Economics 899: Machine Learning & AI in Econ
University of Wisconsin - Madison
Fall 2024
Project Description

Instructions:

- The purpose of the assignment is to have you demonstrate that you can apply the techniques learned in the course to an important real-world setting in order to obtain answers to economic questions of interest.
- You will work in groups. You should form groups of 4-6 people as soon as possible, and fill out the Google form with your group information by **Sept. 30, 2024**.
- This document contains an outline of the project's topic and the expected form of the final report.
- Your group will submit a final report and relevant code by the end of the day on **December 12, 2024**. You should submit only one report for each group.
- The project incorporates strategies from some of the major topics in the course (i.e. data visualization, supervised learning, unsupervised learning) in a way that is appropriate to answer the question of interest. While technical skills in implementing ML techniques are important, the best projects will also motivate carefully the rationale for the analysis, and provide a compelling explanation of the results. See below for more guidelines on evaluation.
- While we are using `R` as a programming language for problem sets during the course, you may use any statistical software to produce the analysis in your report. However, we will not be providing support in all languages — only use a language other than `R` if you already know it well, and will not need support in conducting the analysis.

Topic: Predicting the Price Effects of Airline Mergers

- Industry structure and concentration can have very important consequences on consumer welfare and firm profits. At the core of antitrust policy lies a prediction problem: predicting the price effects of mergers. This is of the highest importance for antitrust authorities that need to evaluate whether to bring enforcement action. In making this decision, an agency needs to rely on the best predictions available. Consult the horizontal merger guidelines for further information on this process in the US.

- The goal of this project is to predict the effects of a counterfactual merger in the airline industry. You will be working with the airline data that you have assembled and worked on during the semester.

- Mergers in the airline industry (and what to do about them) have been intensely debated in the last few years, and continue to be at the forefront of the economic policy debate. Most recently, on July 28th, 2022, Jet Blue reached a deal to acquire Spirit Airlines; however, this was abandoned in spring 2024 as the Justice Department attempted to block the merger.

- Using the data and the tools developed in the course, you will first be constructing an algorithmic model to conduct a merger retrospective, i.e. retrospectively fit a predictive model to a consummated merger, and check its accuracy against the actual path of prices post-merger.

  - To do this task, you should pick one of the mergers from this list that took place between 2005 and 2018. Justify why you picked a certain merger, in light of your assigned counterfactual merger prediction exercise.
  - There is no single technique that is unambiguously best suited for this task; potentially nonlinear supervised learning algorithms such as random forests or neural nets may perform well.

- One of the key conceptual steps you have to take to make predictions in this context is to formulate a model that (i) can be estimated in the pre-merger period, and (ii) can be used to extrapolate predictions in the post-merger period. More concretely, there need to be variables or market characteristics that affect prices and are observable in the pre-period, and can be manipulated to obtain the predictions that correspond to the post-period (which is not in the data). While ultimately you have to choose how to do this in your paper, here are some ideas:

  - You can categorize markets based on the presence of airlines (e.g., duopoly of major airlines, or three majors and two low-cost carriers present, etc.), and use this categorical variable as a predictor; this has the advantage that it is easy to predict how the merger will affect this variable (at least in the short term): if a market is a duopoly of two majors, and these merge, that market will become a monopoly.
  - You may think of using (contemporaneous) quantities to summarize market structure and predict prices; this poses a few challenges. First, you do not know how the merger will affect quantities. Thus, you would need to make assumptions (e.g., does the full pre-merger market share go to the newly merged airline?), but these are bound to be imperfect. Similar considerations apply to using HHI to synthesize market structure, as it is based on market shares.

- Among other important aspects that you have to address are:
  - Analyzing the data in the post-merger period, what are the overall effects of the merger? What are the markets most impacted by the merger? Consult Carlton et al. (2019) for ideas on how to do this.
  - How does your model perform? What are its strengths and weaknesses? Which features of the post-merger prices does it predict well, and which ones does it miss?

- Based on your results in the retrospective exercise, you will then predict a counterfactual merger. Each group will be assigned a different counterfactual merger. Given that Covid has fundamentally affected the airline industry, you should imagine that merger is proposed for October 1, 2019 and that no data exists after that (you would have a very hard time predicting the Covid period!). This is a merger that did not happen in the data; the exercise is meant to mimic the task of regulators who need to evaluate a proposed merger.

- Among the important aspects that you have to address are:
  - Using your predictive model, what would be the markets most impacted by the merger? Does that square with economic intuition?
  - Describing predictions across all markets may be difficult. Instead, you should produce predictions for representative markets in the data that best illustrate the economic effect of the merger on prices. To do this, unsupervised learning

algorithms presented in the course could be useful. Explain and justify how and why you chose your markets.

