Classifying post-secondary school funding type based on student debt and earnings

Shay Lehmann and Michelle Ho

CUSP-GX-5006
Prof. Gustavo Nonato
Machine Learning Assignment # 3

Abstract

In this assignment, the authors explore how 4-year colleges in the United States can be classified into funding types (public, private for-profit, or private non-profit) based on data on the debt and earning outcomes of their students. The exercise aims to compare Support Vector Machines (SVM), decision trees (CART), and Random Forests (with and without boosting) as classification techniques and their in performance in a real life scenario.

1 Introduction

The main goals of this assignment are to compare three classification techniques and to better understand how earning and debt outcomes of students vary by the types post-secondary institutions they attend. The motivation is to ultimately understand what indicates a "high value" education for the final project of this course.

The dataset used in this assignment is provided by the U.S. Department of Education project, College Scorecard. The goal of the project is to provide data necessary for students and their families to compare and assess post-secondary institutions on their costs and student outcomes. The data is compiled from self-reported data from institutions, data on federal financial aid, and tax information for the past 20 years. This rich dataset includes measurements on multiple aspects of a postsecondary institution–including basic identifying information, admissions, student body demographics, degree programs, tuition costs, federal aid, debt repayment, completion rates, and earnings.

Our hope is that students and their families who are making decisions on post-secondary education can make use of our results to compare schools. Institutions themselves can use the results to assess how their students perform compared to peer institutions. Finally, loan granting institutions and the U.S. Department of Education can assess which schools are failing in preparing their students for career success and why.

2 Methods and Data Sets

In order to have data on earnings and repayment outcomes after graduation, this assignment uses only the 2010 - 2011 College Scorecard, since wage and debt data is not yet available for the years after 2011. This dataset contains 7414 institutions and 1743 variables. However, there is not complete coverage for all variables depending on the institution. For this assignment, subsets of the raw dataset was used for analysis, and the variables and data cleaning steps are described below:

- CONTROL: a categorical variable indicating the funding type of a school. Categories are 1="public", 2="private non-profit", or 3="private for-profit".
- REGION: a categorical variable indicating the region of the United States where the school is located. Regions are:
  - New England (CT, ME, MA, NH, RI, VT)
  - Mid East (DE, DC, MD, NJ, NY, PA)
  - Great Lakes (IL, IN, MI, OH, WI)
  - Plains (IA, KS, MN, MO, NE, ND, SD)
  - Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
  - Southwest (AZ, NM, OK, TX)
  - Rocky Mountains (CO, ID, MT, UT, WY)
  - Far West (AK, CA, HI, NV, OR, WA)
  - Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)
- GRAD_DEBT_MDN: Median debt for students who graduate
- WDRAW_DEBT_MDN: Median debt of students who withdrew without completion
- GRAD_DEBT_MDN10YR: Median debt by monthly payment (10 year plan) for graduates
- MN_EARN_WNE_P7: Mean earnings of students working and not enrolled 7 years after entry
- COMPL_RPY_3YR_RT_SUPP: 3-year repayment rate for students who completed degrees, suppressed for institutions of fewer than 30 students
- NONCOM_RPY_3YR_RT_SUPP: 3-year repayment rate for students who did not complete degrees, suppressed for institutions of fewer than 30 students
- LO_INC_DEBT_MDN: The median debt for students with family income between $0-$30,000
- HI_INC_DEBT_MDN: The median debt for students with family income $75,001+
- MD_INC_DEBT_MDN: The median debt for students with family income between $30,001-$75,000
- LOAN_EVER: Share of students who received a federal loan while in school
- The dependent variable being classified and predicted is 'Control' in this assignment's analyses. The authors may occasionally refer to this variable as 'Y'.
- The independent variables are all the others.
- Only institutions classified as ICLEVEL = 1 (eg. 4-year colleges) are used.
- Any observations with null values for any of the chosen variables are dropped. After these drops, the number of observations in the dataset is 2249 and 2433 for the two subsets used. The breakdown school types represented in the second subset is in Table 1, and was similar for the first subset. There were more private non-profit schools than private for-profit and public schools in the dataset.
- Finally, the categorical variable REGION is binarized so that all samples are represented by boolean feature vectors.

