Syllabus

INFO 4940/5940 - Applied Machine Learning: Methods and Applications

Instructor

- Dr. Benjamin Soltoff
- Office: Gates Hall 216
- Email: soltoffbc@cornell.edu
- Office hours: Wednesdays 12-2pm, or by appointment

Course logistics

- Meets TuTh from 1:25pm - 2:40pm for 28 sessions
- 3 credits, offered for a letter grade or S/U
- Prerequisites: INFO 2950, INFO 5001, or equivalent data science experience. Must have experience using R and Git.

Course description

This course provides an introduction to contemporary machine learning methods and their applications. It covers the entire machine learning pipeline, from data collection and wrangling to model evaluation and deployment. The course emphasizes practical applications of machine learning, with additional weight on reproducibility and effective communication of results. We will develop and deploy models using R, the open-source programming language. Our focus will generally be on problems that are solved using tabular data structures and "shallow" machine learning techniques. We will examine deep learning and large language models (LLMs), but predominantly from the context of how to meaningfully and responsibly incorporate them into data science workflows.

Course learning objectives

By the end of the semester, you will...

- Train and evaluate machine learning models using a variety of algorithms.
- Collect and wrangle data for machine learning.
- Deploy machine learning models in a production environment.
- Communicate results of machine learning analyses to a non-technical audience.
- Implement reproducible machine learning workflows using version control and literate programming.

Office hours

- Saif M. - Mondays 10-12pm (Rhodes Hall 404)

- Dr Soltoff - Wednesdays 12-2pm, or by appointment (216 Gates Hall)
- Yuhan T. - Wednesdays 5:30-7:30pm (Rhodes Hall 406)
- Sam G. - Fridays 12:30-2:30pm (Rhodes Hall 404)

Textbooks

All books are **freely available online** or can be accessed electronically through the Cornell library.

- Tidy Modeling with R by Max Kuhn and Julia Silge.
- Introduction to Statistical Learning with R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
- Deep Learning with R, Second Edition by Francois Chollet, Sigrid Keydana, Tomasz Kalinowski, and J.J. Allaire.
- Supervised Machine Learning for Text Analysis in R by Emil Hvitfeldt and Julia Silge.

Course community

We want you to feel like you belong in this class and are respected. Cornell University (as an institution) and we (as human beings and instructors of this course) are committed to full inclusion in education for all persons. If for any reason you feel that we have failed these goals, please either let us know or report it, and we will address the issue.

Services and reasonable accommodations are available to persons with temporary and permanent disabilities, to students with DACA or undocumented status, to students facing mental health or other personal challenges, and to students with other kinds of learning challenges. Please feel free to let me know if there are circumstances affecting your ability to participate in class. Some resources that might be of use include:

- Office of Student Disability Services: https://sds.cornell.edu
- Cornell Health CAPS (Counseling & Psychological Services): https://health.cornell.edu/services/counseling-psychiatry
- Undocumented/DACA Student support: https://dos.cornell.edu/undocumented-daca-support/undergraduate-admissions-financial-aid

Academic accommodations

We want all students to have the opportunity to be successful in this course. Accommodations can help provide some flexibility and equitable classroom access.

Per university policy, this course provides the following accommodations:

- Disability Accommodations
- Religious-Observance Accommodations
- Title IX Accommodations
- Varsity Athlete Accommodations
- Medical Accommodations
- Military Service

- Other Accommodations

Accessibility

If there is any portion of the course that is not accessible to you due to challenges with technology or the course format, please let me know so we can make appropriate accommodations.

Student Disability Services is available to ensure that students are able to engage with their courses and related assignments. Students should be in touch with Student Disability Services to request or update accommodations under these circumstances.

If you have an approved SDS accommodation, please send a copy of this letter to the instructors at soltoffbc@cornell.edu so we can ensure your accommodations are implemented in this course.

Communication

All lecture notes, assignment instructions, an up-to-date schedule, and other course materials may be found on the course website: info4940.infosci.cornell.edu.

Announcements will be posted through Canvas Announcements periodically. Please check Canvas (or ensure Canvas announcements are forwarded to your email) to ensure you have the latest announcements for the course.

Where to get help

- If you have a question during lecture or discussion, feel free to ask it! There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.
- The course staff is here to help you be successful in the course. You are encouraged to attend office hours to ask questions about the course content and assignments. Many questions are most effectively answered as you discuss them with others, so office hours are a valuable resource. Please use them!
- Outside of class and office hours, any general questions about course content or assignments should be posted on the course discussion forum. There is a chance another student has already asked a similar question, so please check the other posts on GitHub Discussions before adding a new question. If you know the answer to a question posted on the discussion board, I encourage you to respond!

