CUNY School of Professional Studies

DATA 624 Predictive Analytics

CUNY SPS Master of Science in Data Science

Fall 2021

Instructor Name: Jeffrey Nieman
Instructor Email: Jeff.nieman@sps.cuny.edu
Class Meetup:
Office Hours:
Degree Program: M.S. in Data Science
Credits: 3 graduate credits
Prerequisites: Data 621
Type of Course: Elective course

Course Description

This course teaches students to use advanced machine learning techniques that are focused on predictive outcomes. Topics will include time series analysis and forecasting, recommender systems, and advanced regression techniques. In addition, students will learn how to evaluate the predictions that result from these techniques, how to assess model quality, and how to improve models over time.

Course Learning Outcomes:

At the end of this course, students will be able to:

Apply advanced regression techniques such as constrained linear (PLS, NIPALS, Ridge, LARS), nonlinear (MARS, SVM, KNN), Trees (RF, Boosted).

Utilize various forecasting techniques to produce reliable and robust forecast models.

Develop recommendation systems using knowledge-based and content-based approaches.

Evaluate the quality of models produced and make recommendations for improvement to models.

Program Learning Outcomes addressed by the course:

- Business Understanding. Students will learn how predictive modeling and forecasting techniques can add value to existing business analytics.
- Data Understanding. Students will learn how to explore data to find patterns that allow for forward-looking forecasts and recommendations.
- Model Implementation. Students will learn to implement models for the various predictive modeling techniques covered in the course, with a focus on recommendations, estimation, and forecasting techniques.

How is this course relevant for data analytics professionals?

Predictive modeling and forecasting are mainstays of the analytics profession. Predictive modeling spans numerous fields and approaches. Indeed, within this course the student will be introduced to a multitude of techniques, some of which fall under the moniker "statistical modeling" while others are referred to "machine learning." For this course it is less important the lineage of a particular technique, but rather the classes of problems to be solved.

Each class of problems introduce multiple techniques. It is likely that the student has encountered many of these approaches in the past. This is both unavoidable and also fortuitous as the bulk of the course can thus focus on applying these techniques to the problem classes as opposed to learning the theory of the techniques.

Assignments and Grading:

Each section comprises an introduction, a reading assignment, and book exercises. In addition, the bulk of the grading is focused on two course-specific projects along with paper submissions detailing methodology and results, including data visualizations.

Book Exercises 30%

Completion of exercises must include working R code along with a brief discussion of the approach and results. All homework sets will be due on Sundays at 11:59 PM. Late homework will be accepted with penalties until the following meetup when the homework will be discussed.

Slack Authoring and Participation 10%

We will create a slack to help the class participate and discuss. This is the primary method to reach out to your classmates and me for help. We will periodically have related topical discussions there as well. You will be graded on how well you participate both in raising issues and responding to each other's questions/discussions.

Project 1 15%

The first student project will be a solo time series and forecasting problem. A professionally written report will be required. Details in Announcements.

Project 2 30%

The second student project will be a group predictive modeling problem. A professionally written report will be required. Part of your grade will be determined by your peers on your contribution to the submission. Details in Announcements.

Lecture 15%

You and your group will be required to prepare and give a lecture for the entire class. We will be meeting in GoToMeetings and you will present from your workstation. There will be 7 presentations, and each will last approximately 20 minutes including Q&A. The 7 topics will be Time Series Decomposition, Exponential Smoothing, ARIMA, Linear Regression and its Cousins, Non-linear Regression, Trees & Rules-based Models and Recommender Systems. Details and Group Assignments on Blackboard.

Required Texts and Materials:

Reading assignments span two primary texts. These are

- Hyndman & Athanasopoulos. "Forecasting: Principles and Practice." https://www.otexts.org/fpp2/
- Kuhn & Johnson. "Applied Predictive Modeling." http://appliedpredictivemodeling.com/ A third book can be used for supplemental reading, which is
- Hastie, Tibshirani, & Friedman. "Elements of Statistical Learning."http://statweb.stanford.edu/~tibs/ElemStatLearn/ Some of the reading will overlap across the two books. Where there is overlap, HA is generally more accessible, acting as an introduction, while KJ is a bit more theoretical. The student is encouraged to exercise judgment as to whether to skip the overlapping content. NOTE: Books are referenced by abbreviation for convenience. Hyndman & Athanasopoulos is abbreviated HA, and Kuhn & Johnson is abbreviated KJ.

Relevant Software, Hardware or other Tools:

This course requires using the R language. Students must be familiar with the language and know how to install packages. All homeworks must be written in R and submitted as code that can easily be cut and copied into R Studio to run. Students must describe in written form their approach and analysis for all problems. The exposition is used to not only determine whether thought processes are sound but also to provide partial credit on problems.

Grading

Grade Distribution

Quality of Performance	Letter Grade	Range %	GPA
Excellent - work is of exceptional quality	A	93 - 100	4
Excellent	A-	90 - 92.9	3.7
Good - work is above average	B+	87 - 89.9	3.3
Satisfactory	B	83 - 86.9	3
Below Average	B-	80 - 82.9	2.7
Poor	C+	77 - 79.9	2.3
Poor	C	70 - 76.9	2
Failure	F	< 70	0

Schedule

Note: Schedule is subject to change.

Dates	Topic
Aug-25 to Sep-05	Welcome and Introductions
Sep-06 to Sep-12	Time Series
Sep-13 to Sep-19	Forecasting
Sep-20 to Sep-26	Decomposition
Sep-27 to Oct-03	Data Preprocessing/Overfitting
Oct-04 to Oct-10	Exponential Smoothing
Oct-11 to Oct-24	ARIMA
Oct-18 to Oct-24	ARIMA
Oct-25 to Oct-31	Project 1
Nov-01 to Nov-07	Linear Regression
Nov-08 to Nov-14	Non-linear Regression
Nov-15 to Nov-28	Trees and Rules
Nov-29 to Dec-05	Recommender Systems
Dec-06 to Dec-13	Project 2

Accessibility and Accommodations

The CUNY School of Professional Studies is firmly committed to making higher education accessible to students with disabilities by removing architectural barriers and providing programs and support services necessary for them to benefit from the instruction and resources of the University. Early planning is essential for many of the resources and accommodations provided. Please see: http://sps.cuny.edu/student_services/disabilityservices.html

Online Etiquette and Anti-Harassment Policy

The University strictly prohibits the use of University online resources or facilities, including Blackboard, for the purpose of harassment of any individual or for the posting of any material that is scandalous, libelous, offensive or otherwise against the University's policies. Please see: http://media.sps.cuny.edu/filestore/8/4/9_d018dae29d76f89/849_3c7d075b32c268e.pdf

Academic Integrity

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the educational mission of the City University of New York and the students' personal and intellectual growth. Please see: http://media.sps.cuny.edu/filestore/8/3/9_dea303d5822ab91/839_1753cee9c9d90e9.pdf

Student Support Services

If you need any additional help, please visit Student Support Services: http://sps.cuny.edu/student_resources/