COLLEGE OF MANAGEMENT OF TECHNOLOGY
MANAGEMENT OF TECHNOLOGY SECTION

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Data Science for Business

MGT-432, Fall 2017

Mondays, 9:15 - 12:00

4 credits

Professor: Kenneth A. Younge, PhD
Office: Odyssea 2.02
Phone: 021 693 00 09
E-mail: kenneth.younge@epfl.ch
Office Hours: By Appointment

Post-Doc: Omid Shahmirzadi, PhD
Office: Odyssea 2.17
E-mail: omid.shahmirzadi@epfl.ch
Office Hours: Fridays 10:00-11:00

https://github.com/epfl-tis-class/mgt-432

Updated: August 15, 2017

COURSE OVERVIEW

This course introduces students to some of the methods and tools used by data scientists to address prediction problems applicable to business. Accordingly, the course objectives are three fold: (1) to develop an understanding of how Data Science methods can support decision making in business environments; (2) to gain familiarity with how Data Science tools function through experience in addressing real-word problems and programming real-world solutions; (3) to evaluate the strengths and weaknesses of alternative prediction strategies. The course is particularly applicable for students interested in working for, or learning about, data-driven companies.

PREREQUISITES

Statistics: Prior to the start of class, all students must complete a comprehensive course in statistics covering descriptive statistics, analysis of variance, and the OLS linear regression model.

Programming: Students must have prior experience with at least one programming language, and it is strongly recommended to take an introductory course in computer programming prior to taking this course. In particular, students should familiarize themselves with the syntax and data structures of the Python 3 programming language. There are numerous online MOOCs and/or tutorials to serve this need. A set of Python 3 tutorials for beginners is available at: https://wiki.python.org/moin/BeginnersGuide/NonProgrammers. Our recommendation for beginners is to finish the following 4-hour tutorial by *DataCamp*, which targets data science applications, available at: https://www.datacamp.com/courses/intro-to-python-for-data-science. You should also familiarize yourself with the official Python 3 documentation at: https://docs.python.org/3/

DIDACTIC APPROACH AND CLASS ATTENDANCE

The course will be taught through a combination of readings, lectures, demonstrations, interactive practical sessions, take-home assignments, coding examples, and a semester-long project. Students are expected to attend every class and to read all assigned materials before the start of each class.

Lectures will cover both abstract concepts and programming code to address Data Science problems. Programming code will be posted before each class and available from a GitHub repository. You therefore are encouraged to come to class with a laptop computer so that you can follow along with the examples, and annotate your own copy of the code to better understand it. Tablets and smartphones, however, are not allowed during class time, and laptops should be used exclusively for coding examples, class notes, and reference sites (not general browsing).

MATERIALS

You must purchase, rent, (or perhaps borrow), the following textbook for the course:

> ***Data Science for Business****: What You Need to Know about Data Mining and Data-Analytic Thinking*
> by Foster Provost and Tom Fawcett
> Published by O'Reilly Media. 1st edition (August 19, 2013) 414 pages ISBN-10: 1449361323

You will also be assigned sections from several other books to read (see the detailed course plan in Readme.md in the course git repo), but there are free versions available online. In particular, lectures will reference material in:

> **An Introduction to Statistical Learning: with Applications in R**
> by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshiran
> Published by Springer. 1st ed. 2013, Corr. 6th printing 2016 Edition, ISBN-10: 1461471370

COURSE SCHEDULE

1.	Sept. 25	Welcome		
2.	Oct. 2	Data	Assignment 1	
3.	Oct. 9	Linear Models	Assignment 2	
4.	Oct. 16	Practical 1		Dr. Shahmirzadi
5.	Oct. 23	Model Evaluation	Assignment 3	
6.	Oct. 30	Similarity Models	Assignment 4	
7.	Nov. 6	Practical 2		Dr. Shahmirzadi
8.	Nov. 13	Midterm Exam		
9.	Nov. 20	Decision Trees	Assignment 5	
10.	Nov. 27	Support Vector Machines	Assignment 6	
11.	Dec. 4	Neural Nets		
12.	Dec. 11	Big Data		Dr. Shahmirzadi
13.	Dec. 18	Final Presentations		
--	TBD	FINAL EXAM		

LEARNING OUTCOMES

By the end of the course, students should be able to formulate prediction problems, tackle those problems with different data science methods, and arrive at the best predictive model to empower data-driven decision making and value generation.

