Econ 626: ML for Economists
Prediction Competition 4: Penalized Regression and Data Visualization

February 6, 2024

Answers are due on Friday Feb 16, 11am.

- 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. Graph for Q2 as calculated from the training data.
  4. Graph for Q3 as calculated from the training data.
  5. Screenshot of an example from ChatGPT/GPT4 interaction.
  6. 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 100,003: 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 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.)

The following introduces the data set.

(The training data are the same as in prediction competition 4. You can also utilize prediction competition 4 test data after it is posted on Monday Feb 11.)

There are two training data sets posted on learn on used car prices and car characteristics: "small" (100,000 observations) and "large" (500,000 observations). The former is not a subset of the latter data set. You can build predictions based one or both training data sets. 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. [7 points] This question is a prediction competition.

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

Utilize only regression algorithms: linear regression, LASSO, Ridge, Subset Selection.. Please do **not** use bagging or boosting. (And please do not use other ML algorithms to transform variables; though you can create new variables and use simple "averaging techniques" to transform and normalize variables).

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] Produce a graph that measures the variable importance of your prediction model (this the same as Q3 in PC2).

Q3) [1 point, challenging] Produce **one graph with multiple panels** that summarizes: (1) the distribution of each variable (a separate panel for each variable) and (2) the link between each feature and the response variable (a separate panel for each variable pair).

When plotting the distribution of a feature, please plot in the same graph the distribution in the training data and in the test data, so that we can see whether those are similar.

Q4) [1 point] Demonstate how ChatGPT/Other LLM can be useful in answering Q1-Q3, either in coding or in designing the approach.