# E401/M518: Machine Learning for Economic Data

Fall 2023


| | | | |
|---|---|---|---|
| **Instructor:** | Stefan Weiergraeber | **Lecture time:** | TR: 3:00pm–4:15pm |
| **Email:** | sweiergr@iu.edu | **Room:** | Wylie Hall 125 |


**Course Pages and Office Hours**

- Announcements and course materials will be posted on Canvas.
- Office hours: Thursday, 10:30am–noon in WY 347 or by appointment.

**Objectives:** This course is an introduction to modern tools from the field of statistical and machine learning. After reviewing basic concepts, such as the bias-variance trade-off, linear regression, and cross-validation, we will cover a broad-range of machine learning methods, for example, shrinkage estimators (ridge regression and LASSO), splines, and random forests. if time permits, we will also cover state-of-the-art methods that apply machine learning techniques to causal inference, for example, double/debiased machine learning, causal forests, and matrix completion methods. Throughout the course, we will use the software package R and economic data to illustrate the discussed concepts and methods. Empirical projects with real-world data and student presentations will be an integral part of the class. We will use the projects to also discuss the full workflow of data science from getting data, importing and cleaning data, visualizing data, to communicating the results of empirical analyses.

This class is cross-listed as E401: Machine Learning for Economic Data and M518: Big Data in Economics. The former is an advanced undergraduate class for Economics majors. The latter is an elective course for M.Sc. students in Economics and Data Science.

**Prerequisites:** Before enrolling in the class you are required to have taken E370 (or equivalent) and E371 (or equivalent). I expect you to have any prior coding experience, although such experience is not strictly required.

**Main References:** *An Introduction to Statistical Learning with Applications in R* (henceforth ISLR) will be our main reference for the theoretical part on statistical learning methods. For the purpose of this course you can either get the first or the second edition. If time permits, we will spend several lectures on recent econometric papers that discuss the application of machine learning methods for causal inference. I recommend *R for Data Science* (henceforth R4DS) as the main reference if you would like to learn more about data science and coding in R. In addition, I will provide a full set of videos on Canvas, in which I introduce you to the fundamentals of R.

- Hadley Wickham & Garrett Grolemund, *R for Data Science*, O'Reilly, 2016.
- James, G., Witten, D., Hastie & T., Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2022.


page 1 of 7

**Learning Outcomes:** Upon completion of this course, you should be able to ...

- have the foundation to become a professional data scientist. In a nutshell this means that you are able to turn raw data (essentially a long sequence of numbers) into meaningful insights about economic processes and relationships.
- use R to master the full practical workflow of data management, in particular, data read-in, cleaning, transformation, visualization, and exploratory data analysis.
- apply a wide range of statistical/machine learning methods to analyze economic data (broadly defined) using econometric models. In particular, you should be able to discuss why a certain method works, when you should apply it and how you should interpret the results. Moreover, you should be able to relate machine learning methods to classical econometrics and assess the advantages and disadvantages of using one or the other for a specific application.
- communicate the results of empirical analyses to non-specialists.

**Software:** For illustrating the methods and the empirical exercises you will need to have R, RStudio, and several R libraries installed on your computer. All of these are free and open source. If necessary, we will have several tutorial sessions to introduce you to the fundamentals of the software.

**Overview of Different Learning Activities:** In this class we will use a variety of learning activities and assignments.

1. In-person lectures: We will meet on Tuesdays and Thursdays, 3:00pm to 4:15pm in WY 125. I will not formally grade attendance, but I expect you to come to class and participate. Many of the concepts that we discuss in this class are fairly abstract and will usually take you several iterations to wrap your head around; therefore, I believe that our in-person meetings and your active participation by asking questions are the most important component of the class. I understand that the current circumstances can make in-person attendance difficult at times for some of you. If you have any concerns in this regard, please let me know. There will be two types of lectures. For the first type I will lecture in a more classical style on several methods from the field of statistical learning. The second type of meetings will be devoted to student presentations of empirical challenges.

   The empirical challenges are 30 to 40 minutes presentations of an empirical case study by pairs of students. The structure of each challenge will be very similar: I provide you with a real-world data set and a relatively open assignment, similar to what your boss might give you when you start a job as data scientist in a company or a policy institution. Broadly speaking each challenge will be about extracting information from raw data using one or two techniques that we recently discussed in class. If you have never conducted a serious data analysis before the challenges can seem like a daunting task, that can take you quite some time to prepare. Since the challenges are very similar to what you could encounter in your future career, I think it will be worth the time investment and I hope that you will find them one of the most useful learning activities of the course.


page 2 of 7

Generally, I will not pay a lot of attention to your presentation style, i.e., do not invest in fancy Powerpoint slides and a polished speech. Going through your R script and discussing the various code chunks, tables, and figures to make your argument, will be totally fine. I expect you to be very clear and show a good understanding of the relevant methods, however. If you are presenting you should be the experts in the room on the relevant topic.