GRADING

12%	Assignments	Six assignments worth 2% each
20%	Written Midterm Exam	Administered during a normal class period
25%	Semester Project	Team-based projects (groups of 2)
8%	Final Presentation	In-class presentations
35%	Written Final Exam	Administered during the final exam period

Assignments (12%)

Students will receive a series of 6 assignments, each worth 2% of the overall grade. All assignments must be submitted electronically via a git repository in a jupyter notebook format (i.e. *.ipynb). You will be graded based on the last git commit **and push** that you make before the deadline. No late assignments will be accepted.

Assignments are graded on a **question-by-question basis**, based on a **Check+ (6)**, **Check (5)**, **Check– (4)**, or **No Answer (3)** basis, provided that you submit at least some answer(s) for that assignment. Question-by-question scores are averaged up to a grade for the assignment overall. If no answers for an assignment are submitted at all, then the assignment will receive a grade of zero (0). Thus, do not skip assignments – try to answer at least something so you earn a score of at least 3. At the end of the semester, the score for each assignment will be averaged across all assignments to determine the 12% final contribution to the course grade.

The answer key for each assignment will be distributed in the class repo on the day the assignment is due. Individual corrections/feedback for assignments can be requested during TA office hours. You may work together in teams of up to 3 persons, but each person must submit **their own solutions,** and you must list all team members working together at the top of each assignment. Please name your assignment **`s-#.ipynb`** using the assignment number. For example, your first assignment should be named: **`s-1.ipynb`**

Deadline: You have two weeks to complete each Assignment after it is assigned. Assignments are due **one hour before the start of class** at which it is due. It is important to **git commit** and **git push** your assignment answers before the start of class, as answers are distributed at the start of class and **late assignments will not be accepted**.

Midterm Exam (20%)

There will be a midterm exam partway through the semester. The Midterm will be administered during a normally scheduled class and will cover both conceptual material and programming code up to that point in the course.

Semester Project (25%)

At the start of the 9th session, we will assign students into groups of two to work together on a project. The topic, dataset, and detailed instructions for the project will be provided in class at that time. Participation of both members in the project is mandatory. The TA will create a separate git repository for each team at the start of the project, and each team should compile all of the code, results, figures and comments for your project into a jupyter notebook and commit it to the git repository by the deadline indicated below. You should make absolutely sure that your results are reproducible and that your code runs entirely from a cloned copy of the repository. (In other words, everything required to run your program must reside within the repository, and/or be installed automatically at runtime, and/or be copied automatically into the working directory at runtime, so that your code will execute automatically from beginning to end with no manual reconfiguration by the instructors.)

Grading: We will grade your last committed version. If you continue to make git commits on your project past the deadline, then your grade will be penalized: you will lose 1% (of the 25%) during the first 24 hours; an additional 2% (of the 25%) during the next 24 hours; an additional 3% (of the 25%) during the next 24 hours; and so on. For example, if your final commit to your project comes 52 hours after the deadline, and if you would have received 22% out of the total possible 25% without the penalty, then your final grade would be adjusted by -1% -2% -3%, for a final penalized grade of 16% out of 25%. Notebooks will be graded based on the quality of the logical flow, explanation, code and results. The instructor may use anti plagiarism tools to check the originality of your answers.

Deadline: Projects are due by midnight, one week before the final presentation.

