CUNY School of Professional Studies

Syllabus

DATA 624: Predictive Analytics

Instructor Name: Scott Burk
Instructor Email Address: scott.burk@sps.cuny.edu
Phone: 254-563-6909
Office Hours: Tuesday Evenings 8:30 to 10:00 PM ET and Saturdays 10 to 11 AM ET

Degree Program: M.S. in Data Science
Credits: 3 graduate credits
Prerequisites: DATA 621 Sound Knowledge of R
Type of Course: Elective course

Course Summary

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/Competencies 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 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:

Course assignments	Percentage of Final Grade	Points
Homework Assignments	40%	400
There will be 19 homework problems assigned to re-enforce course concepts and provide implementation experience. This assignment will be group assignments. And will be submitted in two batches.		
Project A	20%	200
Each group will submit a project.		
The first project will be a time series and forecasting problem. Students should submit a professional written report		
Project B	30%	300
Each group will submit a project.		
The second project will be a predictive modeling problem. Students should submit a professional written report		
Discussion Topic Authoring and Participation	10%	100
You will be responsible to research and submit an engaging discussion topic as author.		
An engaging response to 10 fellow student's discussion posts will be required throughout the semester. Answering the questions of your classmates submitted topics		
These need to be completed by the end of the Term		
Total	**100%**	**1,000**

Each section comprises a reading assignment and book exercises. These assignments will be completed and turned-in by teams. In addition, the bulk of the grading is focused on two course-specific team projects submissions detailing methodology and results, including data visualizations. Finally, you will be graded by quality and volume of participation in the discussion

NOTE: All assignments (except discussions) will be team/group efforts. I will assign the team membership. **NOTE: If you do not like team assignments you should consider dropping and enrolling in another section or for a later time**. Your team will elect a 'point person' or representative. This representative will be the person responsible for team submissions (projects and homework assignments). IMPORTANT: After each team assignment, each team member will be allowed to rate their peers. The team will be graded for each assignment. And, then the peer rating may affect each team members assignment grade. Thus, it is possible for a team to be awarded a B+ on a project and an individual to receive a C if all her/his teammates score them with low ratings.

Book Exercises 40% overall, 20% per submission – Two Team Submissions

Completion of exercises must include working R code along with a discussion of the approach and results. Explicit instructions will be given. Assignments will be submitted in two batches via **Microsoft Word** or **Google Docs, not just R Markdown (you may include this separately if desired)**. NOTE: You will work on these and submit them as a team. These will be collected in 2 batches (midterm and end of term). However, you need to keep up with the assignments weekly! NOTE: you may turn in your work early, but not late without penalty. There will be a 5 point late penalty PER DAY for the midterm assignment and no exceptions at the end of semester (zero). So, if you get sick often or other things impair you meeting deadlines you should consider turning them in early. Scoring criteria will be forthcoming, however, you should note a professional clear write-up is required. R Code must be included in report in Courier for easy copy/paste.

Discussion Topic Authoring and Participation 10% - Solo Author, Multiple Class Responses, Response to Questions and General Participation

There are 2 parts to your discussion grade. First, you will be responsible to research and submit an engaging discussion topic as author. I will provide some examples that you can use or select your own, relevant data science topic. It is best practice to include relevant citations and attribution. You will also respond to other discussions weekly. This is an individual, not team assignment.

Part 2 is general responses to the discussion board, 'Getting Acquainted', and responses to fellow students. Example, one of the discussion boards is 'Ask the Class!', this is meant for scripting, HW and other questions. Taking time to respond to these questions (no explicit HW answers please) will boost your discussion score.

Project A 20% - Team Submission

The first team project will be a time series and forecasting problem that you will tackle as a team. A professionally written report will be required via **Microsoft Word** or **Google Docs, not just R Markdown (you may include this separately if desired)**.. Part of your grade will be determined by your peers on your contribution to the submission. Details in Announcements. NOTE: you may turn in your work early, but not late without penalty. There will be a 10 point late penalty PER DAY. So, if you get sick often or other things impair you meeting deadlines you should consider turning them in early.

Project B 30% - Team Submissions, 2X

The second team project will be a team predictive modeling problem. A professionally written report will be required via **Microsoft Word** or **Google Docs, not just R Markdown (you may include this separately if desired)**.. Part of your grade will be determined by your peers on your contribution to the submission. Details in Announcements. This is an end of term project and must be turned in on time, no exceptions.

Required Texts and Materials:

Reading assignments span two primary texts. These are

- Hyndman & Athanasopoulos. "Forecasting: Principles and Practice." https://otexts.com/fpp2/
- Kuhn & Johnson. "Applied Predictive Modeling." http://appliedpredictivemodeling.com/

Optional and Supplemental Reading

- Hastie, Tibshirani, & Friedman. "Elements of Statistical Learning." https://web.stanford.edu/~hastie/Papers/ESLII.pdf
- Burk and Miner, The Foundations of AI, Big Data and Data Science Landscape **for Professionals** in Healthcare, Business and Government (https://www.routledge.com/Its All-Analytics-The-Foundations-of-AI-Big-Data-and-Data-Science/Burk Miner/p/book/9780367359683)

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 **homework** 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.

