Econ 424. ML for Economists
Prediction Competition 6:
Neural Networks and Text Analysis

November 13, 2023

Answers are due on Wednesday, Nov 23, 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) which type of a neural network model was estimated: describe the model including number of layers, number of nodes and total number of parameters. (4) answer to Q2 (three figures), 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 56,976 observations, the CSV file must have 56,976 lines (not a line less, not a line more). **Even though all response variable only takes integer values, predictions do NOT have to be integer values.**

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.

The data are Glassdoor reviews. Training data sets (small: 110,729 observations, large: 553,678 observations; small is a subset of large) are posted on Learn. The test data set has 56,976 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 56,976 observations in the test set** (and for all observations in the training set you utilize).

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

Construct 5 features from the available data. Then use these 5 features to predict the overall rating **using a neural network model**. You can construct more features but final predictions must be calculated based on

exactly 5 features. You can estimate any type of a neural network algorithm. If you are unable to calculate predictions using a neural network model, you can use another model but grade will be maximum 85%.

You can utilize either training set (small or large) to train the model.

Performance of the algorithm is measured by MSE in the test set. Please report MSE and R^2 for the training set.

Q2) [2 points] Produce a graph that shows the distribution of each feature in the training set and in the test set.

Produce a second graph that shows 1) the distribution of predictions in the training set, (2) distribution of actual values in training set and (3) distribution of predictions in the test set.

Produce a third graph that shows the correlation between each feature and between each feature and the response variable in the training set. This can be a correlation heat-map or a more detailed visual presentation.

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