Econ 626. ML for Economists
Prediction Competition 6: Text Analysis, Economic News and ML Research

February 22, 2024

Answers are due on Wednesday March 20, 6pm.

- 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. Description of approach used for Q1 (one paragraph).
  3. Answer for Q2 (two figures).
  4. Answer for Q3 (one figure)
  5. Answer for Q5 (Screenshot of an example from ChatGPT/GPT4/Other LLM interaction.)
  6. The rest of PDF must include **code** for Q1, Q2 and Q3.
- 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)
  - lines 3 through 15,580: one prediction for every observation in the test set. Please remember: the predictions are either 1 or 0, and **half of the predictions must be 1, half must be 0.**

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 data have 15,578 text snippets from news articles. Half of the original news articles include the words "economic","economy", "economics" or the character sequence "econom" (not case-sensitive) in it.

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

Your task is to predict which articles have the above characteristic. Predict 1 for articles with character sequence "econom" and perdict 0 for articles without character sequence "econom". **Half of the predictions must be 1, half must be 0.**

You can ask ChatGPT/OtherLLM for general help but you are **not allowed** to try find the articles themselves (either with Google or databases).

You are free to utilize almost any supervised or unsupervised algorithm, including neural networks.

Accuracy of your model will be evaluated based on prediction accuracy percentage.

You must produce a prediction for every observation in the data set.

Q2) [1 point] Draw separate wordcloud figures for text snippets with "econom" in it and for text snippets without "econom" in it. Before the analysis, remove common stop words.

Q3) [1 points] Draw one figure that allows you to contrast which words are relatively more common in text snippets with "econom" in it than for text snippets without "econom" in it, and vice versa.

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