The steps taken for this assignment:

1. A classification model is fitted for selected independent variables with a SVM Linear model (Model 1) at varying training sizes
2. Next, a second SVM linear classification model is fitted using a second subset of independent variables (Model 2) with training size of 70% of the dataset
3. Next, decision tree classifier models are fitted on the second subset of independent variables with varying depths of tree allowed. The training size of all trees is set at 70%.

4. Next, decision tree classifier models are done without boosting (Model 3), with bagging (Model 4), with ADA boosting (Model 5), and with gradient boosting (Model 6) at varying depths.

5. All models are assessed via cross validation on training and test sets split from the original dataset. The training and test set sizes are adjusted to assess the effect of training size on quality of classification.

6. Confusion matrices and a plot of in-sample and out-sample errors are created for all models of the dependent variable "Control" to assess the performance of the classifications

3 Results

The first classification model was SVM with a linear kernel, fitted on the variables:
`REGION`,
`GRAD_DEBT_MDN`,
`WDRAW_DEBT_MDN`,
`GRAD_DEBT_MDN10YR`,
`MN_EARN_WNE_P7`,
`COMPL_RPY_3YR_RT_SUPP`, and
`NONCOM_RPY_3YR_RT_SUPP`.

The dataset was split into training sets of 20%, 40%, 60%, and 80% of the whole. The performance was not very good, and misclassified approximately 50 percent of the data for all training sizes (See Table 2). The confusion matrix with training size 80% can be found in Figure 1.

To focus on loan and debt variables, a second smaller subset of variables is chosen only for debt and loan variables:
`LO_INC_DEBT_MDN`,
`HI_INC_DEBT_MDN`,
`MD_INC_DEBT_MDN`, and
`LOAN_EVER`.

Model 2, like Model 1, is a linear SVM classification model. The confusion matrix for the model fitted with 70% of training data can be found in Figure 2. Similarly, this model does not perform especially well outside of the the training set.

We find that decision tree classifier models in general performed better than support vector machines in classifying school types based on debt and earnings data. A regular decision tree model set with increasing tree depths (Model 3) shows that the in-sample error quickly approaches zero (see Figure 3). However, out-sample error does not see much improvement. The decision tree does very well in separating within the training dataset, but it does not do as well with the test data. However, with additional bagging of varying base estimator sizes (Model 4), the problem of overfitting appears to improve. In fact, the out of sample error rate decreases from about 25% to 20%, when comparing decision tree classification without bagging and with bagging at different depths (see Figure 4).

We see a similar improvement with Model 5 (decision tree with ADA boosting), Model 6 (decision tree with gradient boosting) and Model 6 (random forest classifier). The increase of base estimators appear to have the greatest improvement on both in-sample and out-of-sample error when the tree maximum depth is small. In our examples, this is when it is set to 5. Figures 5 - 7 are line plots of the errors from Models 5 - 7 with varying parameter sizes.

4 Conclusions

In summary, we demonstrate that classification with decision trees were able to perform better than SVM on predicting school funding types with this particular dataset. Furthermore, decision trees can be improved with additional boosting and bagging techniques, with improvements depending on the maximum depth of the tree.

Further steps of this assignment would be to vary the training sizes to assess performance, and then possibly add in years of data prior to 2010 to increase size.

Python code used to generate the results, tables, and figures for this assignment can be found at `https://github.com/michellemho/machine_learning_for_cities`.

Table 1: Percentages of each school type represented in dataset (second subset)

Type	Percentage
Public	27.08
Private Non-Profit	45.05
Private For-profit	27.87

Table 2: Percentage of Misclassification, Model 1

Training Size Percent	In-Sample Error	Out-Sample Error
20	49.88	52.56
40	49.27	54.0
60	48.85	55.0
80	49.03	51.11

