

# Prediction and machine learning






## [ EC524/424 ]

Spring 2024 Syllabus









<https://github.com/edrubin/EC524S24>

**Dr. Edward Rubin**

Dept. of Economics, University of Oregon

	<u>Instructor</u>	<u>GE</u>
	<b>Edward Rubin</b>	<b>Ngan Tran</b>
	<a href="mailto:edwardr@uoregon.edu">edwardr@uoregon.edu</a>	<a href="mailto:ntran7@uoregon.edu">ntran7@uoregon.edu</a>
	Start subject with "EC524:".	
	PLC 530	PLC 428
	?	2p-3p
	<a href="https://edrub.in">https://edrub.in</a>	

**Email note:** We will do our best to respond promptly to your emails. Our responses may be slower over weekends/holidays. There may be times that our responses take up to 48 hours. Please do not repeatedly send the same email unless it has been more than 48 hours.

	<u>Lecture</u>	<u>Lab</u>
	Mo. & We., 12:00p-1:20p	Fr., 12:00p-12:50p
	<a href="#">195 Anstett</a>	<a href="#">360 Condon</a>
	Ed	Ngan   Ed
	Our class: <a href="https://github.com/edrubin/EC524S24/">https://github.com/edrubin/EC524S24/</a>	
	Last class: <a href="https://github.com/edrubin/EC524W23/">https://github.com/edrubin/EC524W23/</a>	
	The previous class: <a href="https://github.com/edrubin/EC524W22/">https://github.com/edrubin/EC524W22/</a>	
	The previous <sup>2</sup> class: <a href="https://github.com/edrubin/EC524W21/">https://github.com/edrubin/EC524W21/</a>	
	The previous <sup>3</sup> class: <a href="https://github.com/edrubin/EC524W20/">https://github.com/edrubin/EC524W20/</a>	

## Course summary

**Description** Following the first course on econometrics and causal inference in our sequence, EC524 turns to examining the **tools available and best practices for predicting outcomes**. Put simply, we are now focusing on  $\hat{y}_i$  rather than  $\hat{\beta}$  from the model  $y_i = \alpha + \beta x_i + \varepsilon_i$ .

Learning statistical programming is inherent to practicing applied econometrics. Consequently, throughout this course we will also teach the statistical programming language R.

### Objectives

1. **Distinguish** between settings that require **causal inference** vs. settings that want **prediction**.
2. Understand the main **themes and best practices** in modern **prediction** methods.
3. Develop **familiarity** with common machine-learning algorithms—and their strengths/weaknesses.
4. Build **intuition** for prediction—especially the bias-variance tradeoff.
5. Expand **R expertise**.

**Prerequisites** This course requires the previous course in our sequence—*i.e.*, Economics 423/523. I also assume you are comfortable in R.

## Books

I know you are busy and reading for class is often difficult. However, **if you are actually here to learn, then read these books**.

*Note* Each book (except two of the recommended books) is available for **free online**. The physical copies are also very reasonably priced—I suggest you buy physical versions for books that you like.

### Required books

1. **Introduction to Statistical Learning** *ISL*
2. **The Hundred-Page Machine Learning Book** *100ML*
3. **Data Visualization** *Data Viz*

### Suggested books

1. **R for Data Science** *RDS*
2. **Introduction to Data Science** *IDS* (not available without purchase)
3. **The Elements of Statistical Learning** *ESL* (the big brother of *ISL*)
4. **Data Science for Public Policy** *DSPP* (e-book available via UO library)

## Software and tools

- We will use the statistical programming language **R**.
- We will use **RStudio** to interact with R.

Learning R will require time and effort, but it is a powerful and versatile tool that is valued by many employers. Put in the requisite effort and time, and you will be rewarded.

## Labs, assignments, projects, and exams

**Attend the lab** This course includes a lab, which is **integral to learning** the material in (and passing) this course. The lab includes both general econometrics instruction and computing resources necessary to complete the course and learn/master its topics.

### Assignments

- You will submit **typed assignments via Canvas**, generally in one of two formats (we'll tell you what we want):
  1. An **R notebook** that is hosted somewhere on the web
  2. A link to a **Kaggle notebook**
- We will grade on a **complete/incomplete scale**.  
Low-quality work will be returned to be re-submitted as late.

**Late submissions** Students whose assignments are occasionally late will be penalized half a letter grade. Students whose assignments are frequently late will be penalized a full letter grade.

**Group work** Feel free to work together on the assignments. Unless explicitly stated, each student is required to write and submit independent answer sheets. This means that word-for-word copies will not be accepted and will be viewed as academic dishonesty. If you work with other students, you must list the students in your study group at the top of your assignment. If you fail to do so, you will receive a score of zero.

**Project** We will have one major project. You will choose a topic that we have not covered in class—but that is related to topics covered in class—e.g., spectral clustering or time-series prediction. You will then:

- Learn about this topic on your own,
- Write a 'wiki' that explains this topic using math and examples,
- Make and give a five-minute presentation on the topic (with a simple example)

No duplicates for topics. I will provide a list of ideas on the course site. The idea here is that you extend/apply the course's ideas to situations that **interest you**.

### Exams

- We will proctor an **in-person final** on Friday, June 14<sup>th</sup>, 2024 from 10:15am–12:15pm Pacific.
- A **take-home final** exam will be due Friday, June 14<sup>th</sup>, 2024 by 10:15am.

