Harvard University

Data Literacy in the Age of Machine Learning

MGMT E5072
Course Syllabus – Fall 2017

Course Logistics

- Web Conference Component: Monday 5:10-7:10 pm Eastern Time. Specific Dates: August 28, September 11, September 18, September 25, October 2, October 16, October 23, October 30, November 6, November 13, November 20, November 27, December 4, December 11.
- Mandatory On-Campus Component: December 1, 2, 3
  Attendance is required at the entire weekend in order to earn credit and pass the course. You may not arrive late or leave early. If you are traveling from afar, please plan accordingly giving yourself plenty of time to make it to Cambridge by the start of class. No exceptions can be granted.

Instructors and Teaching Assistant

Dr. Mukul Kumar
Chief Innovation Officer,
Hult International Business School
mukul@reinventedu.com

Dr. Jitendra Subramanyam
Data Science Team Leader,
Synaptiq AI
jsub@synaptiq.ai

Mr. Sasha Karimi
Principal Strategist
A&P Capital
sahandskarimi@gmail.com

Office Hours: After class and by appointment

About the Course

Course Description and Overview

Buzzwords like big data, data science, predictive analytics, machine learning, and deep learning seduce and mystify. As a business manager in this age of digital business, you need to know enough about these topics to make good decisions. This course gives you practical knowledge and tools to think creatively about using data and machine learning – in collaboration with your data science team -- to advance your business goals.

The course is divided into four parts.

- *Part 1: The Mechanics of Prediction*. In Part 1 we'll dive right into machine learning, unpacking the key concepts (spoiler: there are just a few and they're simple) and demystify what really happens when machines learn. We'll apply these concepts to make *predictions* from real datasets. We'll cover the basic techniques of machine learning – regression and logistic regression – and get a feel for the practical things that data scientists do.
- *Part 2: The Science of Machine Learning*. In Part 2 we'll learn to systematically evaluate the performance of machine learning models. We'll understand how to define performance and measure it. We'll use this knowledge to not only build the right machine learning models but build them right.
- *Part 3: The Art of Machine Learning*. In Part 3 we'll tackle the art of machine learning – how to get the most predictive bang for our data buck. In other words, we'll learn about how to make the most out of the data we have.
- *Part 4: Select Topics in Machine Learning*. Finally, in Part 4 we'll cover select topics in machine learning: segmenting customers, spotting fraud, detecting spam, and recommending movies.

Key terms demystified (just buzzwords for now!) in this course include: Machine Learning, Descriptive Statistics, Correlation, Predictive Analytics, Regression, Logistic Regression, Non-Linear Regression, Supervised Learning, Unsupervised Learning, Clustering, Bayesian Inference, and Deep Learning.

Prerequisites

We do not use any advanced mathematics in this course. If you've taken the SAT or the GRE you've already come across math that is much more advanced than anything you will need for this course. Alternatively, if you're comfortable working with spreadsheets (nothing fancy, just basic formulas and manipulations like sorting rows), you will be comfortable with all of the mathematics used in this course. Hands-on learning is encouraged using the Orange data science platform (https://orange.biolab.si/) – a visual way to solve machine learning problems without programming. For those with some programming knowledge of Python, we provide Jupyter notebooks that can be used to build, run, and experiment with machine learning models. Please note that Python knowledge is NOT a prerequisite for the course. The course assignments (homework, group presentation, and the final exam) do NOT require any Python programming.

Note: This course is a practical introduction meant to help business executives understand key concepts and techniques in data science and immediately apply them to business problems. It is not for engineering or computer science students seeking to learn the theoretical (and mathematical) underpinnings of machine learning.

Course Format

This course will be taught in a **hybrid model**, with an intensive – and mandatory – three-day residency and the rest of the course conducted through live web conference. Please see dates above, under "Course Logistics."

This is not a traditional lecture-based course. Conceptual material will be illustrated and applied to the "real world" through rigorous class discussion of business cases, examples, group and individual exercises, and your own business and consulting experiences. Your classmates and your instructors expect you to attend and be well prepared for each class, having read the required conceptual material and completed any group and/or individual exercises ahead of time. We also expect you to play an active role in class discussions. If all class members prepare for and actively participate in each class discussion, we will all learn more from each other and enjoy the course more.

