NORTHWESTERN UNIVERSITY
MLDS 420: Predictive Analytics II, Winter 2024

Instructor: Prof. Daniel Apley; Rm. TECH M235; 847-491-2397; apley@northwestern.edu; office hours: Wed 12:30-1:30 M235.

TA: Wei Liu (weiliu2025@u.northwestern.edu); Office Hours: Th 3:00-4:00 in room MEC 429. HW Grader: Ruiqi Wang (ruiqiwang2025@u.northwestern.edu).

Lecture: M/W 2:00-3:20, North Garage, Krebs Rm 1440. All lecture material is also prerecorded and should be viewed to supplement and reinforce in-class lectures. Links to the videos will be posted on Canvas (under the Panopto link). Video viewing schedules are in the weekly course schedule.

Lab: Wed 11:00—12:00, Krebs Classroom. In the lab sessions, the TA will cover Python implementations of the various supervised learning models (the videos and Powerpoint slides focus on R implementations). The midterm may possibly be during a lab session.

Canvas: Lecture notes, handouts, homework assignments, data sets, announcements, lecture material videos, course schedule, etc. will be posted regularly.

Prerequisite: MSiA 401 or equivalent

Texts:

- (JWHT) *An Introduction to Statistical Learning with Applications in R, second edition*, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, Springer Texts in Statistics. **ISBN-13:** 978-1071614174. Available for free at https://www.statlearning.com/ or https://hastie.su.domains/ISLR2/ISLRv2_website.pdf (you may have to run NU VPN to have access)
- (HTF) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed*., by Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2009. Available for free from author's website https://web.stanford.edu/~hastie/ElemStatLearn/. (thorough, expansive coverage of statistical learning, but aimed at a statistically mature audience)

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

- (VR) *Modern Applied Statistics with S, 4th ed.* William N. Venables and Brian D. Ripley, Springer, 2002. (classic reference for R).
- (BJR) *Time Series Analysis: Forecasting and Control, 3rd ed*., G. Box, G. Jenkins, and G. Reinsel, Prentice-Hall, 1994, Upper Saddle River, NJ. (classic time series reference)
- (KNN) *Applied Linear Regression Models, 4th Edition* by M. H. Kutner, C. J. Nachtsheim, and J. Neter, McGraw-Hill/Irwin, 2004, ISBN-10: 0073014664, ISBN-13: 978-0073014661. (more readable coverage of extensive linear regression concepts and certain nonlinear and nonparametric regression)

Grading:		
	Midterm	25% (Mon, 2/12/24, in class)
	Final	25% (Mon 3/11/24, 12:00-2:00, Rm 1440)
	HW	25%
	Project	25%
	Participation Bonus	see below

Software: R, which is open source freeware that you can download (along with tutorials, etc) from www.r-project.org. Labs will cover Python.

Course Objectives: This course covers nonparametric modeling of complex, nonlinear predictive relationships in data with categorical (classification) and numerical (regression) response variables. It is a follow-up course to Predictive Analytics I, which covers linear and logistic regression methods. The course learning objectives are to:

1. Understand the most popular and effective modern nonlinear/nonparametric predictive modeling methods and how to expertly use them.
2. Identify which method is most appropriate for a particular data set and predictive modeling problem.
3. Empirically (based only on the data) fit and tune the predictive models for optimum predictive performance.
4. Interpret any black-box fitted model, in terms of how the different predictor variables effect the predicted response and which are the most important.
5. Build the knowledge and experience to be able to quickly and expertly self-learn new predictive modeling methods as they are developed and introduced to the predictive modeling community.