My Contact Info:

Please address me as “Professor, teacher or instructor or Dr. Burk” during the course/semester. You are encouraged to ask me questions on the “Ask Your Instructor” forum on the course discussion board where other students will be able to benefit from your inquiries. I generally check the forums within 48 hours, but **if you** do not get an answer, please email me. (there will be an additional forum in Blackboard, “Ask the class” **which may** get you a quicker (or better) response).

I am available by email (scott.burk@sps.cuny.edu). We can also set up a call or interactive session if needed. For the most part, you can expect me to respond to questions by email within 24 to 48 hours. If you do not hear back from me within 48 hours of sending an email or have an emergency, you may send a text message to call my mobile phone at 254-563-6909 (text is better and please point out you are part of CUNY – not a marketer ☺). If you don’t text 1st, it will likely go to voicemail and I can call you back.

Course Outline:

This will be a compressed, rapid paced semester, compressing 16 weeks into a 7-week summer

session. This is the planned course schedule, there may be minor modifications required, therefore, be sure to keep up with Blackboard announcements and attend touch points (meetups). We will plan to use Collaborate for our touch points with GoToMeeting as a backup.

You will greatly benefit by discussing and working together with your classmates. The projects are a major part of the class and will require early planning and organization.

Here is a planned schedule for the summer (tentative and modified if needed). Touch Points will meet at 8:00 PM Eastern Time (ET) on Tuesdays as follows:

Date	**Event**	**Forum**
Tuesday, June 1, 2021	Class Starts	BlackBoard
Tuesday, June 8, 2021	First Class meetup	Collaborate
Tuesday, June 15, 2021	Second Class meetup	Collaborate
Tuesday, June 22, 2021	Third Class meetup	Collaborate
Sunday, June 27, 2021	HW #1 due by Midnight ET	Group email, Confirm Receipt
Sunday, June 27, 2021	Project #1 due by Midnight ET	Group email, Confirm Receipt
Tuesday, June 29, 2021	Fourth Class Meetup	Collaborate
Tuesday, July 6, 2021	Fifth Class Meetup	Collaborate
Saturday, July 10, 2021	Discussions CLOSED	BlackBoard Discussion CLOSED
Tuesday, July 13, 2021	Final Class Meetup	Collaborate
Saturday, July 17, 2021	HW #2 and Project # 2 due by Midnight ET	Group email, Confirm Receipt

Schedule

Week	Week of	Topics	Reading	Homework	Due on
1	1-Jun	Welcome and Introductions / Time Series and Decomposition	KJ#1, KJ#2, HA#1, HA#2, HA#6	HA 2.1, 2.3, 6.2	6-Jun
2	7-Jun	Data Pre-Processing and Exponential Smoothing	KJ#3, HA#7	KJ 3.1, 3.2 HA 7.1, 7.2, 7.3	13-Jun
3	14-Jun	ARIMA Models	HA#8	HA 8.1, 8.2, 8.6, 8.8	20-Jun
4	21-Jun	Linear Regression and its Cousins	KJ#6	KJ 6.3 Project 1 and first Batch of HW due	27-Jun
5	28-Jun	Nonlinear Regression Models, Regression Trees and Rules-Based Models	KJ#7 KJ#8	KJ 7.2, 7.5, 8.1, 8.2, 8.3, 8.7	4-Jul
6	5-Jul	Recommender Systems and Case Study	KJ#10 Assigned via announcements	Finish 8-chapter problems, Case Study, Assigned via announcements	11-Jul
7	12-Jul	Project #2 and second batch of homework due		Discussion closes, project #2 due, second batch of HW due	17-Jul

There is more information in Blackboard (under assignments), but here are some comments about Homework:

Completion of exercises must include working R code (that I can cut and paste into R studio along with a discussion of the approach and results). You must NOT turn in JUST code. You will clearly (a big grade component) denote the problem including

1) Replicate / Copy and Paste the problem you are working on. What specific part you are answering, I do not want to go back and refer to the book - explicitly state the question/problem.

2) Tell me what you what approach you are taking to answer the problem. 'We first checked the time series in R for stationarity....."

3) Then the code (in courier so I can copy/paste and run myself if I desire).

4) Code Results

5) Interpretation

Please hand in a Microsoft Word readable document.

If you want to create a professional markdown version that is your option. But you should submit a report as HW, not what you would turn in for a high school class. I am looking for what you would turn into your boss as a professional data scientist.

First Half Semester Homework

There are 11 problems to be submitted in the 1st Batch of HW as follows all KJ and HA.

- Read HA#1, HA#2 (HW HA 2.1 and HA 2.2)
- Read HA#6 (HW HA 6.2)
- Read KJ #3 (HW KJ 3.1 and 3.2)
- Read HA #7 (HW 7.1, 7.2 and 7.3)
- Read HA #8 (HW 8.1, 8.2, 8.6. 8.8)

Second Half Semester Homework

- Read KJ #6 (HW 6.3)
- Read KJ #7 (HW 7.2 and 7.5)
- Read KJ #8 (HW 8.1, 8.2, 8.3 and 8.7)
- Read KJ #10

Recommender Reading Assignment and HW To Be Announced, see announcements Unit Do”

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/