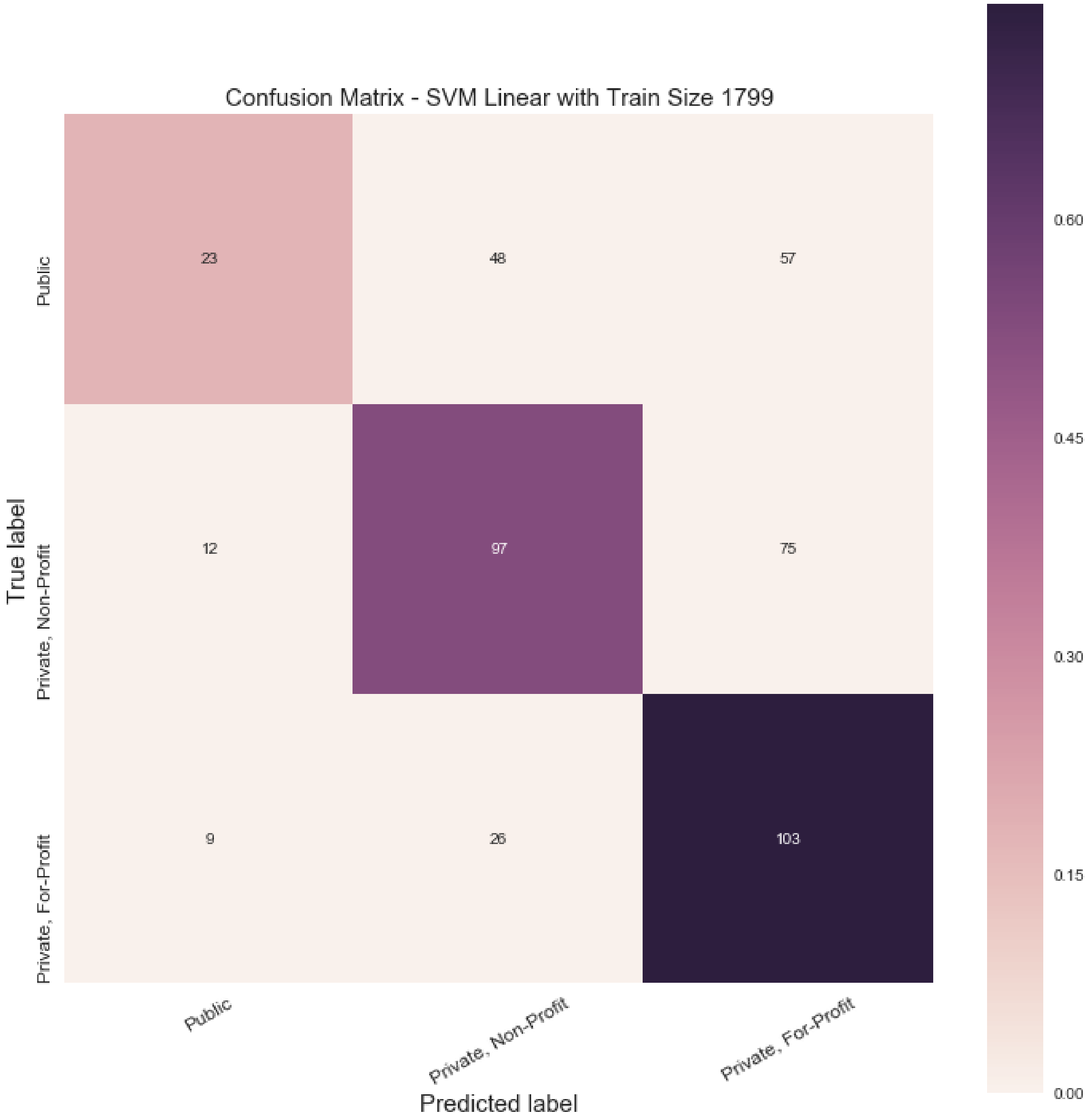


Figure 1: Model 1 Confusion Matrix

