Ruhr Graduate School in Economics

Introduction to Data Science and Predictive Analytics for Economists

Fall 2023

In person Course

Instructor: Steven Lehrer **E-mail:** lehrers@queensu.ca
Office hour Thursday: 16:00-17:00 and Friday 14:00-15:00

Course Description: This course aims to provide an understanding of the fundamental concepts and frameworks in the interdisciplinary field of data science and data analytics. The primary goal of this course is for students to learn data analysis concepts and techniques that facilitate making decisions from a rich (and potentially large) data set as well as practical computational skills. The course is also designed to act as a primer for continued study. It is not specifically an introduction to computer science or machine learning or data mining course, nor a class on high-dimensional econometrics and statistics; rather, like a good data scientist, the class borrows from multiple disciplines. Techniques covered include an advanced overview of linear and logistic regression, model choice and false discovery rates, information criteria and cross validation, regularized regression and the lasso, bagging and the bootstrap, causal estimation and treatment effect heterogeneity, binary regression, classification, latent variable models, principal component analysis, topic models, decision trees and random forests, neural networks, deep learning, boosting, support vectors, text analysis and natural language processing. Many of these methods have the potential to dramatically improve the public welfare by guiding policy decisions and interventions, and their incorporation into intelligent information systems could improve public services in domains ranging from medicine and public health to law enforcement and security.

Throughout the course, heavy emphasis is placed on analysis of actual datasets and the course will provide an opportunity to utilize an open source data analysis tool, R, for data manipulation, analysis, and visualization. Finally, in this course we have the option to discuss diverse issues around data including tradeoffs in ICTC policies including the GDPR.

Course Prerequisites: Students are expected to be comfortable with econometrics, statistics, and microeconomics. Students are expected to be able to undertake data analysis with R.

That said, any course that seeks to provide knowledge in data science, will contain mathematics and statistics and require the ability for algorithmic thinking. Individuals who are unfamiliar with math notation, symbols and basic algebra rules should not take this course. To be clear, even though this course is not designed to transform you into a data scientist, your success will be tied to your foundation and ability to develop from your foundation in programming and both mathematical and statistical concepts/theory.

If you have any questions about these requirements, please see me.

Website: Additional information, lecture outlines, links to readings, the calendar/reading list, and other useful information about this class can be found on the course website

Objectives: After completing this course, you will be able to:

- Understand the distinction between supervised and unsupervised learning
- Improve critical thinking skills
- Be able to translate data science jargon and provide the intuition of how a variety of algorithms work. Explain the trade-offs between alterative algorithms and the role of hyperparameters within each
- Improve programming skills
- Be able to assess whether the goal of an exercise is to yield a prediction or undertake causal inference with machine learning methods.
- Be able to explain why results from econometric strategies differ from machine learning algorithms.
- Be able to explain what cross-validation is and why it is crucial.
- Gain familiarity with causal machine learning methods
- Realize that data science has many practical applications within economic research.

Required Readings: A reading list should appear on the course calendar, but until I get a sense of the class, I will not finalize the exact readings. A copy of all the readings listed on the calendar in this syllabus will be placed on reserve at the library. You may be assigned additional readings dealing with topics covered in class**. Specifically assigned** supplementary readings are integral parts of the course, and therefore, exam questions dealing with those readings are **highly likely**.

In general, readings are assigned for each lecture period, and the material in these readings will be discussed in class. Please complete all readings prior to the class in which they will be discussed. The lectures will also cover material not included in the readings. To a large degree, the readings and lectures are not substitutes – they are chosen and designed to complement each other.

A solid primer to much of the material we will cover is An Introduction to Statistical Learning, by James, Witten, Hastie, and Tibshrani. However, it takes a very different approach from us and only partially overlaps on material. We will also rely on the appendix to my recent paper with Tian Xie that was published in Management Science titled "The Bigger Picture: Combining Econometrics with Analytics Improves Forecasts of Movie Success", since it provides a clear intuitive explanation of many methods we will cover. This material can be found at https://pubsonline.informs.org/doi/suppl/10.1287/mnsc.2020.3911.

Supplementary Readings: You may be assigned additional readings in the form of articles dealing with topics covered in class**.** Specifically assigned supplementary readings are integral parts of the course.