Email

If there is a question that's not appropriate for the public forum, please email us at soltoffbc@cornell.edu. Barring extenuating circumstances, we will respond to INFO 4940/5940 emails within 48 hours Monday - Friday. Response time may be slower for emails sent Friday evening - Sunday.

Activities & Assessment

The activities and assessments in this course are designed to help you successfully achieve the course learning objectives. They are designed to follow the **Prepare, Practice, Perform** format.

- **Prepare**: Includes reading assignments and lectures to introduce new concepts and ensure a basic comprehension of the material. The goal is to help you prepare for the in-class activities during lecture.

> Tip
>
> The class meetings should be highly engaging and hands-on. I plan to give students ample opportunities to apply the techniques we are learning in class where you have direct instructor support. This only works if you come to class prepared having completed any required readings.

- **Practice**: Includes in-class application exercises where you will begin to apply the concepts and methods introduced in the prepare assignment. The format and design of these activities will vary throughout the semester. The activities will graded for completion, as they are designed for you to gain experience with the statistical and computing techniques before working on graded assignments.
- **Perform**: Includes homework and the project. These assignments build upon the prepare and practice assignments and are the opportunity for you to demonstrate your understanding of the course material and how it is applied to analyze real-world data.

Readings (Prepare)

Before each lecture, you are expected to complete the assigned readings. These readings will introduce you to the concepts that will be discussed in the lecture. The readings are designed to help you prepare for the in-class activities during lecture. I don't assign an absurd volume of readings, nor do I expect you to fully and completely understand every single thing you've read. But you should have a familiarity with the concepts and techniques prior to attending class.

Lectures (Prepare)

Part of the class time will be lectures that introduce new concepts or review topics from the preparation materials. Lectures will **not** repeat everything in the readings, they will instead highlight important and known to be complex concepts and will be supplemented with live coding activities. You are expected to attend every lecture.

Application exercises (Practice)

A majority of class meetings will be dedicated to working on Application Exercises (AEs). These exercises will give you an opportunity to apply the statistical concepts and code introduced in the prepare assignment.

Alternatively, we may devote in-class time to work on team projects. If we do not have an official application exercise for the class period, you will submit a brief written reflection of what you have accomplished during the class period.

Based on their format, AEs have different submission deadlines:

- Computer-based AEs are due by the end of the day of the corresponding lecture period. Specifically, AEs from Tuesday lectures are due Tuesday by 11:59 pm, and AEs from Thursday lectures are due Thursday by 11:59 pm. Submission means pushing your work in your repo to GitHub.
- Written AEs are due at the end of the class period.

Because these AEs are for practice, they will be graded based on completion, i.e., a good-faith effort has been made in attempting all parts.

The four lowest AE grades will be dropped at the end of the semester.

Homework (Perform)

In homework, you will apply what you've learned during lecture to complete machine learning tasks. You may discuss homework assignments with other students; however, homework should be completed and submitted individually. Homework must be typed up using Quarto and GitHub and submitted as a PDF in Gradescope.

Homework assignments are due 11:59 pm on the indicated due date.

The lowest homework grade will be dropped at the end of the semester.

Project (Perform)

The purpose of the project is to apply what you've learned throughout the semester to solve some sort of real-world problem. The project will be completed in teams, and each team will present their work at the end of the semester.

More information about the project will be provided during the semester.

Grading

The final course grade will be calculated as follows:

Category	Percentage
Homework	50%
Project	40%
Application Exercises	10%

~~The final letter grade will be determined based on the following thresholds:~~

Letter Grade	Final Course Grade
A+	>= 98
A	93 - 97.99
A-	90 - 92.99
B+	87 - 89.99
B	83 - 86.99
B-	80 - 82.99
C+	77 - 79.99
C	73 - 76.99
C-	70 - 72.99
D+	67 - 69.99
D	63 - 66.99
D-	60 - 62.99
F	< 60

ℹ Revised letter grade thresholds

Letter Grade	Final Course Grade
A+	>= 95
A	90 - 94.99
A-	87 - 89.99
B+	84 - 86.99
B	80 - 83.99
B-	77 - 79.99
C+	74 - 76.99
C	70 - 73.99
C-	67 - 69.99
D+	64 - 66.99
D	60 - 63.99
D-	57 - 59.99
F	< 57

Graduate requirements for INFO 5940

Students in INFO 5940 have additional expectations in the course:

- INFO 5940 homework will at times be graded against a more stringent rubric
- INFO 5940 students will be grouped together for all projects

The final letter grade will be determined using the same thresholds as for INFO 4940.

Course policies

Academic honesty

TL;DR: Don't cheat!

Please abide by the following as you work on assignments in this course.

Discussing assignments

You may discuss individual homework assignments with other students; however, you may not directly share (or copy) code or write up with other students. Unauthorized sharing (or copying) of the code or write up will be considered a violation for all students involved.