Course Outline: (Since JWHT targets a broad audience and often lacks technical details, and HTF is often dense and difficult to read, the lecture notes are an important complement)

1. Some fundamental concepts in supervised learning (5.5 lectures)
   - Maximum likelihood estimation (Class notes; HTF Secs. 2.6.3, 4.4.1, 8.2.2; JWHT Sec. 4.3.2; and many other places in specific contexts; search the textbooks for "likelihood")
   - Nonlinear least squares (Class notes)
   - Quantifying statistical uncertainty: Bootstrapping (JWHT Sec. 5.2; HTF Sec. 7.11), Information matrix (HTF Sec. 8.2.2 and other specific contexts)
   - Model selection and evaluation criteria: C_p, *AIC*, *BIC*, and $PRESS_p$ (JWHT Secs. 2.2, 3.2.2, 6.1; HTF Sec. 7.5), Cross Validation (JWHT Sec 5.1; HTF Sec. 7.10; CV handout)
   - Ideal Bayes classifier and predictor (JWHT Sec 2.2.3; HTF Sec. 2.4)
2. Nonparametric/generic regression and classification 1 (5.5 lectures)
   - Neural networks (HTF Ch. 11, JWHT Ch 10)
   - Interpreting black-box supervised learning models (Class notes)
   - Shrinkage and bias/variance tradeoff, ridge regression, LASSO, stepwise (JWHT Sec 2.2 and Ch 6; HTF Sec. 3.4)
   - Classification and regression trees (JWHT Secs. 8.1, 8.3; HTF Sec. 9.2)
3. Nonparametric/generic regression and classification 2 (5 lectures)
   - Nearest neighbors (JWHT Sec 2.2 and 3.5; HTF Sec. 2.3.2—2.3.4 and 13.3)
   - Local linear regression and kernel smoothing (JWHT Ch 7; HTF Sec. 2.8.2 and 6.1—6.5), generalized additive models (JWHT Sec 7.7; HTF Sec. 9.1), projection pursuit regression (HTF Sec. 11.2), and basis functions (JWHT Sec 7.3; HTF Sec 2.8.3).

- Ensemble/committee methods: Bagging (JWHT Secs. 8.2, 8.3; HTF Sec 8.7), stacking (HTF Sec 8.8), boosting (JWHT Secs. 8.2, 8.3; HTF Ch. 10), and random forests (JWHT Secs 8.2, 8.3; HTF Ch. 15)
- Overview of other methods and concepts (MARS, SVMs, Naive Bayes, etc)

4. Time series analysis and forecasting (2 lectures)
    - Components of time series (class notes)
    - Additive and multiplicative models (class notes)
    - MA and EWMA smoothing and forecasting (class notes)
    - Holt, and Holt-Winters forecasting (class notes)
    - Decomposition methods (class notes)
    - ARMA modeling (class notes; BJR; Ch 14 of VR)

Exams: The midterm and final exams will be closed book and closed notes, except for the following. You are allowed one page of "cheat sheet" notes (8-1/2" by 11", front and back) for the midterm and two pages for the final exam. The exams will stress both concepts and proficiency in interpreting R output results for the various predictive modeling analyses we cover in class. Be comfortable with all of the material we covered in class, in the homeworks, and in the lecture notes, including the "Discussion Points and Questions" slides.

Homework: Homework will be assigned roughly biweekly and will be conducted and turned in individually. It will typically involve analyzing a set of data using the methods covered in lecture and drawing appropriate conclusions. Homeworks must be turned in on time, and late homeworks will not be accepted. The following are some guidelines on the formatting of the homeworks and what to include or not include:

1. Do not include any irrelevant information in submitted homework. This includes things like software warning messages, extra irrelevant plots, message printed out when loading packages, code commented out and not used, traces and iterates of training models, etc.
2. Take some time to format your answers nicely. For example, make sure plots are large enough to read the text, but not so large that you can only fit one plot per page so that a set of plots extends across many pages.
3. Make sure to answer every part of the questions, especially questions like "interpret the result", "compare with other methods", etc.

Project Teams: The projects will be conducted in teams of 4 students (3 or 5 students may be OK, if there is a compelling reason, but it needs approval from the instructor). At the end of the quarter, there will be peer evaluations of the contributions of all team members, and the evaluations will be factored into your project scores.

