STAT 4710/5710: Modern Data Mining

Linda Zhao, Spring 2026

E-mail: lzhao@wharton.upenn.edu
Office: WARB 403
Class Room: JMHH 255

Web: TBD
Office Hours: 3:00-5:00 pm Wednesdays or by appointment
Class Hours: TR: 10:15 - 11:44am (401); 12:00 - 1:29pm (402)

Modern Data Mining

Course Description

In an era shaped by data, algorithms, and rapid advances in artificial intelligence, Modern Data Mining equips students with a deep and principled understanding of statistical and machine-learning methods for real-world data analysis. Rather than a tour of techniques, the course focuses on why methods work, when they should be used, and how modeling choices shape empirical conclusions and decisions. Following the full data-science pipeline—from data acquisition and exploratory analysis to reproducible reporting in RMarkdown—the course develops core modeling ideas through dimension reduction, clustering, regression, and classification, before advancing to modern and high-dimensional methods such as LASSO, neural networks (via Keras), and large language models (LLMs) like ChatGPT and Hugging Facetools. Model-free approaches such as random forests and boosting are also studied, with every method grounded in substantive case studies drawn from finance, healthcare, social sciences, and entertainment. Topics in data-driven decision making, including bandits and reinforcement learning, will be introduced if time permits.

The detailed lecture materials, organized in topic-based modules, resemble comprehensive book chapters. Each module begins with clear objectives and a relevant case study, then addresses the purpose and necessity of each method, dives into mathematical modeling and estimation, and culminates in practical implementation and analysis. These materials offer a rich, integrated learning experience, combining narrative, visualizations, and executable R code, all within RMarkdown format.

Through sustained hands-on programming in R, students develop both theoretical insight and applied fluency without requiring prior coding experience. By the end of the course, students will be able to analyze complex, modern datasets with confidence and to reason critically about models, assumptions, and their real-world implications.

This course is cross-listed as STAT 4710 for undergraduates and STAT 5710 for graduate students.

Prerequisites

Two semesters of statistics courses, familiarity with multiple regressions is assumed. To prepare yourself with the data science workflow, R for Data Science is a good reference.

Methods covered (mostly)

Part I: Acquiring, preparing, exploring and visualizing data

- R/Rstudio/Knitr
- Study design and data acquisition/preparation
- Exploratory Data Analysis (EDA)
- Principal Components Analysis (PCA)
- Clustering
- Matrix completion (Recommendation system)
- Missing data

Part II: Model-based supervised learning

- Multiple regression
- Robust standard error estimation
- AB testing/Multiple testing
- Step-wise regression (Cp/AIC, BIC)
- Training and testing errors
- k-fold cross validation
- Bootstrap
- Penalized regression: LASSO, Ridge Regression, Elastic Net
- Logistic Regression/Multi-Nomial regression
- Classification/ROC/AUC and FDR

Part III: Machine learning

- Neural network/Deep learning
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs) / Time Series
- Large Language Models (LLMs), Transformers, and Attention Mechanisms
- Text mining/Natural Language Processing (NLP)
- Image Processing
- Tree based methods (Bagging, Random Forest and Boosting)

Part IV: Special topics

- Bandits
- Reinforcement Learning
- Time Series

Case study/Datasets

Most of the following cases will be covered:

- Billion dollar Billy Beane
- Wharton Business Radio Audience Estimation via Amazon Mturk
- Discrimination against women in STEM fields?
- Gene expression data miracles
- What controls housing price?
- COVID-19: Lock-down and Compliance
- Who tweets for Trump?
- Framingham heart disease study

- Diabetes/Health care (Predicting Readmission Probability for Diabetes Inpatients to Save Health care Cost)
- IQ=Success?
- Boost return by 80% in Lending Club?
- Using Yelp reviews to predict the rating (text mining)
- Which patent is more likely get approved? LLM and beyond
- Handwriting recognition (image recognition)
- Can we do something to reduce crime rates?
- What can we do to improve education – Texas third graders?
- Whose political bill is more likely to be approved in the sea of bills proposed by politicians?
- McGill Billboard – how long a song can sit on the board?
- Out of 502 stocks can we do better than S&P500?
- How to be successful at Kickstarter
- Chinese Annual Industrial Survey
- Hunting for important gene expression positions to help out with HIV positive patients

And more!

Course Materials

Software

The free and open source statistical computing language R is used through RStudio(Poist). There are infinitely many new packages available for us to use; a pretty interface to explore the publicly available R packages is available via Poist now. (used to be hosted by Microsoft) We will use RMarkdown for all materials to ensue reproducibility. Students are encouraged to use Git and GitHub for version control and collaboration.

Throughout of the semester, we use the free RStudio, an interface for writing R documents and working with data.

Install the following software: R, RStudio and RMarkdown. Detailed instructions are available in canvas.