Learning Objectives

By the end of this course you will be able to:

- List the types of problems that can be solved using machine learning.
- Understand the seven key steps to solving any machine learning problem.
- Apply machine learning techniques such as regression and classification to solve a variety of business problems using real-world data.
- Build strong intuitions about machine learning techniques by implementing them in a hands-on interactive programming environment.
- Determine efficient and effective ways to improve the results produced by machine learning models.
- Collaborate productively with your data science team.
- Keep up with the rapidly progressing field of machine learning and AI.

Course Materials

This course is taught in a hybrid model and requires students to work continually throughout the semester. It entails a fair amount of reading, working with data, reflection and discussion. Required readings will consist of a variety of blog posts and other articles/videos that can be accessed over the internet. You will also need to download and install the software to run the *Orange* machine learning platform on your local computer (https://orange.biolab.si/). All the materials required for this course are free.

Listed below is an **optional** book for the course. It is NOT required for the course but can serve as an alternative source for learning about the topics we cover in this course. The book can be purchased from many bookstores, including the Harvard Coop and online booksellers. You can also access the book online with your Harvard library credentials.

Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking
Provost, Foster and Fawcett, Tom (O'Reilly Media, Inc.)
ISBN: 978-1-36132-7

Grading

A student's final grade in this course will be based on the following weighting:

15%	Class Participation
40%	Homework Assignments
25%	Final Group Presentation (Intensive Weekend)
20%	Final Exam

Grades reflect the quality of a student's work submitted throughout the term according to the Harvard Extension School's grading standards (http://www.extension.harvard.edu/exams-grades-policies/grades).

This is a graduate-level course and graduate-level work, which includes active participation in class discussions and activities and high-quality written work, is expected. Much of a manager's success depends on communication; therefore effective written and oral communication will constitute a significant portion of a student's grade. Written work should be clear, logical, grammatically correct, spell-checked, persuasive, supported by examples, and backed up by citations for any data, ideas or other content used. It should represent the student's best effort.

Please note that all homework assignments are due in the relevant course assignment folder (on the course Canvas website) at the indicated time. Late assignments will be penalized significantly.

Assignments

Homework Assignments (approximately every other week)

These assignments will vary in length and content. Their objective is always the same though – to ensure that you get a working knowledge of the material covered and enable you to explore topics that are not covered explicitly during class.

Data Exploration and Machine Learning Application Project (Intensive Weekend)

Across the intensive weekend you will have an opportunity to learn and practice key data science skills like problem framing, data gathering, visualization, modeling, and communicating recommendations. The weekend will conclude with each team presenting the results of their data science analysis. These recommendations will be in the form of a short presentation supported by slides.

This assignment is designed to allow your group to practice data science skills. We will provide you with datasets from which you can choose. Alternatively, you can work on a dataset of your choice; please let your instructors know what you choose and get their feedback before proceeding.

We urge you to choose something that you find interesting and that will be helpful to your professional career and/or your personal skills. You will be graded on the quality and depth of your approach, the logic underlying your conclusions, and the clarity and professionalism of the presentation.

The presentation can be no longer than 10-minutes (using PPT slides), followed by Q&A and feedback. There is no limit on the number of slides you use. You must also include a bibliography. We recommend sharing as Appendix slides any intermediate work product that does not make it into the final presentation.

You must select your topic for this assignment and submit a brief (no more than 1 page) written proposal by October 30, 2017. Please don't hesitate to contact your instructors to discuss potential ideas for datasets.

Class Participation

Even if you are convinced about the business return on a data science project, you will often be in the position of having to "sell" it to your business colleagues and bosses to get their acceptance and support. In this course, the classroom provides a laboratory in which you can test your ability to convince your peers of the appropriateness of your approach to data science projects. Furthermore, it tests your ability to carefully listen to others' perspectives and understand why they may reach a different conclusion. Before you can effectively sell your ideas to others, you must understand what is motivating them, what issues they feel are important, and what assumptions they are making that may be different from your own.

When evaluating your contribution to the class, then, we will consider how effectively you put forth your own arguments, as well as how well you listen to, understand, and build upon (or refute) the arguments of others. We will look for high quality (which is not always the same as high quantity) arguments, analyses and questions. While we encourage you to speak up at any time, keep in mind that comments that are redundant, tangential or seemingly irrelevant to the discussion at hand will have a negative impact on your class contribution grade.