Software: All computing is conducted in R, a platform for statistical analysis. R, which is available for free via www.r-project.org. You can download and install the software following directions at cran.us.r-project.org (do this ASAP). R is a widely used and hugely flexible analysis platform. It has a command line interface (you type commands to get what you want). Some students find the learning curve for such 'programming' to be very steep. I provide limited software instruction, in-class demonstration, and code to accompany lectures and assignments. I assume that you have not used R in a previous class. However, this is not a class on R. Like any language, R is only learned by doing. You should install R as soon as possible and familiarize yourself with basic operations. Ideally, you would start this course able to replicate any analysis from previous classes in R. A great way to start learning is to buy a book and start working through tutorials. A good guide is Adler's R in a Nutshell. They have many tutorials to help you get up to speed. You can browse other options by searching 'R statistics' on Amazon. To make it possible to focus on data science concepts, I strongly encourage students to learn the basics of the language and software BEFORE starting the course. See the next section for detail and resources.

R studio is a free platform for both writing and running R, available at www.rstudio.org. Some students find it friendlier than basic R (especially in windows OS)

Additional R resources

- Tutorials at data.princeton.edu/R are fantastic (and there are many others out there).
- youtube intros to R
- Me and your classmates: work together, and chat on the discussion board.

Grading: This course is designed to be very rigorous and demanding. You are expected to work hard, actively participate in class, ask questions when you have any doubts, and perform to the very best of your ability. The course grade will be computed using the following weights

Class Participation/Attendance	15%
Assignments	85%
Total	100%

On class participation: Class discussion is important for both individual and collective learning. The quality of a student's participation is at least as important as the quantity, and the following points characterize effective participation: . Do comments draw on the text and materials from this and other courses? Do they show evidence of analysis? Does the student distinguish between positive and normative analysis? Does the student distinguish between opinion and well-supported analysis? . Are the points made substantive? Are they linked to the comments of others? Do they advance or deepen the discussion? Do they deepen the analysis? Do comments clarify and highlight the important aspects of earlier comments and lead to a clearer statement of the concepts being considered? Is there an attempt to synthesize the discussion?

On assignments: Assignments will be created and graded to ensure that each student can better assess their progress in the course. In order to maximize the benefits of this course, there will be many short programming assignments. I will grade these and try to provide you with direct feedback on your coding. Note, if we cover text/image analysis, that assignment will involve Python.

If you need help: If you find that you are having difficulty with any of the material in this course:

(1) DO NOT let it build up. The material is very cumulative in nature and you are likely to find yourself only falling further behind.

(2) DO contact me, either after class or by making an appointment. Be forewarned: I expect that you have read the appropriate sections of the textbook and reviewed your notes BEFORE any discussion.

Attendance and Lateness: All students are expected to attend class regularly. Although attendance will not be taken each and every class, be warned that you are responsible for all material covered in class including that, which is not in the text. You are expected to make every effort to be on time to class.

Teaching Style: There are many ways to learn. And different styles are more effective for some students than others. Therefore, we will utilize several different approaches: straight lectures, Powerpoint slides, problem sets, and exam preparation.

Accommodation after the fact: Once a student has submitted an assignment, they may not subsequently be granted accommodation such as being offered a second opportunity to write assignment or have it count for less than originally specified in the course syllabus (reweighted). Students who cannot perform to the best of their abilities due a serious, extenuating circumstance must inform their instructor before attempting an assignment to arrange appropriate accommodation.

References: In general, I am happy to provide references for employers or write letters of reference for students who plan to attend graduate school. The strength of my recommendation remains positively correlated with your performance in my course. For job references, please email me with a heads up that a potential employer might call or email. Please also let me know if there are any skills of yours that I should highlight in my reply to them. Naturally, make sure that these claims are credible as my reputation is on the line. If you would like a letter of reference for graduate school, please provide me with a short note explaining what the reference is for and when it is due. Also attach a statement of purpose (if relevant) as well as a current CV/transcript. Please allow 3 weeks for the completion of letters.

Tentative Calendar and Reading List

Date	Topic	Assigned Readings to be Announced Shortly
Wednesday 9:30-12:30	Course Introduction Regression Review and Binary Responses Model Selection and Model Averaging	
Wednesday 14:00-16:00	Penalized Regression and Variable Selection Methods including Lasso, Ridge and Elastic-Net	
Thursday 9:30-12:30	Regression Trees, Bagging, Forests and Boosting. Hybrid Approaches Introduction to Neural Networks	
Thursday 14:00-16:00	Nonparametric Approaches including Support Vector Machines, Neural Networks and Deep Learning	
Friday 9:30-12:30	Causal Machine Learning Unsupervised Learning: Clustering Methods Text Analysis and Topics on the Research Frontier	