Final Presentation (8%)

Each team should prepare a short presentation of their analysis to be delivered during the last session of the course. We will call teams in random order, and teams will present to the instructors and teaching assistant privately for 5 to 10 minutes (depending on the number of teams – details to be announced), while other teams wait outside. Presentations should not exceed 10 slides. Each presentation will be followed by up to 5 minutes of Q&A. Slides

MUST be presented in (Adobe Acrobat) PDF format (you can export to PDF from either in Keynote or PowerPoint). You MUST git commit your presentation PDF file to your repository BEFORE THE START of the final class, and then bring printed copies of the presentation to the final class for note taking and review by the instructors. Your presentation will be graded based on the quality of your presentation slides, your ability to communicate your findings and the limitations of your analysis, and your ability to field questions and answer correctly.

Final Exam (35%)

The final exam is comprehensive, across the ***entire*** course, and covers all material presented in assigned readings, lectures, demonstrations, interactive practical sessions, take-home assignments, coding examples, and your semester-long project. It will be taken during the final exam period – location and time to be determined by EPFL.

ABOUT YOUR PROFESSOR

Kenneth Younge is an Associate Professor in Technology and Innovation Strategy at the College of Management of Technology (CDM) at the École Polytechnique Fédérale de Lausanne (EPFL). His research examines resource-based strategy, innovation patterns in patent data, and employee mobility between firms – research areas that require the use of methods from Big Data, cloud computing, machine learning, high-dimensional visualization, and Python/Data Science programming. Prior to returning to academia, Professor Younge worked for 14 years in industry as a Director of Development, Chief Technology Officer, and new venture President.

Before joining EPFL, Professor Younge was an Assistant Professor at Purdue University, a post-doctoral scholar at the University of California Berkeley and a doctoral student at the University of Colorado Boulder. He is a past winner of the Academy of Management's Business Policy and Strategy Division Outstanding Dissertation Award, the Strategic Management Society's Best Conference Paper Award, several Distinguished Teacher awards from Purdue University, and the Leeds Outstanding Teaching Award for a Doctoral Student. He graduated *Magna Cum Laude* and Phi Beta Kappa from Brown University, and then began his career as a Strategic Management Consultant with Mercer Management Consulting (now Oliver Wyman). Later in his career went on to found four firms.

ABOUT YOUR TEACHING ASSISTANT

Omid Shahmirzadi is a post-doctoral researcher in the Technology and Innovation Strategy laboratory, under the Chair for Corporate Entrepreneurship at the École Polytechnique Fédérale de Lausanne (EPFL). His research applies big data and machine learning techniques to the analysis of patent data. Omid completed his MA and PhD in computer science with a focus on distributed systems. After graduation, he turned his focus to data science, big data analysis, and statistical methods. He completed a post-doc at the Swiss Institute of Bioinformatics (SIB) where he worked on the statistical modelling of genetic data.

[intentionally left blank]

DETAILED COURSE OUTLINE

1. Welcome

Course Admin: Syllabus, evaluation, structure, etc.
CRISP and the value of data science for business

2. Data

Core concepts
Preprocessing data
Text as data
ASSIGNMENT 1

3. Linear Models

Linear regression
Linear classification
ASSIGNMENT 2

4. Practical 1

Prediction using linear models
DISCUSSION: Over-fitting and under-fitting

5. Model Evaluation

Performance metrics
Model comparison
ASSIGNMENT 3

6. Similarity Based Models

Similarity, distance, and k-NN
k-means clustering
ASSIGNMENT 4

7. Practical 2

Analysis of text documents
DISCUSSION: Supervised vs. unsupervised methods

8. Midterm Exam

Administered during the normal class time

9. Decision Trees

Informative features based on entropy & variance
Classification and regression trees
ASSIGNMENT 5

10. Support Vector Machines

Maximal margin classifiers
Non-linear boundaries using Kernels
ASSIGNMENT 6

11. Neural Networks

Artificial neural networks
Deep learning

12. Big Data

Big data technologies
Distributed machine learning
FINAL PROJECT: due the midnight before class

13. Final Presentations

Timing and details to be communicated in class

FINAL EXAM

Administered during the final exam period in January.
Location and date to be determined later and communicated in class