Econ 626. ML for Economists
Prediction Competition 2: Regression Trees, Cross-Validation, and Variable Importance

January 18, 2024

Answers are due on Wednesday Jan 31, 5pm.

- Your submission must consist of two parts: CSV file and PDF file.
- **The top of first page** of PDF must include (in this order!):
  1. Anonymized name (please do NOT write your own name anywhere – I can see it on Learn).
  2. R^2 and MSE for the training set.
  3. Graphs for Q2, Q3 and Q4 as calculated from the training data.
  4. Screenshot of an example from ChatGPT/GPT4 interaction.
  5. The rest of PDF must include **code** for Q1, Q2 and Q3 answers.
- The CSV file must include the following:
  - line 1: student id number (so TA can connect your predictions to your name)
  - line 2: anonymized name (for the class leaderboard)
  - line 3: R^2 in the training data (typically a number between 0.00 and 1.00, though could be negative if model really bad). Remember: it does NOT matter how high/low this number is.
  - lines 4 through 10,003: one prediction for observation in the test set.

  Again, the CSV file must have one prediction for every observation in the test set, and nothing else (e.g. no variable names or other headers). Hence, given that the test set has 10,000 observations, the CSV file must have $3 + 10,000$ lines (not a line less, not a line more).

Best PDF answers will be distributed to class. Students whose answer is selected will receive 10 percentage point bonus.

You can use any programming language/statistical software package.

Collaboration is encouraged but **everyone must run their own code and write up their own answers.** You are always free to use ChatGPT/GPT4/Other LLMs in any way you consider useful (to help in writing, coding, analysis, etc.)

Q1. [6 points] This question is a prediction competition.

Use the regression tree algorithms to build a prediction model for predicting 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 predictions have been submitted. Accuracy is measured by MSE in the test set.

Data:

- **The training data** set is posted on Learn. This data set has 20,000 observations on housing units (a subset of the Mullanaithan and Spiess sample). In addition to the logarithm of housing values, there are 13 other variables that you can use as model features or to build model features.

  The size of the data set may make your estimation very slow. Machine learning is always about making compromises because of limited computing resources. So please be prepared to potentially drop variables or observations when you estimate the model.

- **Data on the test set** is also posted on Learn, but without observations on the response variable. The TA will calculate this MSE based on the predictions in your CSV file and the actual values of the response variable in the test set (which have not been distributed to students).

- Remember, **model accuracy is evaluated based on MSE between your predictions for these test set observations and the actual values of the response variable in the test set.**

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

Grade will be an increasing function of the algorithm's performance in the test data set. To receive credit for a submission, the code must be reported clearly enough to enable replication. Please include a brief explanation of each step so that the reader can follow the logic of code easily.

Final note: you have to produce a prediction for all 10,000 observations in the test set. Hence, you cannot skip observations for which values of some features are missing. This means that when estimating the model using training data, you want to make sure the model gives a prediction for all 20,000 observations in the training data – including those observations for which values of some features are missing.

Q2. [1 point]

Construct a graph that shows actual values of the response variable on the horizontal axis and predicted values of the response variable on the vertical axis.

Q3. [1 point, challenging]

Construct a graphical illustration of the relative importance of different features in your prediction model. The importance is measured based on each variable's importance in predicting housing values **in the training data**.

Hint: You are trying to do replicate ISLR Figure 8.9 for your model and this data set.

Q4. [1 point, challenging]

Estimate regression trees of varying depth and construct a graph that show "training error" (MSE) and "test error" (MSE) as a function of model depth. Here you calculate both using the available training data (please divide data into two sets for this demonstration).

Q5. [1 point] Ask ChatGPT/GPT4 for help on how to best answer Q1, Q2, Q3 or Q4. Report a screenshot of the most useful/interesting interaction.