## Recommendations

1. **Be kind.**
2. **Take responsibility** for your own education and try to **learn** as much as you can.
3. **Do your own work.**
4. Develop your **intuition**—e.g., why would method  $x$  work in one situation and fail in another?
5. **Learn R.** Struggle while you try—and use **Google** to figure things out.
6. Come to **office hours**.<sup>1</sup>

## Honesty and academic integrity

**You must do your own work.** Do not claim credit for any work other than your own. **Your work should not be identical to others' work.** Cheating or plagiarizing of any sort on any component of this class will result in a failing grade for the term and a report of the offense to the university. Please acquaint yourself with the [Student Conduct Code](#).

**Large language models:** ChatGPT, GitHub Copilot, and the related AI 'assistants' are great tools. I am totally fine with you using them—and even encourage it. However, you still need to submit work **in your own words**, and you need to **understand the code** that you submit. Anything less is plagiarism, lazy, and a loss of opportunity to actually learn valuable material/tools.

## Accessibility

If you have a documented disability and anticipate needing accommodations in this course, please make arrangements with me during the first week of the term. Please request that the [Accessible Education Center](#) send me a letter verifying your disability.

## Grading

Grades will be assigned as follows.<sup>2</sup>

<u>Grade</u>	<u>Assignments</u>	<u>Project</u>	<u>Final exam</u>
<b>A</b>	<i>Incomplete</i> on 0 assignments.	$\geq$ <i>Professional</i>	$\geq$ 80%
<b>B</b>	<i>Incomplete</i> on $\leq 1$ assignments.	$\geq$ <i>Minor revision</i>	$\geq$ 70%
<b>C</b>	<i>Incomplete</i> on $\leq 2$ assignments.	$\geq$ <i>Moderate revision</i>	$\geq$ 60%
<b>D</b>	<i>Incomplete</i> on $\leq 3$ assignments.	$\geq$ <i>Major revision</i>	$\geq$ 50%

Recall that assignments are graded as *Complete* vs. *Incomplete*—the standard for *Complete* is much higher than simply submitting.

---

<sup>1</sup>Two related articles from NPR on office hours: [College Students: How to Make Office Hours Less Scary](#) and [Uncovering A Huge Mystery Of College: Office Hours](#).

<sup>2</sup>Undergraduates are allowed to miss one additional assignment in the scheme.

## **COVID-19 and safety**

The University of Oregon (UO), in accordance with guidance from the Centers for Disease Control, Oregon Health Authority, and Lane County Public Health requires faculty, staff, students, visitors, and vendors across all UO locations to use face coverings when in UO owned, leased, or controlled buildings. This includes classrooms. Please correctly wear a suitable face covering during class. Students unable to wear face coverings can work with the Accessible Education Center to find a reasonable accommodation. Students refusing to wear a face covering will be asked to leave the class.

If the professor or GE is made to feel threatened or uncomfortable by a student aggressively or repeatedly refusing to properly wear a mask, the student will be reported to the university and asked to withdraw from the class (or the student will receive an F).

## **Academic disruption**

In the event of a campus emergency that disrupts academic activities, course requirements, deadlines, and grading percentages are subject to change. Information about changes in this course will be communicated as soon as possible by email, and on Canvas. If we are not able to meet face-to-face, students should immediately log onto Canvas and read any announcements and/or access alternative assignments. Students are also expected to continue coursework as outlined in this syllabus or other instructions on Canvas.

In the event that the instructor of this course has to quarantine, this course may be taught online during that time.

# Tentative, overly-ambitious, predicted outline

Note: Stay up to date on our class [class's Github page](#).

## 0. An introduction to prediction and statistical learning

1. What are we doing? **Readings** *ISL* Introduction, Ch1
2. Prediction vs. causal inference **Readings** *Prediction Policy Problems* by Kleinberg *et al.* (2015)
3. Modeling decisions and assessment **Readings** *ISL* Ch3

## 1. Exploratory data analysis

1. Building insights from graphics **Readings** *Data Viz* Preface, Ch1
2. `ggplot2` **Readings** *Data Viz* Ch3

## 2. Supervised learning

1. An introduction to machine learning **Readings** *100ML* Preface, Ch1–Ch4; *ISL* 2.1–2.2
2. Resampling methods and other best practices **Readings** *100ML* Ch5; *ISL* Ch5
3. Why don't we stick with regression? **Readings** *ISL* Ch3
4. LASSO and Ridge regression **Readings** *ISL* 6.1–6.3, 6.6
5. Classification and logistic regression **Readings** *ISL* 4.1–4.3
6. Decision trees **Readings** *100ML* 3.3; *ISL* 8.1
7. Ensembles: Bagging, random forests, boosting **Readings** *ISL* 8.2–8.3 *100ML* 7.5 and Ch8
8. SVM **Readings** *100ML* 3.4; *ISL* 9.1–9.4
9. Neural nets **Readings** *100ML* 6
10. Additional topics **Readings** *100ML* Ch7 and Ch11

## 3. Unsupervised learning

1. Introduction to unsupervised learning **Readings** *100ML* Ch9; *ISL* 10.1
2. Principal components analysis **Readings** *ISL* 10.2; *100ML* 9.3
3. Nearest-neighbor matching, *K*-means, and hierarchical clustering **Readings** *100ML* Ch9; *ISL* 10.3

## 4. Extensions

1. Bias and fairness **Readings** [Hao \(2019\)](#)