# Econ 424. ML for Economists
# **Final Prediction Competition**

November 15, 2023

Answers are due on Thursday, December 14, 5pm.

- Your submission must consist of two parts: CSV file and PDF file.
- **The top of first page** of PDF must include:
  - anonymized name to be shown on class leaderboard
  - prediction accuracy in the training data
  - which type of model was estimated
  - confusion matrix for training data (a table)
  - answer to Q2 (5 figures)
  - answer to Q3 (brief explanation and a table or figure)
  - answer to Q4 (ChatGPT interaction screenshot)
  - 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). Given that response variable only takes integer values, and predictions are measured by number of correct predictions, **all predictions should be integer values 1, 2, 3, 4 or 5**.

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; the training data are the same data as in prediction competition 6. Training data sets (small: 110,729 observations, large: 553,678 observations; small is a subset of large) are posted on Learn. The new 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 features from the available data. Then use these features to predict the overall rating **using any machine learning model**.

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

Performance of the algorithm is measured by the number of correct predictions (accuracy percentage). Please report prediction accuracy in training set and the confusion matrix.

Q2) [1 point] Produce graphs to illustrate the following (we have done all 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 other words, contrast distributions for $y$ vs. $\hat{y}$, or distributions for $y$ versus $u$).
- One more useful graph selected by student/ChatGPT.

Q3) [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.

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**.

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