You should think of me and your fellow students as participants in a board or division meeting that wants to use your analysis to improve the decision making of the company or the policy institution. Depending on how many students take the class I might also assign pairs of students to act as boss/client/policy maker whose role is to skim the presented analysis in advance and prepare a short feedback discussion about parts that you liked, points you would criticize etc.

I also expect the non-presenters to ask questions and I generally expect the presenters to be able to answer them. The question can refer to both the coding techniques and the empirical methodology used as well as the specific application. So we will use the challenges also as an opportunity to review concepts and answer remaining questions about the course content that is covered in the pre-recorded videos. It is unlikely that the presenters will have a perfect answer all the time (and probably I won't have either). In this case we will try to figure out the answer together. I would like each student to present at least one challenge.

My hope is that by presenting an actual case study you will also gather a little bit of presentation experience, which is a very valuable skill on the job market. I will upload all challenges by the end of the first week of the semester and would like you (in groups of three students) to sign up within the first three weeks of the semester for at least one presentation.

2. Pre-recorded lectures: If you do not have a lot of experience with data work in R, I will upload recordings of the core lecture material with roughly 90 minutes lectures per module. I recommend that you watch these videos on the timeline indicated in the module structure.

   Each lecture will be accompanied with a short discussion thread in which I ask you to briefly post some insights and questions that came to your mind while watching the lecture. This is completely optional, and only for a little bit of extra credit. It is totally fine to back another students opinion or question in these threads. However, I ask that you try your best to come up with an insight or question that has not been brought up in the specific discussion yet.

   In addition, there will be short multiple-choice quizzes for each lecture and recording. The quizzes will count as extra credit towards your final grade (for more details on the weighting of the different assignments, see below). I will not formally grade the discussion participation. However, I will monitor them and use them as the basis for our in-person lectures as well as a tie-breaker in case you are in between grades at the end of the course.

   There will be two types of recordings. First, lectures about managing data and coding in R. These will be very practical lectures in which I demonstrate various tasks in R. I strongly encourage you to implement that examples presented on your own computer and internalize these concepts by doing. You may have to pause the


page 3 of 7

video a couple of times and take your time to figure out what exactly is going in a certain code chunk to replicate it and make it work on your computer. Second, I will upload tutorials on how to implement the statistical learning methods that we discuss in class in R. The material from the videos will cover the skills you need to work on the empirical challenges.

My main reason for outsourcing the coding topics to recorded lectures is that based on my experience in previous semesters it is very hard to teach coding in an actual classroom to a group with a heterogeneous background. Inevitably, I would go too fast for those of you who have never done any coding. At the same time, those of you who have coding experience would easily be bored. The recordings will allow each of you to engage with the basic material at your own pace. This will free up some valuable class time for talking about the application of the coding techniques in real-world applications, i.e., the empirical challenges and hopefully more interesting and focused Q&A sessions.

I realize that this is an increase in workload compared to covering everything in the lectures exclusively. I believe that this totally manageable. I would compare watching the recordings to doing some required advance reading that you would have to do for other classes. I will post lecture recordings within the first three weeks of the semester to give you maximum flexibility.

3. Problem sets: There will be approximately 8 problem sets throughout the semester. Please work on and hand in each problem set in groups of at most three students. I will upload answer keys for each problem set, and I will ask you to self-grade your answers. Please note that I reserve the right to change your self-assigned grade.

4. Mid-term: There will be one mid-term, which cover the material covered in class, mostly in the form of short-answer or essay questions. The exams will be in-class and closed-book. I tentatively scheduled the midterms for September 28.

5. Method presentation in video form: Instead of a second midterm, I will ask you to prepare another group project. This project will be similar to the empirical challenge presentation in class. However, the topics will be more advanced and not covered in the lecture material. In addition, I ask you to record your presentation on video and upload it to Canvas.

   I will provide you with a specific list of project assignments by the first week of the semester, and I am open to your suggestions. My default is to have you explain a method from the field of causal inference and apply it it to a real-world data set, which I will provide. For example, you can work on matching, RDD, IV, panel data, diff-in-diff, synthetic control, text analysis or factor models.

   We will not have a class on November 9 to give you time to record your group video. The tentative deadline for uploading the video to Canvas is November 15. After each group has uploaded their videos, each group is asked to watch the video presentations of at least two other groups and write a brief evaluation of the presentation. The evaluation reports will be due December 1.

6. Final project: At the end of the semester you will have to hand in an empirical project in which you gather some data, run some analysis and present the results


page 4 of 7

in a short report. You can work on this project alone or work with a partner. You are free to choose the topic of your project as long as it relates to some of the topics discussed in class. The project cannot exceed the limit of 15 pages. I strongly encourage you to get started on your project very early in the semester as it can take time and potentially multiple attempts to find a topic that really works for you. Several of the problem sets will contain more specific guidelines and intermediate steps that I would like you to follow when you develop your project. You are required to hand in a two-page project proposal by October 15. The final project will be due on December 15, 2023.