Tutorials

- Basic R tutorial (Available in canvas)
  - Get_staRted.Rmd/Get_staRted.html
- An advanced R tutorial
  - advanced_R_tutorial.Rmd
  - covering `dplyr`, `data.table` and `ggplot`

Lecture notes

Over the years we have been developing our own lecture notes. They are organized by topic and written in reproducible RMarkdown format which combines R codes, visualizations, and narrative text. Real case studies are deployed throughout. The methods are explained through insightful ideas with minimum mathematics. Some deeper explanations and useful materials are postponed in Appendices as references. Students are urged to read through before classes and put hands on line by line at some point.

Textbooks

While we suggest you to read through thoroughly our own lecture notes which often cover more materials, we also suggest you to have the following two books:

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Application in R (ISLR)*, Available freely online, Second Edition, 2023, Springer New York.
- Garrett Grolemund & Hadley Wickham, *R for Data Science*, 2023, O'Reilly. Available freely online.

A reference for general classical statistics method you may check:

- Ramsey and Schafer, *The Statistical Sleuth*, Third Edition, 2013, Brooks/Cole (an e-version is available in the canvas site)

Course Policies

Communication

Communication will be through Canvas and through Ed Forum. Files will be uploaded to Canvas, including datasets, homework, and lecture notes. Ed is a useful forum for students to ask/answer questions.

Laptop Policy

A **laptop** is a must for the course. You are encouraged to bring the laptop to classes so that you may run the lecture code simultaneously with the professor. However, it is not allowed to use the laptop for other purposes during the lectures. Cell phones must be turned off.

Assignments and Exams

Homework: We will give 5 homework assignments. These can be done in groups of up to 3 people; see the Group Policy for more details.

Quizzes: There will be three in-class, individual quizzes, administered on paper and consisting of multiple-choice questions. Quizzes are designed to assess conceptual understanding and modeling judgment rather than computation. *No makeup quizzes will be offered. The higher grade of the first two short quizzes will be doubled.* Contact the instructor in advance for special cases.

- **Quiz 1: Tue 2/10, in class**
- **Quiz 2: Tue 3/17, in class**
- **Quiz 3: Thu 4/23, in class**

Midterm: Mon 3/30, 7:00-9:00 PM This exam will be an in class, *individual*, open-book and done on the computer. You will be given an exam in RMarkdown format to work through. All TAs will be available to answer questions. Previous exams are available on Canvas. **Section 401: SHDH 350, Section 402: SHDH 3510**

Generative AI tools of any kind are strictly prohibited during all teh quizzes and the midterm exam. Use of such tools constitutes a violation of academic integrity.

Final Project:

- **10 minute presentation on Fri 5/1**
- **Final report: Sun 5/3, before 11:59 pm**

The final project is designed to give students experience conducting a complete, modern data-driven study—from problem formulation and data acquisition to modeling, interpretation, and communication of results. Students will work in groups of up to three members.

Each group is required to give a 10-minute project presentation before a member of the TA team, prior to submitting the final written report. The presentation is an integral part of the project and is intended to:

- evaluate students' ability to clearly communicate findings orally,
- help ensure balanced contributions among group members,
- and reinforce the importance of explaining modeling choices and conclusions.

A complete written report is required and should be suitable for inclusion in a professional portfolio or CV.

Project requirements:

- A well-motivated and relevant research question
- Originality, appropriate complexity, and analytical depth
- Clear interpretation of results and limitations
- A complete written report
- Maximum length: 15 pages (excluding references and appendices)

Potential data sources include (but are not limited to):

- Kaggle is a good place to find a data set.
- Google provides public dataset through BigQuery on Google Cloud Platform.
- gapminder
- UCI Machine Learning Repo
- TO ADD MORE SOURCES

Late Work Policy

It is imperative that you manage your workload properly for this course. We will allow late assignments up to 3 days late, with a 15% penalty per day. Note that lateness will be determined by the timestamp on Canvas submissions, i.e. 12:01 AM is considered late.

Group Policy

The homework and the final project can be done by groups of up to three people (can be from either sections). Sign up for groups on Canvas as soon as possible but no later than **Sat 1/24**. We will help out for those who need to find a group, with searches on Ed.

Please note that at no time may a group have more than 3 members. In addition, while those within a group will submit a single homework file for the group, students must follow the code of academic integrity in regards to classmates outside their group. Finally, students do not have to complete the final project in the same group as for homework. They may form a new group

though again no more than 3 people may be in a group. We prefer you keep the same groups through the semester but it is not required.

Grading Policy

- Homework: 30%
- Quizzes: 15% (5% for quiz 1 & 2; 10% for quiz 3)
- Midterm Exam: 35%
- Final Project: 20%

Professor Zhao may make adjustments for those who actively contribute to the class throughout the semester.

Generative AI/ChatGPT Use Policy

ChatGPT, a state-of-the-art language model developed by OpenAI, has emerged as an invaluable tool in various fields, especially in Data Science. Its ability to understand, generate, and analyze natural language text makes it particularly useful for data scientists. In our course, ChatGPT can assist in exploratory data analysis, hypothesis generation, code synthesis, and even complex problem-solving, allowing students to engage with the material more deeply and intuitively.

