Econ 626. ML for Economists
Assignment/Prediction Competition 1: Regression Trees

January 12, 2021

PLEASE CHOOSE ANY TWO OF THE FOLLOWING 4 SETS OF QUESTIONS. Questions 1 and 2 are easier than questions 3 and 4.

Question 4 is the prediction competition. Thus, in this assignment 1 the prediction competition is optional (in others it is not).

Due Friday Jan 22, 5pm. Please submit via Learn. Please submit both your code and the answer. Please submit your answers in a **PDF** file. Question 4 (the prediction competition) has separate instructions for how to submit the answers.

For each question, **please submit both the answer and the code.** Throughout the semester, students will learn from one another's prediction assignment/competition submissions. For this reason, please try to document your code so that the reader can follow the logic of. your submission without too much effort. Best answers to Q1, Q2 and Q3 will be distributed to class - students whose answer is selected will receive 1 bonus point for each answer. For this reason, please do not include your name anywhere on the submission (unless you want to). LEARN automatically adds your name to the filename when I download the submissions. If I share the file, I anonymize the filename.

Late submissions are ok (these are unusual times, I understand that). If I have started grading the other assignments, there will be a small penalty for a late submission.

You can use any programming language/statistical software package.

Collaboration is encouraged <u>but</u> **everyone must run their own code and write up their own answers.**

1a) [6 points] Using the Boston data set [ISLR], predict median house values

based on the average number of rooms per house using linear and polynomial models.

Use the validation set approach to compare the two prediction models: first divide data into two, a training sample and a validation sample. Report the training and validation sample MSEs for linear and polynomial models. Which of these models is the most accurate?

Hint. The necessary commands are found in ISLR 2.3.4, 3.6.2 and 5.3.1.

1b) [1 point] [challenging] Construct a graph that illustrates the numbers reported in 1 as a function of the flexibility of the model.

1c) [1 point] Now repeat 1a, but using leave-one-out-cross-validation.

Hint. The necessary commands are found in ISLR 5.3.2.

1d) [1 point] Now repeat 1a, but using k-fold cross validation.

Hint. The necessary commands are found in ISLR 5.3.3.

2a) [8 points] Using the Boston data set [ISLR] and a regression tree, predict median housing values based on the following three predictors: average number of rooms per house, average age of houses, and percent of households with low socioeconomic status.

First divide data into training sample and validation sample. Report a graph of the resulting regression tree and training and validation set MSEs.

Hint. All necessary commands for regression trees are found in ISLR 8.3.2.

2b) [1 point] Use cross-validation to examine how model accuracy varies as a function of the complexity penalty α. Provide a summary of the results and an assessment of which model is the most accurate.

Hint. All necessary commands are found in ISLR 8.3.1-2. (Note that 8.3.1 is on classification trees, which we study in more depth later in the semester; while some aspects of that section will be somewhat opaque now, reading it can still help complete the task).

2c) [1 point] Repeat 2a while allowing all available variables to be considered as

predictors.

3a) [5 points] Download the house value data for the analysis in Mullainathan and Spiess (2017), divide the data to training and test data (you choose how) and report mean and median house value for training and test samples.

3b) [5 points] Using the same data as in 3a, estimate a regression tree model and a linear model. Compare their performance relative to one another and relative to the results reported in the Mullanaithan and Spiess (2017) paper.

4) **Prediction Competition**

[10 points] This question is a prediction competition. Your task is to build a regression tree model to predict housing values. Please do **not** use bagging, boosting or random forests (we will study those later). Your algorithm will be evaluated based on the its predictive performance in a test set that will only be revealed after the algorithms have been submitted.

Timeline:

- The training data set is already posted on Learn. This data set has 10,000 observations on housing units (a subset of the Mullanaithan and Spiess sample). In addition to the logarithm of housing values, there are six other variables.
- Submission 1: Your model submission is due on **January 22, 5pm**. All you need to submit is the code that you use to train the model. **Please submit your answers as a PDF file.**
- The test data set will be posted on **January 22, 5.01pm**.
- After the test data set is posted, you will then calculate the accuracy of your model in the test set. Model accuracy is evaluated based on MSE. Please also calculate and report R^2.
- Submission 2: Accuracy of Algorithm 1 in the test dataset is due on Monday **due January 25, 5pm**. **Please submit your answers as a PDF file.**

The **top of Submission 2** must follow the following format:

- Name of submission [INSERT HERE The anonymized name of your submission, such as **BellKor** or **Joe97**. Please do not include your name or student id anywhere on the submission.]
- Accuracy of Algorithm 1 in test set: [INSERT HERE]
- Accuracy of Algorithm 1 in training set: [INSERT HERE]

  (Note: When reporting accuracy, please report both MSE and R^2)

  The rest of your PDF submission should include the code.

Notes on the data sets:

- The training and test data sets are comma separated. The response variable is the first variable (logarithm of housing value). You include any of the other variables as features in your model.