Econ 424. ML for Economists
Final Prediction Competition: Deep Learning, Text Analysis, and Causal Insights

October 23, 2024

Answers are due on Monday, December 9th, 5pm.

- Your submission must consist of three parts: CSV file, PDF file, and .py (or .R) file.
- The PDF must include (in this order):
  - anonymized name to be shown on class leaderboard
  - MSE and R^2 in the training data
  - which type of model was estimated
  - answer to Q2 (1 figure)
  - answer to Q3 (5 or more figures)
  - answer to Q4 (brief explanation and a table or figure)
- The .py (or .R) file must include:
  - code for replicating answers to questions Q1-Q4
- 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; remember this number does not matter for leaderboard position).
  - line 4: Name of ML algorithm(s) used; if a Deep Learning model is used, please specify which kind of DL model it is.
  - lines 5 through 100,004: one prediction for every observation in the test set. Predictions must be numbers between 1 and 5 but do **not** have to be integers. Predictions must be in the same order as the observations are in the "test set without response variable" data set.

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.

The data are Glassdoor reviews. Training data sets (small: 100,00 observations, large: 500,000 observations; small is not a subset of large; you can use both training sets) are posted on Learn. The test data set has 100,000 observations. Test data set without the response variable is also already posted on Learn.

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**.

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

Construct features from the available data. Then use these features to predict the overall rating **using any machine learning model**, including neural networks.

You can utilize either or both training data sets to train the model.

Performance of the algorithm is measured by the mean squared error (MSE) in the test set. Please report also R^2 in the training set.

Q2) [1 point] Draw a figure that shows both the structure of the model and what features are used as inputs.

Q3) [1 point] a) Illustrate in one graph (may have multiple panels), how the content of the available text varies across different review star-levels. One possibility is to calculate frequency of words that are typical of each category: 5-star reviews, 4-star reviews, 3-star reviews, 2-star reviews and 1-star reviews, and then draw 5 word clouds. But there are likely better options.

b) Produce graphs to illustrate the following (we have drawn all of these in previous prediction competitions):

- Importance of each feature.
- Correlation between features.
- Distribution of each feature in the training data versus in the test data.
- How large and common are prediction errors at different points of the distribution of the response variable in the training data (in other words, contrast distributions for y vs. $\hat{y}$, or distributions for y versus u).

Q4) [1 point, challenging] Economist Sarah Bana trained a convolutional neural network model to predict wages based on text. (Paper is available here.) The analysis then uses the machine learning model to calculate the value of different certifications for a worker, and how the valuations for these certifications vary across the wage distribution.

Consider the Glassdoor data. Use your own model to produce **actionable insights** into what firms at different points of the quality distribution (as measured by average overall rating) can do to improve their rating.

Example 1: Perhaps most of the complaints on low-quality workplaces are about lack of restrooms.

Example 2: Perhaps most of the complaints about high-quality work-places are about lack of sushi.

Example 3: Perhaps most of the praise for high-quality restaurants are about ability to work from home.

In your answer, please briefly explain your methodology and then show a figure/table that summarizes the results neatly.

Also discuss whether your approach likely reveals a true causal link between a variable (such as number of restrooms) and the average perceived quality of the employer. Please explain why or why not in one paragraph. It's important to try justify the answer ("yes, we uncovered a causal link" or "actually, it's not necessarily a causal link") well.

The approach used for this Q4 does not have to be related to approach used in Q1-Q3.