Encouragement of Wise Use

While we recognize the immense potential of ChatGPT, we encourage students to use it wisely and ethically. This means understanding its limitations, being aware of potential biases in generated content, and using it as a tool to augment, not replace, your critical thinking and problem-solving skills. Students should strive to use ChatGPT as a collaborative partner in their learning journey, leveraging its capabilities to enhance their understanding and creativity.

Policy on Incorporating ChatGPT in Coursework

- Homework and Final Projects:

Students are allowed and even encouraged to use ChatGPT for homework assignments and final projects. It can serve as a valuable resource for brainstorming, coding assistance, data analysis, and more.

Transparency Requirement: Any work submitted must include a clear, detailed explanation of how ChatGPT was used. This includes specifying what parts of the work were assisted by ChatGPT, how the tool's output was integrated, and any modifications or interpretations made by the student. Quizzes and Midterm:

- Quizzes and the midterm:

To ensure a fair and accurate assessment of individual understanding and skills, the use of ChatGPT **will not** be allowed during quizzes and the midterm exam. These assessments are designed to evaluate your independent critical thinking, problem-solving abilities, and mastery of the course material. It's essential to develop and demonstrate your knowledge without the assistance of AI tools in these scenarios.

Commitment to Ongoing Support and Integration:

ChatGPT is new to all of us, and its potential in education is just beginning to be explored. As instructors, we are committed to:

- Providing Helpful Resources:

We will provide links and resources to help you understand and effectively use ChatGPT. This includes best practices, ethical considerations, and innovative ways to integrate the tool into your learning.

- Incorporating into Course Design:

We will continuously explore how ChatGPT can be integrated into homework and other assignments. This may involve creating tasks specifically designed to utilize AI in problem-solving or analysis.

- Adapting and Learning Together:

As we all learn more about the capabilities and limitations of ChatGPT, we will adapt our teaching methods and the course content accordingly. We encourage feedback and discussion about how the tool is used and its impact on learning.

Class Schedule

Tentative and subject to change.

Week 01, 01/12 - 01/18:

- **Thu 1/15**: Data acquisition and preparation
- **Thu 1/15**: TA onsite Office Hours, 4:00 - 5:30, Place: JMHH 270

Week 02, 01/19 - 01/25:

- **Tue 1/20**: Exploratory data analysis (EDA)
- **Thu 1/22**: Dimension Reduction/Principal Component Analysis (PCA)
- **Sat 1/24**: **Grouping due on Canvas**

Week 03, 01/26 - 02/01:

- **Tue 1/27**: PCA/SVD
- **Thu 1/29**: Clustering
- **Sun 2/1**: **Homework 1 due**, before 11:59 PM to Canvas

Week 04, 02/02 - 02/08:

- **Tue 2/3**: Linear regression
- **Thu 2/5**: Continued topics

Week 05, 02/09 - 02/15:

- **Tue 2/10**: **Quiz 1**. Continued topics
- **Thu 2/12**: Continued topics

Week 06, 02/16 - 02/22:

- **Tue 2/17**: K-fold Cross Validation / LASSO
- **Thu 2/19**: LASSO
- **Sun 2/22**: **Homework 2 due**, before 11:59 PM to Canvas.

Week 07, 02/23 - 03/01:

- **Tue 2/24**: Logistic regression, MLE
- **Thu 2/26**: Continued topics

Week 08, 03/02 - 03/08:

- **Tue 3/3**: Classification (ROC, AUC, FDR). Bayes rule
- **Thu 3/5**: Classification (ROC, AUC, FDR). Bayes rule

Week 09, 03/09 - 03/15:

- **Tue 3/10**: Spring Break
- **Thu 3/12**: Spring Break

Week 10, 03/16 - 03/22:

- **Tue 3/17**: **Quiz 2**.Text mining
- **Thu 3/19**: Neural Network/Deep Learning/Keras
- **Sun 3/22**: **Homework 3 due**, before 11:59 PM to Canvas. –>

Week 11, 03/23 - 03/29:

- **Tue 3/24**: Deep Learning/ CNN
- **Thu 3/26**: Deep Learning/ RNN

Week 12, 03/30 - 04/05:

- **Mon 3/30**: **Midterm Exam** 7:00 - 9:00 PM. 401: SHDH 350, 402: SHDH 351
- **Tue 3/31**: Large Language Models (LLMs)/ Transformers and Attention Mechanisms
- **Thu 4/2**: LLMs/Huggingface
- **Sun 4/5**: **Homework 4 due**, before 11:59 PM to Canvas.

Week 13, 04/06 - 04/12:

- **Tue 4/7**: Decision trees
- **Thu 4/9**: Bagging/Random Forest

Week 14, 04/13 - 04/19:

- **Tue 4/14**: Random Forest
- **Thu 4/16**: Boosting
- **Sun 4/19**: **Homework 5 due**, before 11:59 PM to Canvas.

Week 15, 04/20 - 04/26:

- **Tue 4/21**: Special topic (Bandit/RL)
- **Thu 4/23**: Quiz3, Special topic (Bandit/RL)

Week 16, 04/27 - 05/03:

- **Tue 4/28**: **Final project presentation**
- **Fri 5/1**: **Final project presentation**.
- **Sun 5/3**: **Final project due before 11:59 PM to Canvas**