MLDS 401: Predictive Analytics 1
Fall Quarter, 2023
Professor Malthouse
Mondays and Wednesdays, 3:00-4:20pm, Krebs Hall
Labs: Tuesdays, 15:00–16:00, Krebs
Instructor: Professor Edward C. Malthouse
Office: Fisk 304D, 1845 Sheridan Road
Phone: 847-467-3376
email: ecm@northwestern.edu
TAs: Yayu Zhou and Mengfan Xu

This is the first course in a two-course sequence (401, 420) covering predictive analytics. Broadly speaking, 420 covers data-driven, (mostly) non-parametric models for prediction and exploration models. This course (401) focuses on theory-driven, (mostly) parametric, confirmatory models. It can be divided into two main parts, each followed by an exam: (1) fundamentals of linear regression and (2) generalized linear models. The parts will build on each other. The course will cover different models, how and why they work, estimation, model evaluation and metrics, and issues involved in applying the models to real problems.

Prerequisites: undergraduate linear algebra and calculus; two-semester, calculus-based, 300-level sequence in probability and statistics at the level of Montgomery and Runger or Devore. Your previous statistics courses should have covered multiple regression in some detail including the underlying math, although I will start from the beginning and move fast. Predictive models are a vast subject and my selection of topics will be eclectic, drawing from many fields. Even if you were a statistics major there will be new ideas and approaches covered in this class. My emphasis is on the art and science of applying the models to real problems rather than just the math (which is often the focus of math and stats courses on the subject).

Office hours.

1. Ed: Mondays and Wednesdays, after class and by appointment. My office is located on the south end of campus in Fisk Hall (15 minutes by foot from Krebs), and so it might be more convenient to meet after (or before) class in the MSiA space. It is best to send me an email if you want to meet at the MSiA space. I have blocked 4:20–5:00 in my calendar every Monday and Wednesday. Send me an email if you want to meet before class.

2. Yayu: Tuesdays 16:00–17:00

3. Mengfan: Thursdays, 16:00–17:00

Software: I will teach and use R and Python in class. You may also use Matlab, SAS, etc. My focus is less on software and more on understanding the fundamental models, how they work, and how they are applied to solve real problems.

References

Course Materials

1. Course packet.
2. Handouts, homework assignments, data sets, announcements, etc. will be posted to Canvas.
3. **(ACT)** Tamhane (2019), *Predictive Analytics: Parametric Regression and Classification Models*

Useful References: (if you want to dig deeper or see alternative presentations of material)

- **(ISL)** James, Witten, Hastie, and Tibshirani (2021). *An Introduction to Statistical Learning*, Vol. 2. New York: springer. This is a classic, but it is more of an undergraduate textbook
- **(KNN)** *Applied Linear Regression Models*, any Edition by M. H. Kutner, C. J. Nachtsheim, and J. Neter, McGraw-Hill. This is a classic linear models textbook.
- **(ESL)** *The Elements of Statistical Learning: Data mining, inference and prediction*, Hastie, Tibshirani and Friedman, Springer. This gives a more advanced treatment of the topics. It is excellent and free.
- Myers, Montgomery, Vining and Robinson, *Generalized Linear Models, with applications in engineering and the sciences*. This is good statistics textbook at the right level, and with many topics we will not get to.

Student evaluation. Your course grade will be determined as follows:

1. *Midterms:* 30% in total
2. *Final exam:* 35%. The final will be comprehensive, but with more emphasis on part 3 (generalized linear models)
3. *Homework:* 20%. Roughly weekly problem sets worked in groups.
4. *Project:* 15%. Work in groups. The project will involve real data sets. There will be some focused research questions. You will have to build appropriate models to answer the questions and present the findings to a mixed audience of data scientists and domain experts/decision makers.

For homework and project, Work in assigned teams of 3–4 students, submit one copy with all names on it.

Exam policies:

- I do not give makeup exams. Please attend the exam at the scheduled time.
- You may use a pocket calculator, one 8.5 by 11 inch sheet of notes on the midterm and two sheets of this size on the final.
- I will distribute practice exams given during previous years.
- You must show your work on the exam to receive credit. I will give full credit if you get the right answer and show your work.

Honesty, Plagiarism, and Cheating. The code of student conduct for all Northwestern University students is contained in the Student Handbook. Questions of academic dishonesty, cheating, plagiarism and other violations, and the terms and conditions are all listed in the handbook.