- There is a well-developed merger-simulation toolkit used by economists working in Industrial Organization. This toolkit has been used to evaluate mergers in the airline industry. Read Peters (2006) and references therein to familiarize yourself with this approach.
  - You should *not* use this approach in your paper; we are providing you with this reference to contrast a different, more theory-driven approach to merger simulation, to the more data-driven exercise you will execute.
  - Write a brief non-technical summary of the economic approach to merger simulation.
  - When you compare the economic approach in Peters (2006) and the predictive model you constructed, what are the respective strengths and weaknesses? When do you expect one method to work better than the other?

Carlton, D., M. Israel, I. MacSwain, and E. Orlov (2019): "Are legacy airline mergers pro-or anti-competitive? Evidence from recent US airline mergers," *International Journal of Industrial Organization*, 62, 58–95.

Peters, C. (2006): "Evaluating the performance of merger simulation: Evidence from the US airline industry," *The Journal of law and economics*, 49, 627–649.

Final Report:

- The report should contain the following sections

    1. Abstract summarizing the question and results (no more than 100 words)
    2. Introduction
    3. Data section summarizing the data set, providing relevant summary statistics, etc.
    4. A section on the merger retrospective, including a description of the merger, of the methodology, and of the results. Explain clearly the techniques that you used and why you used those techniques. Discuss the results, and relate them to economic intuition.
    5. A section on the counterfactual merger prediction results, presenting and explaining your main findings, *including a discussion of the caveats in interpreting the results.*
    6. A section on the comparison with the economic approach to merger simulation.
    7. Conclusion
    8. Do not include printouts of your code, but attach to your submission the programs necessary for replication

- **The report should be around 15 pages long (not including the appendix)**
- Follow all of the style guidelines that appear at the end of this document
- Only include material that is relevant to the argument in your paper. Some students have a tendency to include as much course material as possible in the hopes of showing every fact learned from lecture. True mastery of course material involves not only being able to recall facts and figures from lectures, but knowing when they are actually relevant and applying them appropriately. You should attempt to build a tight argument in the paper, without extraneous material.
- Similarly, when deciding the methods you want to use, more sophisticated does not always mean best. Part of demonstrating knowledge is being able to choose the best tool available to answer the research question, even if it is not the most sophisticated. You should not aim to include every technique you have learned in the course, instead choose the ones appropriate for answering your question.

Evaluation Guidelines

- Successful projects fully develop the assigned topic, including merger retrospective, counterfactual prediction, and comparison with the economic approach.
- You should work hard on both form and substance: the report should be clear and in line with the standards of professional writing for economists. See the style guidelines below for hints, and compare the style and formatting of your paper to the references in the syllabus.
- While precision and technical expertise is necessary in choosing and estimating your model, and producing predictions with it, we particularly emphasize that mature understanding of what you are doing, and why, is the hallmark of a successful project. When evaluating a project that employs the most sophisticated tools, but with limited understanding of the underlying data and poor interpretation of the results, versus a project that uses basic tools, but demonstrates full understanding of the data, technique and results, we will almost surely give a higher score to the latter.
- We give substantial weight to the quality of the discussion of the results that you obtain: what are their limitations and strengths? Why? Are there other avenues of analysis that, although not pursued in the current report, may yield interesting results?
- The best project reports reflect a journey of learning: if you worked hard at the project, you learned many interesting and insightful things about the topic of interest along the way. Make sure you share with the reader your most interesting findings, and guide them through a well defined line of reasoning.
- Successful projects always result from the full involvement of everyone in the group. Specialization of different individuals to different tasks is efficient and necessary, but everyone should also be on board with the overall direction of the project, and contribute their views and expertise where needed. Lack of cohesion or free-riding within the group results in projects that read like collection of disconnected pieces, and these typically get low scores.

Style Guidelines

These style guidelines are a slightly modified version of those for Econ 706.

In reading the papers, we pay considerable attention to style (correct spelling and grammar, clear exposition, good organization). So, too, do most people: reports that are difficult to read routinely get ignored, even if they contain good ideas. Thus, it will pay to develop the habit of working hard to craft a clear explanation of your ideas.

Many of the following suggestions are standard good practice. Others are matters of taste.

1. In writing up a research report, one should have an audience in mind. We suggest that you take the audience to be your fellow students in this course. They'll know most of the relevant economics and methodology, but won't know your data set or your approach.

2. Include a cover page with the following information: Title; date; your name; your e-mail address; the word "Abstract"; an abstract of 100 words or fewer. If you have acknowledgments to make (thanking a fellow student for helpful comments, for examples), put these on the bottom of the cover page. The text of the paper begins on the next page.

3. Your paper should be divided into sections, to help guide the reader.

4. Number the pages. No plastic covers or binders, please.

5. In the introduction, present an overview of your paper and summarize your findings. In the conclusion, give suggestions for future research.