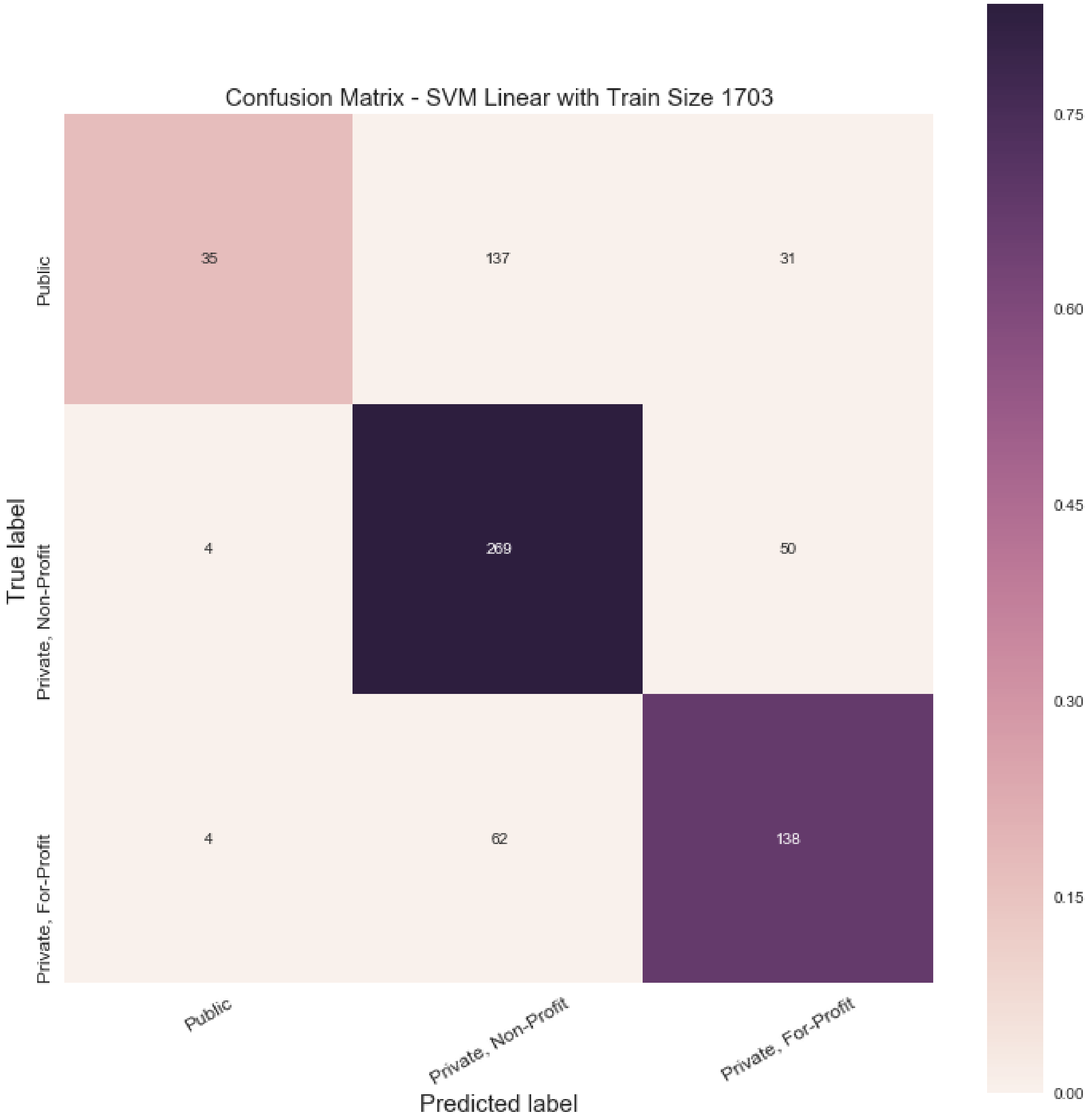


Figure 2: Model 2 Confusion Matrix