Course Outline:

Daily class schedule

1. *Wednesday, Sep 20, 15–16:20.* Overview of supervised learning: modeling processes, deductive versus inductive reasoning, typology of applications (e.g., confirmatory, exploratory, descriptive, predictive, prescriptive), typologies of models (e.g., regression versus classification; parametric versus nonparametric, supervised versus unsupervised, count, censored, time series, panel, etc.), scientific versus statistical models
   - Skim Breiman (2001), Statistical modeling: The two cultures. *Statistical science*, 16(3), 199-231. **Read** DR Cox's response (pages 216–218). You may want to skim Efron's comment and Breiman's rejoinder.
   - Cross-industry standard process for data mining
   - Supplementary reading: **ACT Ch. 1; ISL Ch. 1.**; KNN §1.2, 1.4, 9.1
2. *Monday, Sep 25, 15:00–16:20.* Overview of using linear regression: model, interpretation, standard errors, CIs, PIs, hypothesis tests, R^2 (**ACT §2.1–2, 3.1–3; ISL §3.1-3.2.**)
3. *Wednesday, Sep 27.* Least squares theory: matrix notation, objective function, normal equations with closed-form solution, key properties including unbiasedness, variance, normality, BLUEness (**ACT §3.8.**)
4. *Monday, Oct 2.* Model assumptions, effects of violations, diagnostics and remedies: residual plots, QQ plots, variance stabilizing transformations, Tukey's ladder of transformations (**ACT §4.1–3; ISL §3.3.**)
5. *Wednesday, Oct 4.* Quadratic models, multiplicative models, outliers, leverage, Cook's distance (**ACT §4.4, 5, 7.**)

6. *Monday, Oct 9.* Newfood case. Multicollinearity: definition, detection, effects and remedies; VIF; omitted variable bias, pipes, forks, colliders (**ACT §4.8.**)

7. *Wednesday, Oct 11.* Quality control case. Simpson's paradox, Berkson's paradox.

8. *Monday, Oct 16.* ANOVA decomposition, extra/partial sums of squares, F test (**ACT §3.3.**)

9. *Wednesday, Oct 18.* Dummy variables and interactions and (ACT §3.6)

10. *Monday, Oct 23.* Linear models for prediction and exploration: overfitting, bias-variance tradeoff, test/validation sets, k-fold cross validation, evaluation metrics (**ISL §5.1, 2.2.**

11. *Wednesday, Oct 25.* Midterm: covers material through Oct 18. ACT Chapters 1–4 but not §3.7. Or ISL Ch. 3 but not §3.5

12. *Monday, Oct 30.* Automated model selection/regularization (stepwise, lasso, ridge) (**ACT 5.1, 5.2, 6.1–2; ISL §6.1–2.**)

13. *Wednesday, Nov 1.* Principal components analysis (PCA) and regression (PCR) (**ACT §5.3; ISL §6.3, 12.2.**)

14. *Monday, Nov 6.* Binary logistic regression: model, interpretation, odds ratios, CIs, Wald test (**ACT 7.1–2; ISL §4.1–3.**)

15. *Wednesday, Nov 8.* MLEs, deviance, LRT, AIC, model evaluation (accuracy, AUC, recall, precision), class imbalance (ACT 7.3–5)

16. *Monday, Nov 13.* Multinomial, softmax, multi-class evaluation metrics (e.g., micro precision)(**ACT 7.6; ISL 4.3.5.**)

17. *Wednesday, Nov 15.* Count data: Poisson regression (**ACT 9.4; ISL 4.6.**)

18. *Monday, Nov 20.* Finish Poisson models: residuals, evaluation, overdispersion, mention model extensions, e.g., NBD, ZIP

19. *Wednesday, Nov 22.* This is the day before Thanksgiving and the university closes after our class ends. I will be in class and will give a lecture on customer lifetime value models. This topic will *not* be on the homework or exams, but it is an area of application for survival models and is important. The topic also provides context for the final project. This is also one of my favorite lectures and I have given this to many different audiences over the past 20 years. Those who are traveling and want to be hear the lecture can watch my videos on the topic.

    - 1. Introduction to CLV (22.2 min) (Youtube) (Panopto)
    - 2. Simple retention model (30.0 min) (Youtube) (Panopto)

- 3. Simple retention model in Excel (7.4 min) (Youtube) (Panopto)
- 4. General retention model (33.6 min) (Youtube) (Panopto)

20. *Monday, Nov 27.* Survival analysis: Survivor and hazard functions, Kaplan-Meier estimate (**ACT §9.1–2; ISL 11.1–11.4.**)

21. *Wednesday, Nov 29.* Survival analysis: Cox model (**ACT §9.4; ISL 11.5.**)

22. Final project presentation. Each team should schedule a 20-minute time to present the project to me on Dec 1, 4, 5 or 6. I will provide more details later, but here are some basic guidelines. In real life you would want to provide background, establish the importance of the topic, give information about data you used, etc. Given the time constraints you can skip these parts (I don't want to hear about the data 16 times). I want to see the following: (1) conceptual framework with justification of why there might be a causal relationship, (2) information about any *additional* data you gathered beyond what I provided, (3) your model with estimates, (4) summary of your conclusions, and (5) discussion of the managerial implications. Do *not* tell me everything you did—show me your final model (you can provide additional robustness checks in an appendix). You might want to have some basic descriptive statistics such as a correlation matrix and univariate statistics (n, mean, SD, skewness, min, max) on the variables handy. Stay on time and do not run over! Plan for 15 minutes plus 5 minutes for discussion and feedback.

