Analytics and the Digital Economy-ADVANCED

OIDD 245, Spring 2018

DRAFT SYLLABUS - Subject to change

Instructor

- **Professor Prasanna (Sonny) Tambe**
- tambe@wharton.upenn.edu, JMHH 558

Teaching Assistants

- To be determined

Office hours

- To be determined

Course Objectives

The goal of this Advanced segment is to further immerse students in the world of data science projects. Specifically, we focus on working with large, unstructured data sources and gain experience with introductory machine learning concepts. Students who take this segment of the course will spend time inside and outside of the classroom combining data and code to develop data products for a number of new industries, including finance, the restaurant industry, and health care.

At the end of the course, students will be expected to complete an advanced data project, which involves acquiring data from an online web property (e.g. Uber, Facebook) through an API and developing an interactive data visualization. Students who complete this course should have the necessary tools to begin building a portfolio of data science projects that they can share online through platforms such as GitHub or with future employers.

Course Overview

Over the last decade, there has been a dramatic rise in the use of tech skills and data analytic thinking to solve business problems in many domains, including finance, HR, policy, and strategy. As a result, the modern "analytic leader" increasingly requires the use of technology, statistics, and data analysis skills to facilitate business analysis. This includes knowing how to a) effectively frame data-driven questions, b) analyze data, and c) use a new generation of tools that are becoming available to acquire, analyze, interpret, and communicate insights derived from data. Students that take this course will engage with the world of data analysis using tools such as Tableau and R that are becoming increasingly popular in industry.

The Intro segment of the course is designed for students with limited experience with data analysis projects, and while familiarity with R, via courses such as STAT 405 or STAT 470, will be ideal preparation, students with other programming exposure can pick up the required skills via review sessions and self-instruction. The second 0.5 CU, Advanced, course will extend students' experience to industry applications of text mining and machine learning and require students to work with more unstructured data. In contrast to the first course, the Advanced module will rely heavily on R and will require the completion of STAT 405, STAT 470, or equivalent preparation.

Throughout the semester, each week of the course will be devoted to analysis of a data set from a particular industry (e.g. HR, sports, fashion, real estate, music, education, politics, restaurants, non-profit work), which we will use to answer business questions by applying analytic techniques. Beyond applications of data tools and methods, a learning goal of this course is exposure to how data is changing decision-making in different industries. The course is *extremely* hands on, and each week focuses on the application of a particular set of tools or analytic methods. Limited time will be devoted to lectures. Most class time will be devoted to supervised work on weekly data projects. Through these exercises, students are expected to become proficient at applying data to business decisions and at effectively analyzing big data sets to inform decisions about business problems using data analysis tools.

Course web site

We will be using Canvas to submit assignments and receive grades. All course information will be posted on the course website. Course communication will be primarily through Slack.

Required textbooks and software

There is no textbook. Occasional readings will consist of selected online content which will be posted on the course site. As part of your homework, you will also be expected to complete some online courses that supplement what we do in class. The majority of the homework requirements involve working on data analysis projects.

Deliverables and grading

During this course, you will be assigned a number of hands on data projects which you will spend time on both in class and out of class. You are expected to participate in classroom discussions (there is more information about participation below). The breakdown of points is as follows:

Data Labs	25%
Individual Homeworks	25%
Final Project	25%
Professionalism + Participation	25%

With each project, you will be provided with a set of guidelines. You can expect to use various data analysis tools extensively, including R and Tableau. We may also, to a limited extent, explore the use of Python and SQL for data analysis/visualization.

In corporate America, you will be expected to present your analytic findings and make a recommendation. Therefore homework deliverables may include short, informal analyses and an accompanying recommendation.

Group projects will be completed in small groups (two to three students, no more than three). You may also be asked to evaluate the contribution of each of your team members after the group project.

The classroom presentation and discussion presents a unique opportunity for you to develop and enhance your confidence and skills in articulating a personal position, sharing your knowledge, and reacting to new ideas. All of you have personal experience that can enhance our understanding of this subject, and we want to encourage you to share that experience.