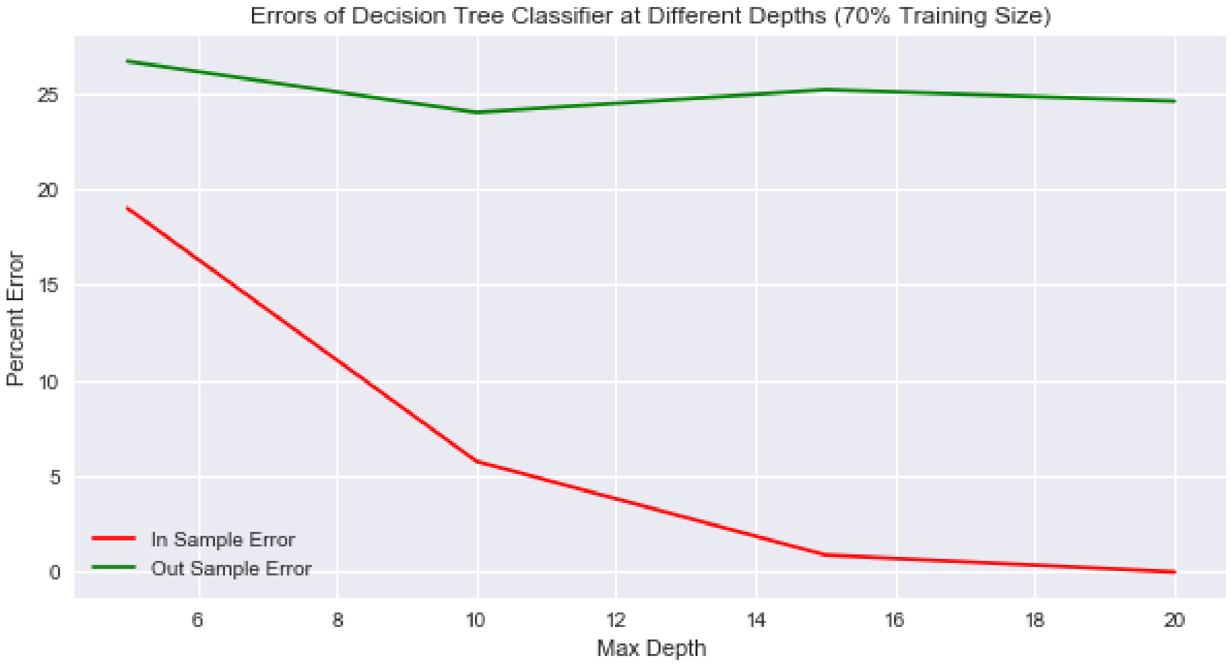
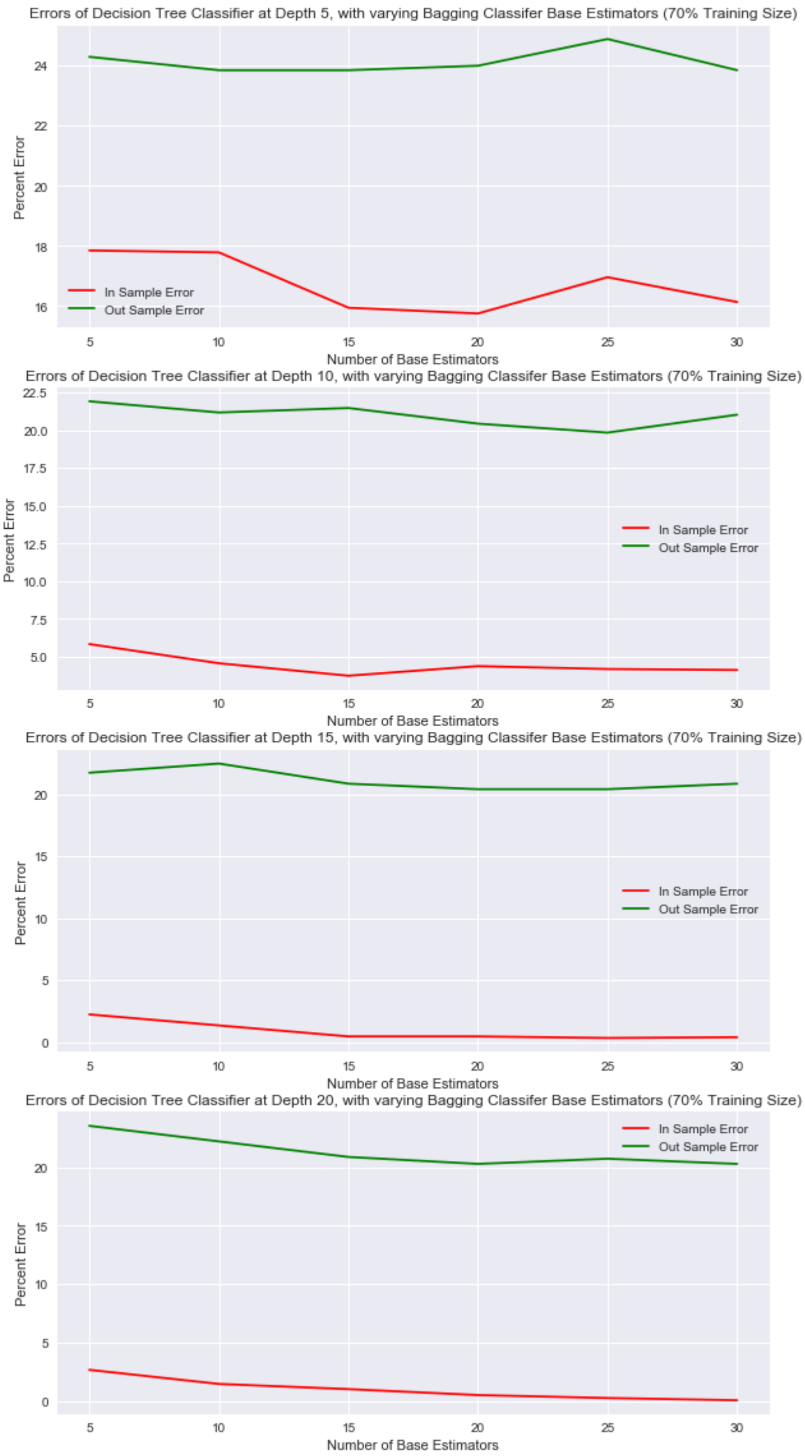


