Applied Machine Learning – Project

Instructor: Prof. Yannis Bilias (bilias@illinois.edu), TA: Julian W. Oolman (jjpwade2@illinois.edu)

Final Project

An important goal of Applied Machine Learning class is to have students develop hands-on experience with the techniques discussed in the course. To this end, each group (of up to 4 students) is expected to write a paper to address a question of interest using the techniques that are covered in this class. We will greatly appreciate papers which analyze economic related questions or provide motivation based on economic reasoning.

(*Just a hint*: In **choosing a topic**, you may want to consider datasets you found in other empirically oriented subject areas (for instance, in applied econometrics or applied statistics), with rather large size and sufficiently large number of predictors. You can analyze such a data set using machine learning approaches. Or you can consider working on a topic and a paper you have been exposed to before to analyze it using a data set of new measurements (e.g., from a different country, or a different time-period).

In the *References* section at the end of this document, you can find a paper and a book that have been written by economists and use machine learning techniques. This material and the references listed therein might be helpful in developing ideas about your project paper.)

A substantial portion of your grade - 35% in total - will be based on your final paper. The term paper should be **returned in R Markdown format** and the upload should follow the same procedure you use when you return and upload the assignments. Given the experience of the assignments that are returned in R Markdown format, it will be very easy to exceed the 10 printable pages. As a result, you can exceed the 10 printable pages, but we **recommend** that you do not exceed the 20 printable pages. The paper is due on **April 28 on canvas. A research proposal is due on March 24.**

Logistics of Submitting your project

Each student should submit her/his **own** copy. However, you need to mention **all the members** of your group on the cover page.

Designing your project – Project Proposal (deadline March 24)

At the stage of designing your project, you should think what your main question will be, what data you will use, and why this is interesting.

The project proposal should be uploaded on canvas by March 24. The proposal should be no more than a page outlining the question you will study, a description of the dataset you will use (sample size, response, predictors) **with a link to it**, and whether it is a regression problem or classification problem.

Evaluation

This list is by no means exhaustive. A solid, coherent, creative paper will receive an A. Here are some tips on what is expected:

Project Question: It is important to describe the question you plan to research with clarity. This should be included in the abstract and in the introduction. It will be valuable to explain why this question is important and why people should care about this topic. In addition, you may want to explain why machine learning can be helpful in researching this question.

Review of Literature: What have other researchers said on the subject? You must include citations. Citations can be originated from popular news sources (such as *The Economist*, *The New York Times*, or *The Wall Street Journal*) and/or academic papers.

Data: Once you have a question, you must locate data to answer the question. Good resources for economic data are the U.S. Census (www.census.gov), the Bureau of Labor Statistics (www.bls.gov), and the Bureau of Economic Analysis (www.bea.gov), FRED Economic Data, IMF (www.imf.org), PENN World Tables, World Bank (www.worldbank.org) although there are many, many other data sources out there. You can also access lots of databases through the library system online. Another way to locate data is to google for "*data for machine learning*". In your paper, you should explicitly mention the source of the data and provide a link to the dataset or submit the dataset with your paper. Bear in mind that your results must be replicable.

Empirical Application: Much of this class is devoted to regression and classification with a primary view to obtain learning methods with improved prediction, for instance variables subset selection, regularized methods, tree-based approaches. You should be able to apply these techniques, and others, using the data set you have chosen. To receive a good grade, these concepts must be applied *correctly*. In addition, the steps of analysis to reach the final best candidate model (or models) should be clear and coherent.

Interpretation: You will **not** receive a good grade if you just present a bunch of numbers and leave it to the reader to come up with their own conclusions. You need to thoroughly explain the results to the reader with text well-written between the code-output chunks.

Computing

To write a paper using data, you will need to use a statistical software program. The default programming language of the course is R. People familiar with the other language (like python) are free to use it as long they provide explanation of the libraries and routines they use.

Part of the lecture time is devoted to presenting R Labs of the text. These are very well organized and very informative. They illustrate the essential R routines needed for your analysis. You don't need to be an experienced R programmer to make use of these routines.

If you have a question or problem with your R code, please make every effort to find the solution yourself. There are ample resources available, for free online, to familiarize yourselves with these programs in addition to the in-class labs. Working through the code diligently on your own is the only way to truly learn, and these skills will be invaluable on the job market.

References

Sendhil Mullainathan and Jann Spiess (2017), "*Machine Learning: An Applied Econometric Approach*", Journal of Economic Perspectives, pp87-106. (Available https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87)

Matt Taddy (2019), "*Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*", McGraw Hill.