23. **Final exam: Dec 4, 15:00–17:00** (comprehensive with emphasis on material covered starting Oct 23, i.e., ACT Chs. 5, 6, 7, 9, 10; ISL Chs. 4, 5.1, 6, 11)

Academic Integrity Statement

Students in this course are required to comply with the policies found in the booklet, "Academic Integrity at Northwestern University: A Basic Guide." All papers submitted for credit in this course must be submitted electronically unless otherwise instructed by the professor. Your written work may be tested for plagiarized content. For details regarding academic integrity at Northwestern or to download the guide, visit: here

Accessibility statement

Northwestern University is committed to providing the most accessible learning environment as possible for students with disabilities. Should you anticipate or experience disability-related barriers in the academic setting, please contact AccessibleNU to move forward with the university's established accommodation process (e: accessiblenu@northwestern.edu; p: 847-467-5530). If you already have established accommodations with AccessibleNU, please let me know as soon as possible, preferably within the first two weeks of the term, so we can work together to implement your disability accommodations. Disability information,

including academic accommodations, is confidential under the Family Educational Rights and Privacy Act.

Religious Observance Statement Northwestern is committed to fostering an academic community respectful and welcoming of persons from all backgrounds. To that end, the policy on academic accommodations for religious holidays stipulates that students will not be penalized for class absences to observe religious holidays. If you will observe a religious holiday during a class meeting, scheduled exam, or assignment deadline, please let me know as soon as possible, preferably within the first two week of class. If exams or assignment deadlines on the syllabus fall on religious holidays you observe, please reach out so that we can discuss that coursework.

Diversity, equity and inclusion This course strives to be an inclusive learning community, respecting those of differing backgrounds and beliefs. As a community, we aim to be respectful to all students in this class, regardless of race, ethnicity, socio-economic status, religion, gender identity or sexual orientation.

Exceptions to class mobility

Class sessions for this course will occur in person. Individual students will not be granted permission to attend remotely except as the result of an Americans with Disabilities Act (ADA) accommodation as determined by AccessibleNU.

Maintaining the health of the community remains our priority. If you are experiencing any symptoms of COVID do not attend class. Follow the steps outlined on this site for testing, isolation and reporting a positive case. Next, contact your instructor as soon as possible to arrange to complete coursework.

Students who experience other personal emergencies should contact the instructor as soon as possible to arrange to complete coursework.

Should public health recommendations prevent in-person class from being held on a given day, the instructor or the university will notify students.

Guidance on class recordings

Portions of this class might be recorded by the instructor for educational purposes. I will communicate how members of the class can access the recordings. Portions of the course that contain images, questions or commentary/discussion by students will be edited out of any recordings that are saved beyond the current term.

Prohibition of recording of class sessions by students

Unauthorized student recording of classroom or other academic activities (including advising sessions or office hours) is prohibited. Unauthorized recording is unethical and may also be a violation of University policy and state law. Students requesting the use of assistive technology as an accommodation should contact AccessibleNU. Unauthorized use of

classroom recordings – including distributing or posting them – is also prohibited. Under the University's Copyright Policy, faculty own the copyright to instructional materials – including those resources created specifically for the purposes of instruction, such as syllabi, lectures and lecture notes, and presentations. Students cannot copy, reproduce, display, or distribute these materials. Students who engage in unauthorized recording, unauthorized use of a recording, or unauthorized distribution of instructional materials will be referred to the appropriate University office for follow-up.

Support for Wellness and mental health

Northwestern University is committed to supporting the wellness of our students. Student Affairs has multiple resources to support student wellness and mental health. If you are feeling distressed or overwhelmed, please reach out for help. Students can access confidential resources through the Counseling and Psychological Services (CAPS), Religious and Spiritual Life (RSL) and the Center for Awareness, Response and Education (CARE). Additional information on all of the resources mentioned above can be found here:

- https://www.northwestern.edu/counseling/
- https://www.northwestern.edu/religious-life/
- https://www.northwestern.edu/care/

The Writing Place

When working on writing assignments for this class, I encourage you to visit the Writing Place, Northwestern's peer writing center. You will work with juniors and seniors who have been trained to provide you feedback and assistance on any type of writing at any stage in the writing process. They will not edit your work. Rather, they will work with you to brainstorm ideas, organize or outline an essay, clarify your argument, document your sources correctly, or refine grammar and style.

To book an appointment, register for an account here.