Figure 3: Model 3



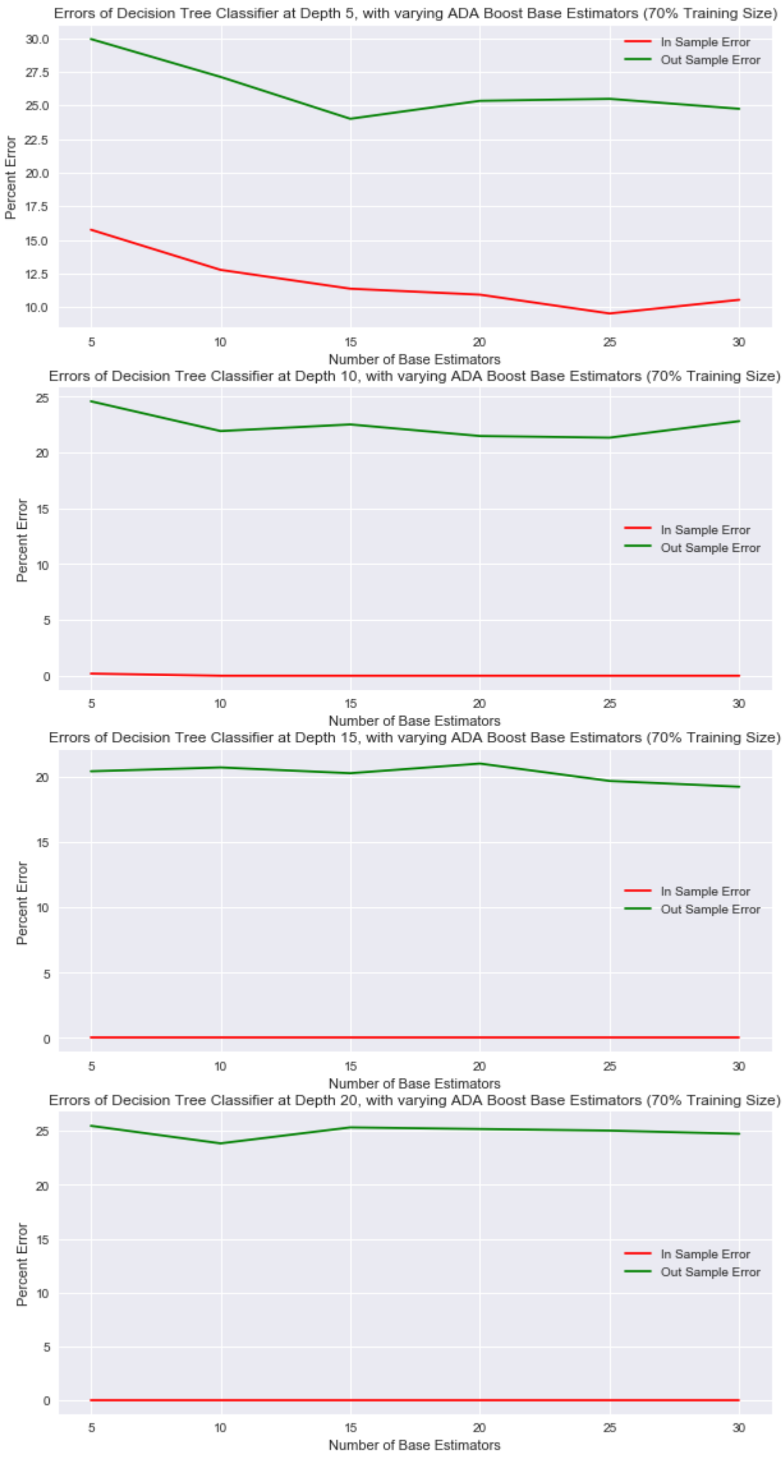
Errors of Decision Tree Classifier at Depth 5, with varying Bagging Classifer Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 10, with varying Bagging Classifer Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 15, with varying Bagging Classifer Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 20, with varying Bagging Classifer Base Estimators (70% Training Size)