Project: The project will be conducted, presented, and turned in as a team assignment (see project teams above). The project can involve any subset of the methods covered in class, or perhaps related methods if more relevant. The project should revolve around a particular predictive analytics problem and questions that you would like to answer, and the problem should dictate the methods that you use (as opposed to vice-versa). Ideally, this will be your practicum project, but you can choose another relevant analytics problem and data set if appropriate. You may be able to find an appropriate problem and data from one of the publicly accessible data repositories (e.g.,

http://archive.ics.uci.edu/ml/datasets.html [my favorite], http://www.kaggle.com/competitions, http://www.kdnuggets.com/datasets/index.html, http://lib.stat.cmu.edu/, http://kdd.ics.uci.edu/databases/, http://www.statsci.org/datasets.html, etc). Time permitting, teams will present their results in class, the last week of class.

Final project reports are due Friday, March 8 by 11:59 pm (post an electronic version on Canvas). Biweekly project status updates are due every other Friday (Jan 26, Feb 9, Feb 23) by 11:59 pm. Post updates as pdf files on Canvas.

Some guidelines for the projects:

- Make sure to have an interesting problem or question to be answered, and an appropriately sized data set that will help to analyze the problem.
- The project should involve some "response" variable that you want to predict, as a function of a number of other "predictor" variables that will be available at the time of prediction. This is the most generic supervised learning setting and is the primary subject of this course.
- On the surface, many practicum projects sound like unsupervised learning projects (e.g., the client is asking you to find meaningful market segments or clusters in the data). However, this is often a miscommunication or misunderstanding, and the project is really better handled via supervised learning. For example, when clients ask you to identify "clusters", they usually want those clusters to have a predictive impact on some other available variable like spend, attrition, late payments, etc. If you want to find clusters or combinations of a group of variables (call them predictor variables) that have a predictive impact on some other available variable (call it a response variable), this is a supervised learning problem, by definition. Fit a supervised learning model to predict the response as a function of the predictors and sort the data into groups according to their predicted response (aka their "score"), which is a function of the predictors. If you use unsupervised learning methods to find clusters in the predictor variables, then even if you are able to find meaningful clusters, they may have no relationship with the response variable of primary interest.
- If you obtain your own data, you will almost certainly need to invest significant time in data "pre-processing", including defining appropriate predictor variables from unstructured data repositories, cleaning bad/missing data, transforming heavy tailed variables, etc. This is important and should be discussed in your report, even though it is not the subject of this course. In particular, you should have a lengthy discussion on why you considered the predictors that you considered.
- Most/all projects will involve finding the best possible supervised learning model, in terms of giving the most accurate response prediction. This will involve trying and comparing many different models, from sophisticated to simple (e.g., linear and logistic regression). If linear regression, as opposed to a more sophisticated method, ends up being the best predictor, that is OK. But you should provide compelling evidence that this is the case. If your conclusion is that there is no response predictability (i.e., if the global average is the best predictor), that is OK, too. But try to avoid selecting a project for which this is the case.

- One of the major goals of this course is to build sufficient background for you to pursue further study on your own of other methods that we did not cover. There are a great many supervised learning methods out there. Hence, every project must include trying at least one new method that we did not explicitly cover in class.
- For the final project report, include your R code as Appendices in the electronic (pdf) version of your report. But do not include the code in the hard copy printout. For the electronic version, please label the different parts of your code and refer to the labels when discussing the results in the main body of the text.
- A very important part of the final project report is a clear and detailed description of your data: What were your raw data, what exactly was the data array to which you fit the supervised learning model, what does each row correspond to, what was the response variable for each row, what were the predictor variables for each row, etc. It should be described in enough detail that, given your raw data, someone could reconstruct the data array you used for regression/classification.

Class Participation Bonus: This is purely an opportunity to raise your grade and will in no way lower your grade. If you have good participation, and your final grade is close to the border between two letter grades, then you will be bumped up to the higher of the two grades. The lower boundary of the bump-up margin will be extended for exceptional class participation (e.g., down to the midpoint of the range for the lower grade). Inadequate class participation will not affect your grade. Participation means coming to class (on time), asking questions, and volunteering answers. I may occasionally collect in-class feedback on how well students are absorbing the material via short written answers to a question. Being present and writing a reasonable answer will count towards your participation.

Miscellaneous Announcements:

-