Econ 626. AI in Economics

Prediction Competition 4: Ensemble Models and the Prediction Error Distribution

January 6, 2026

Submission deadline is posted on the Learn Dropbox.

- Your submission must consist of three parts: CSV file, .py (or R) file and PDF file.
- **The PDF must include** (in this order):
  1. Anonymized name (such as "BellKor97"; please do NOT write your own name anywhere – I can see it on Course Website).
  2. Graph for Q2.

The CSV file must include the following:

- line 1: anonymized name (for the class leaderboard)
- line 2: student id number (so TA can connect your predictions to your name)
- line 3: Prediction accuracy in the training data (R^2; typically a number between 0.00 and 1.00; Remember: it does NOT matter how high/low this number is.
- line 4: Name of algorithm used (this also does not influence grade; only accuracy matters).
- lines 5 through 100,004: one prediction for every observation in the test set, **in the same order as the observations are in the "test data without response variable" file.**.

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 100,000 observations, the CSV file must have $4 + 100,000$ lines (not a line less, not a line more).

Best answers will be distributed to class. Students whose answer is selected will receive 10 percentage point bonus. An LLM may be used to evaluate student answers.

Anonymized answers to Q2 may be shared with class for instructional purposes.

You can use any programming language/statistical software package.

Collaboration is encouraged, even sharing code is okay. 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 sets.

There are two training data are the same as the training data for PC3. You can also utilize PC3 test data as training data.

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.

Predict the **natural logarithm of the price** of used cars.

Important constraint: You can use any algorithm **except** neural networks. You can also use any combination of algorithms.

Performance of your model will be evaluated based on the **MSE** in the test set.

Q2. [2 points] Draw a figure that demonstrates the how prediction accuracy varies across the distribution of the response variable **(in the training data)**. That is, draw a figure where values of the response variable y are on the horizontal axis and predictions $\hat{y}$ are on the vertical axis. Or, draw a figure where values of the response variable y are on the horizontal axis and prediction errors $\hat{u}$ are on the vertical axis.