Figure 4: Model 4

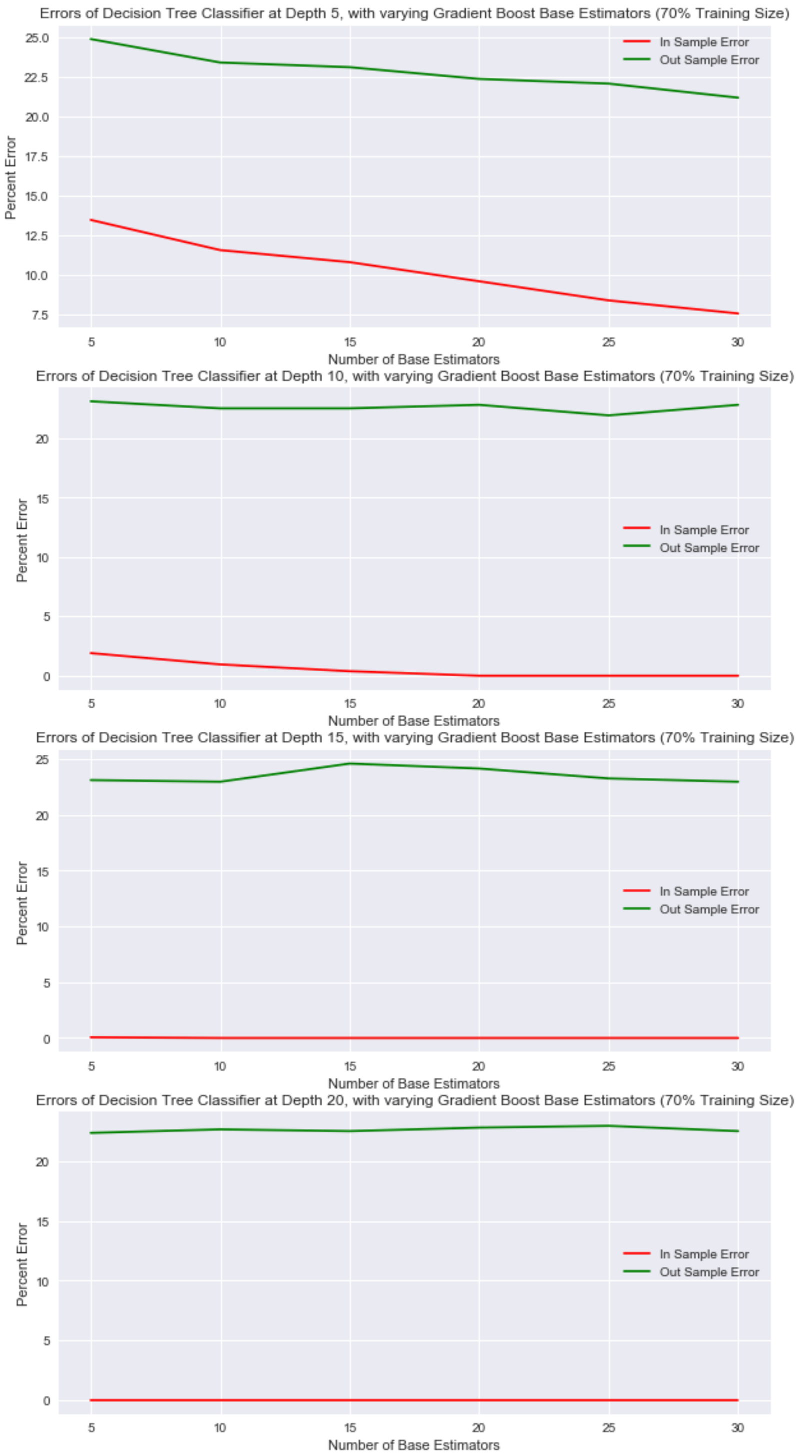

Errors of Decision Tree Classifier at Depth 5, with varying ADA Boost Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 10, with varying ADA Boost Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 15, with varying ADA Boost Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 20, with varying ADA Boost Base Estimators (70% Training Size)