Participation and Professionalism

This course, like many other courses at Wharton, uses learning methods that require active involvement (e.g. attendance, participation in discussions, and in-class exercises). Not only is this the best way to learn, but it also develops your communication and presentation skills. Regular attendance, participation, presentations, and in general, presenting yourself professionally are all very important, and are an important part of your grade. Active participation requires good preparation—thoughtful completion of homework before class is essential. We recognize that expressing viewpoints in a group is difficult, but it is an important skill for you to develop. We will do what we can to make this as easy as possible. Remember though that only regular and insightful contributions will be rewarded.

The grade we assign for your class participation and attendance is a careful, subjective assessment of the value of your input to classroom learning. We keep careful track of attendance, your contributions towards each class session, and these contributions can include (but are not restricted to) raising questions that make your classmates think, providing imaginative yet relevant analysis of a situation, contributing background or a perspective on a classroom topic that enhances its discussion, providing thoughtful feedback on the presentations of other students, and simply answering questions raised in class. A lack of preparation, missing classes without justification, negative classroom comments, or improper behavior (such as talking to each other, sleeping in the classroom or walking in and out of the class while the lecture is in process) can lower this grade.

Grading Guidelines

At Wharton, we strive to create courses that challenge students intellectually and that meet the Wharton standards of academic excellence. If you believe that an assignment or project grade you received was unjustified, you can appeal the grade. To appeal the grade you must write a one-page explanation as to the reason for your appeal and hand it along with your graded assignment back to the TA responsible for that assignment. Please think twice before appealing a grade: the TA will completely re-grade the assignment, which may increase your grade, but may also lower it (e.g., if the TA catches more mistakes the second time around). If after re-grading you feel that your grade was again unjustified, you can appeal the grade with the instructor.

Overview of Course Schedule for Advanced Module (Q4)

Session	Topic	Date	Due
1	**Datathon 1: in-class challenge**	Mar 12	
2	**Mini-review of regression**	Mar 14	
3	**Applications of text mining**	Mar 19	HW 1: Rats!
4	**Applications of text mining**	Mar 21	
5	**Lab 1B: Yelp reviews**	Mar 26	
6	**Lab 1B: Yelp reviews**	Mar 28	
7	**Interactive data visualization**	Apr 2	Lab 1B
8	**Datathon 2: in-class challenge**	Apr 4	
9	**Applications of machine learning**	Apr 9	HW 2: News analytics
10	**Applications of machine learning**	Apr 11	
11	**Lab 2B: Peer-to-peer lending**	Apr 16	
12	**Lab 2B: Peer-to-peer lending**	Apr 18	
13	**Datathon 3: in-class challenge**	Apr 23	Lab 2B
14	**Final presentations**	Apr 25	Final project

Details of Individual Sessions

Session 1: Datathon 1

- Session Objectives: This session introduces students to "datathons", in which they are provided with a data set and a some time to build a compelling data product. These exercises emulate data-driven consulting engagements, and are becoming an increasingly important assessment tool during interviews. Asking students to apply their skills in such contexts also strengthens their confidence in their data skills. For the first datathon, students are given a a single class period and asked to use Tableau to develop an interesting finding about patterns of use for NYC taxis. Winners receive a small prize.
- Datathon 1 details
- Some examples of recent datathons:
  - *A datathon hosted by MOMA*
  - *A recent open datathon at Cornell Tech*
  - *Citadel is using datathons for recruiting*

Session 2: Mini-review of regression with applications

- Session Objectives: In this session, we discuss applications of how linear and logistic regression are being used in data science settings, with an eye towards building a foundation for a more detailed discussion of machine learning later in the semester. We also review how R can be used to perform statistical analysis. Students are led through a variety of in-class exercises in which they generate correlations and run regressions. It is assumed that students have already taken a statistics course where linear regression is covered, and an in-depth treatment of the statistical foundations of regression is left for other courses.
  - Due:
    - Review the goals of logistic regression
    - Begin Homework 1