Reusing code

Unless explicitly stated otherwise, you may make use of online resources (e.g. StackOverflow) for coding examples on assignments. You may not directly copy and paste from these sources, but instead you need to adapt the code to fit your specific task. You must explicitly cite where you obtained the code using a code comment # immediately near the appearance of the reused code in the file. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.

Use of generative artificial intelligence (GAI)

Cornell's report on Generative Artificial Intelligence for Education and Pedagogy outlines many of the potential benefits and drawbacks to using GAI in the classroom. In this course, we see the value of coding assistants such as GitHub Copilot and ChatGPT to generate code from text. However as an introductory course, we need to ensure that GAI is not used as a substitute or replacement for student learning. GAI should not be used by students to replace your ability to think clearly. Students who use GAI should use it to **facilitate**, rather than **hinder**, learning.

- **GAI tools for reference purposes:** You may make use of the technology as a reference tool, similar to looking up the documentation for a function or Googling your problem. For example, I hate writing regular expressions. Absolutely loathe it. Say I have a dataset where I need to clean a character column to remove all words that are within double asterisk symbols. I might ask ChatGPT

  How do I make a scatterplot using **ggplot2** in R?

- **GAI tools for writing my code/analysis:** You may use GAI tools to assist in writing code in this class. You are expected to understand how any/all submitted code works. Any assignment for which you use GAI as more than a reference tool will require a written self-reflection to consider how you used GAI tools, what skills you acquired through the assignment, and how you believe you demonstrate mastery of the learning objectives for the course.

  You may not make use of the technology as a substitute for critical thinking. For example, you may not upload your data file to a GAI platform and ask it to create charts and statistical models for you. You are taking this course, not a GAI tool. I reserve the right to orally assess any student on their submissions to verify they meet the learning objectives for the assignment; **students who fail to satisfactorily demonstrate they have met the learning objectives may receive a grade penalty of up to 100% on the assignment.**

- **GAI tools for narrative:** unless instructed otherwise, you may not use GAI to write narrative on assignments. In general, you may use generative AI as a resource as you complete assignments but not to answer the exercises for you.

You are ultimately responsible for the work you turn in; it should reflect *your* understanding of the course content.

> ⚠ Warning
>
> Any violations in academic honesty standards as outlined in the Cornell University Code of Academic Integrity and those specific to this course will result in a 0 for the assignment (or possibly more) and will be reported to the College of Engineering Academic Integrity Hearing Board.

Extra credit

Students can earn up to a maximum of 1 percentage point towards their final grade through the extra credit assignment. This is the only opportunity for extra credit in the course.

Late work & extensions

The due dates for assignments are there to help you keep up with the course material and to ensure the course staff can provide feedback within a timely manner. We understand that things come up periodically that could make it difficult to submit an assignment by the deadline. Note that the lowest homework and lab assignment will be dropped to accommodate such circumstances.

Late work

- A **slip day** allows you to submit an assignment 24 hours after the deadline and still receive credit without a late penalty. You are provided with a total of **4 slip days** for the entire semester. Slip days may be used on **homework assignments**. You can use up to 1 slip day for a given homework assignment.

To use your slip days, just submit your assignment late. No need to email telling us you are submitting using your slip days. Check Canvas to see how many of your slip days you have used before submitting an assignment late.

If you use a slip day, **do not submit anything to Gradescope before the submission deadline**. We may begin grading before the slip day deadline and we will grade whatever submission we see in Gradescope.

If you run out of slip days or fail to submit your assignment prior to the slip day deadline without prior permission then your assignment will not be accepted.

- There is no late work accepted for application exercises, since these are designed to help you prepare for homework.
- There is no late work accepted for project components.

Waiver for extenuating circumstances

If you need a bit of extra time, **please use your slip days**. Slip days are specifically intended for legitimate reasons for needing an extension like disability, religious observance, Title IX, student athletics, medical problems, and military service.

If using your slip days for accommodations is not working for you or if you have an SDS accommodation which includes deadline flexibility, you may request a deadline extension in-advance of the deadline. We will work with you to develop reasonable accommodations that align with your individual situation.

To request a deadline extension:

1. Commit and push the work you have completed up to this point on the assignment.
2. Email soltoffbc@cornell.edu. In your email clearly state
    a. The assignment
    b. What you have already completed on the assignment.
    c. What you have left to complete.
    d. Your proposed deadline extension (e.g. *Monday, February 8th at 11:59pm.*)

Regrade requests

Regrade requests can be submitted beginning at noon the day after an assignment's grade is posted, and must be submitted on Gradescope within a week of when an assignment is returned. Regrade requests will be considered if there was an error in the grade calculation or if you feel a correct answer was mistakenly marked as incorrect. Requests to dispute the number of points deducted for an incorrect response will not be considered. Note that by submitting a regrade request, the entire question will be graded which could potentially result in losing points.