Figure 5: Model 5

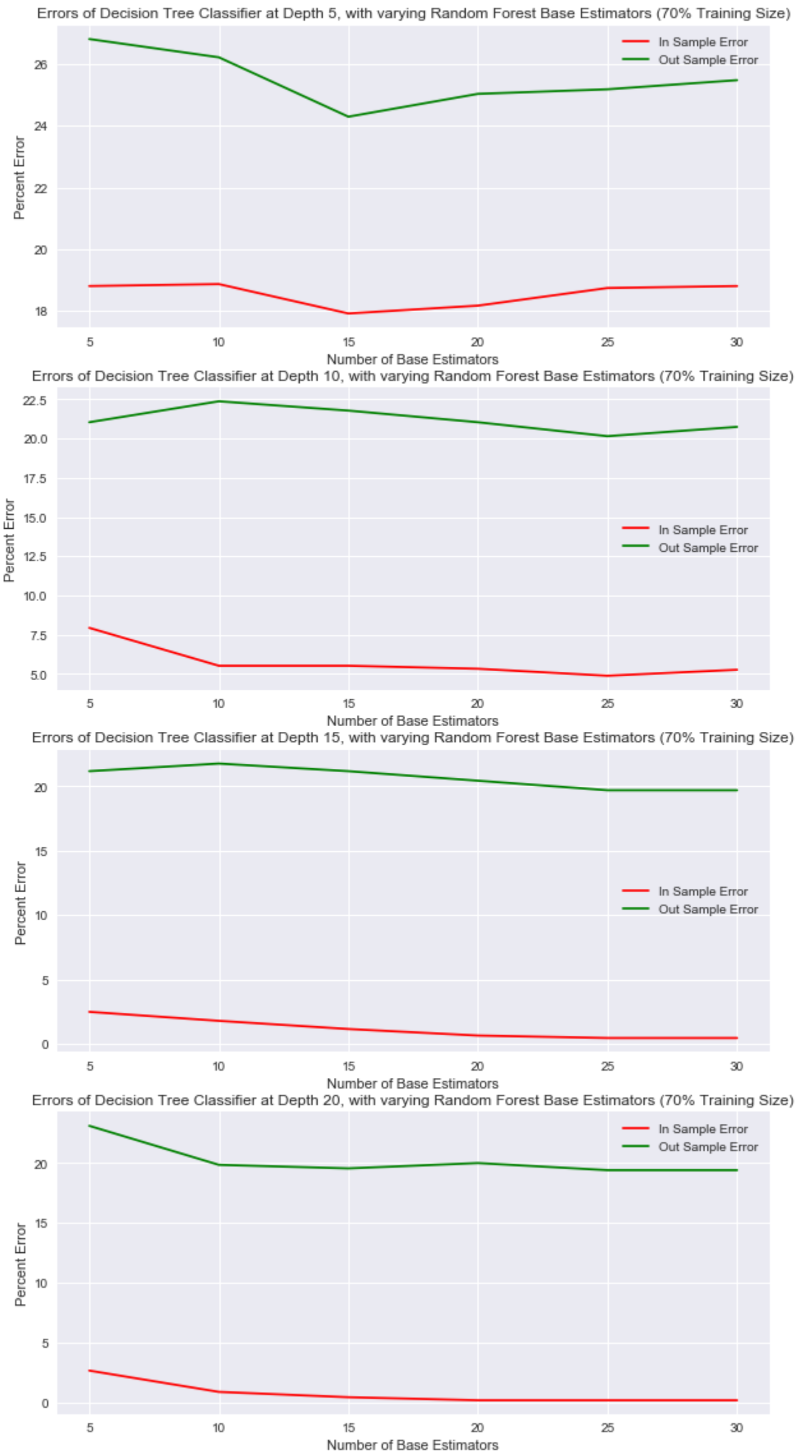

Errors of Decision Tree Classifier at Depth 5, with varying Gradient Boost Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 10, with varying Gradient Boost Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 15, with varying Gradient Boost Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 20, with varying Gradient Boost Base Estimators (70% Training Size)

Figure 6: Model 6



Errors of Decision Tree Classifier at Depth 5, with varying Random Forest Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 10, with varying Random Forest Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 15, with varying Random Forest Base Estimators (70% Training Size)

Errors of Decision Tree Classifier at Depth 20, with varying Random Forest Base Estimators (70% Training Size)

Figure 7: Model 7