6. Be explicit about your data set. State the sample size. For time series, state whether the data are monthly, quarterly or annual, and whether or not they are seasonally adjusted. State the units of measurement. For example, if ?income? is a variable, state whether it is measured in current dollars or constant 2008 dollars, and if it is per capita, say so. For data in logs or log differences, you will usually want to multiply by 100 so that the units will be percent or percent change. Also explain the choice of the sample: why does it start in a given year, or you use only a cross-section from a given year instead of a panel, and so on.

7. You will want to include a plot and/or a table with basic statistics (means, standard deviations) of the data.

8. Number the equations. You can number the equations by section, if you have sections in the paper. That is, the third equation in section 2 can be numbered 2-3, the second equation in section 4 can be numbered 4-2, etc., if you prefer doing this to numbering sequentially through the paper.

9. Tables:

   (a) Number the tables, and on each include a descriptive header ("Means and Standard Deviations of Data," or "Variance Decompositions," for example).

   (b) Tables may appear in the text in the appropriate place, or at the end of the paper.

   (c) Tables should not run over page boundaries, unless they are too long to fit on a single page. That is, if you include a table in the text, you should insure that you place it so that it does not run from one page to the next.

   (d) Make every effort to make each table self-contained, even though this will require you to redundantly present information that is also stated in the body of the paper itself. This is now the standard in the profession and you should look at a paper published recently to see how much detail is included in tables.

      i. In notes at the bottom of each table, define the symbols that are in the table, or give a precise reference to where the definition may be found. It is not adequate to simply state "definitions are in the paper" or "see section 2 of the paper for definitions". Instead say something like "Variable definitions: y =log per capita income in 1992 dollars, r =interest rate on 3 month Treasury bills (end of quarter)," and so on. Alternatively, for many of you it might be best to include a table that defines the symbols, and in subsequent tables say "see Table x for variable definitions" where "x" is the number of the table that defines the symbols. (You will also present such information in the text itself.)

      ii. In tables that present regression results, include a note that describes the estimation technique ("The probit was estimated by maximum likelihood, assuming normality," for example.) (You will also present such information in the text itself.)

      iii. If a given set of variables appears in more than one table–as is often the case–there is no need to repeat the variable definitions. Instead one of the notes to (say) Table 2 can say "Variable definitions are given in notes to Table 1."

iv. Be sure to include the name of the dependent variable somewhere in the table.

v. When possible use words to describe the variables in your model. For example, if years of schooling is a regressor in your model write out "Years of schooling" not "$YRSCH$" if that is the name of the variable in your statistical software.

(e) In all but the simplest tables, number the rows and columns. When the text references a result in the table, cite the row and column: "the t-statistic is 2.12 (row (2), column (4))"

10. Figures:

(a) Number the figures, and on each include a descriptive header ("Parental Income versus SAT Score," for example).

(b) Figures may appear in the text in the appropriate place, or at the end of the paper.

(c) Figures should not run over page boundaries, and must always fit on a single page. That is, if you include a figure in the text, you should insure that you place it so that it does not run from one page to the next.

11. Reporting of estimates:

(a) Do not report more than 3 or 4 digits. Example: report 0.412, not 0.4117678.

(b) Avoid long strings of zeroes at the beginning of a number. You can always retroactively rescale variables and coefficients.

(c) Report standard errors, not t-statistics. Standard errors belong in parentheses under the coefficients. Example:

$$\text{Report} \quad \begin{matrix} 0.412 \\ (0.146) \end{matrix} \quad \text{not} \quad \begin{matrix} 0.412 \\ 0.146 \end{matrix}$$

12. Avoid the use of elaborate acronyms to denote variables (like AUSGDP_{90} for Australian GDP in 1990 dollars). They are rarely helpful to the reader. A single letter, usually with a subscript, ordinarily suffices and is easier to read when used in equations.

13. References:

(a) All references cited in the paper should be listed in a bibliography at the end of the paper. Cite these in the text as Walter (1995) or Walter (1995, p361).

(b) When you reference a specific result, such as a point estimate of a parameter, or a theorem that establishes a particular claim, give the page number, such as Walter (1995,p361). When you reference a general result, for example noting other papers that have studied topics similar to yours, no page number is needed.

14. Computer code: You do not need to include you programs in the paper. We should be able to figure out what you did without seeing it explicitly.

15. Miscellaneous reminders on terminology:

    (a) Hypotheses (not tests) may be "accepted" or "rejected."

    (b) Hypotheses refer to the magnitudes of population parameters, not estimates, and not to statistical significance. The word "significant" should not appear in the statement of a hypothesis.

16. It is a violation of scholarly ethics to repeat a passage, even a sentence, from another source without putting the passage in quotes and citing the source: the usual publication details in case of printed matter, the URL and date in the case of web-only material. This rule applies even when you are describing dreary facts: if you repeat a description from another paper of how data were collected, or the steps in computing an estimate, you must put the passage in quotation marks and cite the original source.