Econ 424. ML for Economists
Prediction Competition 4: Ensemble Methods and Feature Construction

October 17, 2023

Answers are due on Thursday Oct 26, 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 to be shown on class leaderboard, (2) MSE and R^2 in the training data, (3) name of algorithm used, (4) answer to Q2 (one figure and 1-2 sentences), and (5) answer to Q3 (ChatGPT interaction screenshot), and (6) code that can be used to **fully replicate how you constructed the predictions**.
- 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 100,000 observations, the CSV file must have 100,000 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.** As always, you can utilize ChatGPT/other LLMs in any way you wish.

The following introduces the data set.

There are two training data sets posted on learn on used car prices and car characteristics: "small" (200,000 observations) and "large" (1,000,000 observations). The former is a subset of the latter data set. You can build predictions either based on the small or the large data set. The data are comma separated.

The test data without response variable have also been posted. These data have 100,000 observations.

The original data was downloaded from Kaggle. You can use any resource you find on Kaggle, but please do not try to download more data from Kaggle to help with prediction.

Some feature variables may have missing values. In the prediction competition you cannot skip observations with missing values on some features: **you must give some prediction for all 100,000 observations in the test set** (and for all observations in the training set you utilize).

Q1. [**8 points**] This question is a prediction competition.

Utilize either training data set to train a model that predicts the **logarithm of car price** (first variable in the training sets, missing in the distributed test data set without response variable).

Utilize any algorithm. Feel free to use bagging, random forests or boosting.

Utilize any features already included (such as Year and mileage) or any features that can be constructed. .

Accuracy of your model will be evaluated based on **the prediction accuracy** in the test data set, where accuracy is measured by MSE.

As emphasized above, you must produce a prediction for every observation in the data sets that you utilize.

Q2) [1 point] Draw a figure that illustrates how well your model predicts for different values/ranges of the response variable (for example: high, medium, low) in the training set. One way to do this is to draw a figure where values of the response variable are on the horizontal axis and values of predictions are on the vertical axis. You can consider other ways too, please provide only one figure.

Discuss the implications of this pattern in 1-2 sentences.

Q3) [1 point] Demonstate how ChatGPT/Other LLM can be useful in answering Q1-Q2, either in coding or in designing the approach.