**Grading Policy:** Course Assessment: Problem sets (15%), midterm exam (20%), challenge presentation (20%), method presentation in video from (20%), final project (25%). For extra credit: post-lectures discussions and quizzes (up to 20%). For details on the different types of learning activities and assessments, please see above.

Submitting Assignments: Please submit all your assignments via Canvas in one pdf file. Even if you are working on the problem sets in a group, I ask that each group member uploads their own solution. Please indicate clearly the member of your group on the assignment. Please do not submit any answers via email.

How Will I Know How I'm Doing in This Course? I will use the Canvas Gradebook to keep a record of your scores. Assignment grades (in the form of points) will usually be posted in Canvas within 14 days of the due date. I will refrain from committing to a fixed final grading scale at this point. If you have any concerns or need assistance in this regard, please do not hesitate to contact me.

Late Assignments: In order to be fair to your classmates who hand in assignments on time, I will generally not be able to accept late assignment hand ins. I understand that these are still challenging times for everybody and that technology may fail occasionally. I will provide you with the relevant materials at least three to four weeks in advance. This comes with the expectation that you factor in potential technology fails or other disruptions when planning your course work. Starting an assignment the night before the due date is not a viable strategy for this class!

**Tentative Course Outline:** I will organize all course material and topics into weekly Canvas modules. Depending on how the course evolves I may drop or add some topics. So please check the module page regularly for the most up to date information. If you have requests for specific topics that are currently not listed in any module, please do not hesitate to contact me and I will try my best to accommodate your requests.


page 5 of 7

Week 1: Introduction to the course and R .....
Week 2: Review of linear regression .....
Week 3: Linear regression wrap up and challenge presentation .....
Week 4: Basic Concepts in Statistical Learning .....
Week 5: Cross-validation .....
Week 6: Exploratory Data Analysis & Midterm 1 .....
Week 7: Linear Model Selection & Regularization .....
Week 8: Linear Model Selection & Regularization .....
Week 9: Nonlinear models .....
Week 10: Nonlinear models (part 2) .....
Week 11: Tree-based methods .....
Week 12: Tree-based methods .....
Week 13: Catch up week and prepare video recordings .....
Week 14: Machine learning for causal inference .....
Week 15: Recent topics in AI & course wrap-up .....

**Important Dates**

Midterm: In-class closed-book exam on **September 28, 2023**.
Project Proposal: Two page proposal due on **October 15, 2023**.
Group Presentation Video: upload due on **November 15, 2023**.
Final Project: The project will be due on **December 15, 2023**.

**Additional References**


- Athey, S. (2017): *Beyond Prediction: Using Big Data for Policy Problems*, Science.
- Athey, S., and Imbens, G. (2019): *Machine Learning Methods that Eonomists Should Know About.*, Annual Review of Economics, 685-725.
- Einav, L., and Levin, J. (2014): *Economics in the Age of Big Data*, Science.
- Hastie, Trevor, et al. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Springer, 2009.
- Hastie, T., Tibshirani, R., & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, 2015.
- Mullainathan, S. and J. Spiess (2017): "Machine learning: an applied econometric approach," Journal of Economic Perspectives, 31, 87-106.
- Varian, H.R. (2014): "Big Data: New Tricks for Econometrics", Journal of Economic Perspectives, 28(2), 3-28.


**Policies**


page 6 of 7

- **Academic Integrity:** As a student at IU, you are expected to adhere to the standards and policies detailed in the Code of Student Rights, Responsibilities, and Conduct. When you submit an assignment with your name on it, you are signifying that the work contained therein is yours, unless otherwise cited or referenced. Any ideas or materials taken from another source for either written or oral use must be fully acknowledged. If you are unsure about the expectations for completing an assignment or taking a test or exam, be sure to seek clarification beforehand. All suspected violations of the Code will be handled according to University policies. Sanctions for academic misconduct may include a failing grade on the assignment, reduction in your final course grade, a failing grade in the course, among other possibilities, and must include a report to the Dean of Students, who may impose additional disciplinary sanctions.

- **Special circumstances:** Students requiring any type of special classroom/testing accommodation for a disability, religious belief, scheduling conflict, or other impairment that might affect his or her successful completion of this course must personally present the requested remedy or other adjustment in written form (signed and dated) to the instructor, i.e. supporting memorandum of accommodation from the Office of Disabilities Services for Students. Requests for accommodations must be received and authorized by the instructor in written form no less than two weeks in advance of need. No accommodation should be assumed unless so authorized. In the event of needs identified later in the course, or for which an adjustment cannot be made on a timely basis, a grade of "I", Incomplete, for the course will be given to accommodate the unanticipated request.

- **Exam absences:** In the event of a catastrophic (and documented) occurrence which necessitates an absence from a scheduled exam, the student should immediately seek the instructor's *permission to miss an exam.* If approval is granted, the weights of the student's scores for the other exams will be re-adjusted proportionately, so as to make up for the missed exam. If completed documentation is not presented within one week after a missed exam, or if no *permission to miss a exam* has been obtained prior to the exam date, the missed exam will received a score of zero points.


page 7 of 7