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

September 20, 2023

Answers are due on Thursday Sept 28, 6pm.

- Your submission must consist of two parts: CSV file and PDF file.
- **The top of first page** of PDF must include (1) Anonymized name; (2) the R^2 and MSE for the training set, (3) the answer for Q2 as calculated from the training data, (4) Screenshot of an example from ChatGPT/GPT4 interaction. The rest of Submission 1 must include code for Q1 and Q2 answers.
- The CSV file must include 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 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.**

Q1. [**8 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. (Please also report R^2)

Timeline:

- 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. **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 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).

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.

Notes on the data sets: 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.

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, challenging]

Construct a graphical illustration of the relative importance of different features in your final 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 this data set.

Include this FIGURE in your submission, immediately after your report on the accuracy of your model in training dataa

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