Econ 626. ML for Economists
Prediction Competition 5: Ensemble Methods, Algorithmic Bias and AI Alignment

February 27, 2024

Answers are due on Thursday March 7, 6pm.

- 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. **MSE and R^2 in the training set.**
  3. Name of ML algorithm used in Q1.
  4. Answer for Q2 (four numbers numbers, one graph and one sentence explanation).
  5. Answer for Q3 (one screenshot)
  6. Answer for Q4 (two screenshots)
  7. Answer for Q5 (Screenshot of an example from ChatGPT/GPT4/Other LLM interaction.)
  8. The rest of PDF must include **code** for Q1 and Q2.
- 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.
  - line 4: **Name of ML algorithm used for Q1.**
  - lines 5 through 50,004: one prediction for every observation in the test set. **Please remember: the predictions are for the natural logarithm of price.**

  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 50,000 observations, the CSV file must have $4 + 50,000$ 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.)

The following introduces the data set.

The training data are the same car price data as in prediction competitions 3 and 4. You can also utilize prediction competition 3 and 4 test data.

There is also **an additional training data file in PC5 folder**, which you can choose to utilize as well. This new training data have $100,000$ observations and must be utilized in Q2.

The test data file without the response variable is also posted in PC5 folder. These data have $50,000$ 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 50,000 observations in the test set** (and for all observations in the training set you utilize).

Q1. [7 points] This question is a prediction competition.

Utilize any available training data set to train a model that predicts the **natural logarithm of car price**.

You are free to utilize almost any algorithm, including random forests, bagging and boosting.

You are **not allowed** to utilize neural network algorithms.

Utilize any features already included (such as year and mileage) or any features that can be constructed from the available numerical and text data. Obviously, you cannot construct features from the price variable.

Accuracy of your model will be evaluated **MSE** in the test data set (please also report R^2).

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

Q2) [1 point] Train the same algorithm you utilized in Q1 but using only the training data posted in PC5 folder ($100,000$ observations). Set $20,000$ of these observations aside as a validation set.

Now train the same algorithm but using the small training data in PC3 folder ($100,000$ observations). Again, set $20,000$ of these observations aside as a validation set.

Report R^2 from both models **in both validation sets**, so 4 numbers in total, and explain the differences **using one graph** that shows how the distributions of observations differ in these two training samples.

Q3) [0.5 points, challenging] Use a text-to-image generative AI program (e.g. DALL-E, Midjourney, etc) to draw a figure on a situation that is not well presented in the training data. (Examples: horse sitting on a human in a bowl, person writing with their left hand).

Give an example where the program fails (i.e. the image does not capture the spirit of the request).

Please try to be original; points are not given for examples that are submitted by more than 3 students.

Q4) [0.5 points, challenging] Use two different large language models (e.g. GPT, Gemini, Grok) to show that one model can refuse to answer a question that another does answer.

Please make sure the request is **related to economics** in some manner (i.e. economic policy, human behavior, human interactions, economic theory, econometrics).

Q5) [1 points] Demonstate how ChatGPT/Other LLM can be useful in answering Q1-Q4, either in coding or in designing the approach.