negatives into account. Precision is the number of true positives divided by all positives (true and false) returned by the predictor. Recall is the number of true positives divided by all that are supposed to be (true positives and false negatives). ROC is the rate of true positives compared tothe rate of false positives. For a threshold of 1%, decision tree, boosting, and bagging all had 100% in the evaluation metrics described above. This means that whether or not the program is fully funded is predicted accurately 100% of the time using those 3 models.

3.2.2 How do the results change over time? If my model were done correctly, I would assume that the evaluation metrics would change every time period, with the models getting increasingly higher scores on all evaluations. That said, I would assume that while some models are achieving higher scores, other models are getting lower scores. This means that the more data fed into the models, the better some get at predicting and the worse others get at predicting. This would help us determine the exact model to use. However, if any models are too good at predicting, they should also be examined and possibly thrown out as the model of choice for the data. We do not want the model to be too accurate because then it may be overfitting to the data. This means that the model is extremely good at predicting what will happen in this particular dataset, but it would not be generalizable to new data.

3.2.3 What would be your recommendation to someone who's working on this model on what model to go forward with? I would recommend choosing a model that passes the Goldilocks Test: it's not too bad, but it's not too good. Also, we would want to keep in mind what is most important to us: false positives or false negatives. If we care more about false positives (things that are classified as true but should have been classified as false), then we should make sure the model has high precision score, because that is looking at the number of true positives over the number of all positives classified by the model (true and false). If we care more about false negatives (when the model classifies it as false but it really should have been classified as true), then we want to choose a model that has high recall. If we care about both false positives and false negatives, we need a model with a high f1 score. If we care only about a particular portion of the population (say the top 10%), then we should look at the AUCROC and choose a model that that has a high score for the top 10% of the population.

0.0.4 Disclaimer:

While I know you can't give me points for something I didn't do, I just wanted to say I do realize that I didn't remove all columns except fully funded and the project ID from outcomes so the data

weren't biased. Additionally, I know the temporal validation and magic loop are not completely correct and that's why my outcomes are all weird. Stress seriously messes with my intestines and the last couple weeks were a doozy for stress. I decided to take a hit on this homework so my body could relax before the last push through to the end of the quarter. For my own experience and knowledge I will likely re-do this assignment after the quarter is done to make sure I understood it. Would I be able to stop by to get feedback on it then (not for credit)?