Final Project Guidelines
Big Data and Machine Learning for Applied Economics

The purpose of the final project is for you to showcase the skills that you acquired in this course. You can choose three routes for the project (a) a novel research paper, (b) replicate a paper that you find interesting, or (c) deploy a machine learning model in a web application that allows practical business decisions based on data. If you choose either (b) or (c) you should talk to me first.

There are two stages for the final project:

1. Project Proposal. This proposal is due on **October 11th at 6 pm**. The submission should be a brief statement of what you plan to do. I expect that you provide sufficient background so that the reader can understand the problem/research question you want to address. It should include an outline of how you plan to accomplish it, including the data you have/or plan to acquire, the methods, and any other relevant information you deem necessary. It should be at most 4 (four) pages in length. I'll provide feedback and will ask for changes if needed. **This is worth 5% of your final grade.**

2. Final submission. The final document, due on **December 12th at 6 pm**, should not be longer than 8 (eight) pages (not including the title page with abstract, references, and appendix), and contain the following sections:

   - Title

   - Abstract (200 words limit)

   - Introduction. It should contain at least: the problem/research question clearly defined, antecedents, the added value (i.e., why your project is interesting, novel, and different), and a preview of results and takeaways.

   - Data. Treat this section as an opportunity to present a compelling narrative to justify and defend the data choices. Describe it accordingly with descriptive stats, graphs, etc.

   - Model. Present and describe your preferred model specification. Argue why this is the best one for the task and describe it. Explain how it was trained, the hyper-parameters selection, and other relevant information. Be explicit about

how you define "best." If it is, by comparison, present the results of the different specifications and explain them briefly (details of non-preferred models can go into the appendix). In your comparison, be explicit about the metric chosen to compare among specifications.

- Results. Here you present your results. Interpret these as an economist. When reporting the results is always a good idea to follow Nikolov's (2022) advice:

  A good rule to follow regarding reporting results is that less is usually better. Novice researchers (or graduate students) tend to over-include or report many secondary parameter estimates from most regression specifications. Although such a "kitchen sink" approach has some merits (e.g., it shows the audience the extensive analysis performed or that the researcher has examined various aspects regarding the stability of the results), this particular approach has a significant drawback. Pages of parameter estimates usually muddy the main message and story of the paper. It can significantly detract from the comprehensiveness of your central contribution, which is most important in your paper.

    - *The reader will get either lost, bored, or annoyed.*
    - *Present only results and parameter estimates that directly bolster your main takeaway and story.*
    - *Do not present secondary analyses or relegate such analyses to an Online Appendix if needed.*

  I understand that a semester is a short time to have a full paper, so preliminary results are acceptable.

- Conclusions and recommendations. In this section, you state the main takeaways of your work.

- Data and Code availability. State where the reader can find the data and code. Results should be fully reproducibility. If you are using proprietary data, please get in touch with the professor.

This is worth 15% of your grade.