Session 3: Applications of text-mining

- Session Objectives: This section introduces the students to applications of text mining. Students do exercises in which they convert unstructured text data (e.g. from 10-K documents, online reviews, etc.) into quantities that can be used to predict business outcomes. In this session, we revisit web scraping and convert the documents we retrieve into a text corpus. We generate a basic word cloud from the scraped data and do other simple manipulations of the text.
- Continue Homework 1

Session 4: Applications of text-mining

- Session Objectives: This session continues the discussion of applications of text mining. We discuss tokenization, stop words, generating features from a text corpus and building predictive models based on features in the text data set. These tools are used to complete several in-class exercises, including analyzing which words in a Reddit post are most predictive of higher levels of user engagement.
- Continue Homework 1

Session 5: Lab 1B (Yelp reviews)

- Session Objectives: In this lab, students apply methods covered in the prior two lectures on text mining. They are asked to analyze about twenty thousand Yelp reviews to determine which establishments are best to visit at different times of day (e.g. breakfast, lunch). The deliverable involves cleaning the data, reporting summary statistics, and developing a model of which restaurants are likely to be best for lunch based on the words that customers use to describe them in online reviews.
- Yelp lab details
- Homework 1 is due

Session 6: Lab 1B (Yelp reviews)

- Session Objectives: In this session, students continue working on Lab 1B in class.
- Begin Yelp lab

Session 7: Interactive data visualization

- Session Objectives: This session extends the visualization module offered earlier in the semester. This session focuses on the use of *Shiny*, which is a package that allows users to offer interactive visualizations. In class, students build a tool that, if given a Twitter hash-tag (e.g. #nfl or #delta), generates a chart tracking real-time sentiment classification for the last 500 tweets that include that hashtag.
- Read/Review:
  - Browse/review Shiny
  - Review pp. 36-48 of ’R for beginners’
  - Yelp lab is due

Session 8: Datathon 2: in-class challenge

- Session Objectives: This datathon asks students to find "blog-worthy" patterns in the data describing outcomes from a large scale speed dating experiment. Aside from helping students build confidence in their new tool sets while facing time pressure, this datathon introduces the idea of how data science requires workers to combine data and statistical acumen with expertise in a particular domain. We discuss how crafting a "good question" requires a different tool set than analyzing and visualizing data. Winners receive a small prize.
- Datathon 2 details

Session 9: Applications of machine learning

- Session Objectives: This session provides students with a brief overview of how businesses are using machine learning for decision making, and introduces them to some basic machine learning applications. They use some of these techniques to generate solutions for some of the entry level challenges on Kaggle.com, such as the Titanic survivor prediction challenge.
- Read:
  - Visual Introduction to ML
  - Sign up for a Kaggle account

Session 10: Applications of machine learning

- Session Objectives: In this session, we cover some additional applications of machine learning. It is likely that it would be focused on hands-on AI based challenges, rather than adding additional lecture concepts.

Session 11: Lab 2B (P2P lending)

- Session Objectives: This lab asks students to apply basic machine learning concepts to a peer-to-peer lending data set. Based on a number of features in the data, they are asked to predict borrowers who are most likely to default. This lab is meant to reinforce some of the concepts covered in sessions 9 and 10.
- Lending Club lab details
- Read:
  - Peer-to-peer lending

Session 12: Lab 2B (P2P lending)

- Session Objectives: In this session, students continue working on Lab 2B.
- Lending Club lab details

Session 13: Datathon 3: in-class challenge

- Session Objectives: In this session, students directly compete in a Kaggle competition. They use any machine learning technique they prefer and that can solve the prediction challenge, and they submit their solutions and receive scores and leader board positions from Kaggle.com. Winners receive a small prize. This lab is meant as a first attempt to engage with the external world of data science and machine learning, and we discuss the role of crowd sourcing in modern data science.
- Datathon 3 details

Session 14: Final presentations + wrapup

- Session Objectives: Students are asked to share their final projects in class. We also discuss how future career paths are likely to incorporate data analytic skills and what students can do to further improve these skills after they leave class.
- Final projects are due