You may miss one online class session without penalty, but all other absences will result in a negative score for class participation for that session. **Participation in the December 1, 2, 3 on-campus sessions in their entirety is mandatory, and students may not be late or leave early for any of these sessions.** ***Failure to be in attendance for the entirety of the on-campus session will result in removal from the class.***

Academic Integrity

You are responsible for understanding Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid- plagiarism), where you'll find links to the Harvard Guide to Using Sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Additional Information

Workload. The value you receive from this course will be commensurate with the thought and effort that you put into the endeavor. Students should expect to spend 4-8 hours outside of class each week to read the assigned materials, reflect, complete assignments, and prepare for the next class session. More time will be required to do the team project.

On Time. Students are expected to arrive to the online classroom on time and stay for the duration of the class session. If you expect to be late or absent from class – or need to leave early – let the instructor know prior to the start of class.

Deadlines. All assignments must be submitted to the correct assignment drop box on the Canvas course website by the specified day and time and **late submissions will not be accepted**. If you experience any problems uploading your assignment to the Canvas drop box, you should email the document to the instructor. Please note that, if you email the assignment because you cannot upload it, the email and the relevant attachment *must be received on or before the assignment deadline to be accepted*. Should you experience any internet problems, please call/leave a message for the instructor – this call should occur before the submission deadline passes. If you are absent the day an assignment is due, the assignment is still due at the specified day and time. True medical or family emergencies will be dealt with on a case-by-case basis.

Professional Conduct. Professional behavior is expected throughout the class. This means respectful communication both inside and outside of class. During discussions, civil discourse should be maintained at all times and comments should be aimed at moving the discussion forward. This does not mean that students must always agree with others since reasoned, respectful dissention may be part of the discovery process and lead to previously unconsidered options.

Disability Services: The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services- accessibility for more information, or contact the Accessibility Services office at Accessibility@dcemail.harvard.edu or (617) 495-4024

Course Outline and Schedule

Session #	Date	Topic Area / Theme	Read and Do BEFORE the Class Session	Deliverables | Assignments Due | Exams
1	August 28	**Introduction to Data Science and Machine Learning** ▪ The current state of machine learning and where it might be headed ▪ Overview of the course ▪ Visualizing data systematically	▪ Complete pre-course survey (in Canvas) ▪ Make sure that you have a working headset for the online sessions ▪ Read through course syllabus in its entirety ▪ View Frank Chen's video, "The Promise of Machine Learning" (https://vimeo.com/215926017) ▪ View John Launchburg's video, A DARPA Perspective on AI (https://www.youtube.com/watch?time_continue=5&v=-O01G3tSYpU)	**Pre-Course Survey** **(Due before August 25)**
2	September 11	**Hands On - The Nuts and Bolts of Machine Learning** ▪ How machine learning really works ▪ How it compares with other approaches	▪ Download and start up Orange (https://orange.biolab.si/download/) ▪ View the first 3 Orange training videos at https://www.youtube.com/watch?v=HXjnDIgGDuI&list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy ▪ Ensure that you have access to the Jupyter notebooks on GitHub ▪ Read the lecture notes and the Jupyter notebook on systematic visualization.	
3	September 18	***Predicting Numerical Values 1*** ▪ Regression with a single feature	▪ Read the HBR article, The Simple Economics of Machine Intelligence (https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence)	**Homework Assignment 1** **(Due at 10am on September 18)**
4	September 25	***Predicting Numerical Values 2*** ▪ Regression with multiple features ▪ Non-linear regression	▪ Read the lecture notes on single-feature regression. ▪ Run a regression using Orange and a dataset of your choice.	
5	October 2	***Predicting Categorical Values 1*** ▪ Logistic regression with two features ▪ Logistic regression with multiple features	▪ Read the lecture notes on regression with multiple features. ▪ Read about the key types of data science projects to work on (https://www.dataquest.io/blog/build-a-data-science-portfolio/)	**Homework Assignment 2** **(Due at 10am on October 2)**
6	October 16	***Predicting Categorical Values 2*** ▪ Non-linear logistic regression ▪ Particularly useful models for non-linear	▪ Read the lecture notes on logistic regression. ▪ Skim the following dataset repositories to get ideas for your team project:	**Homework Assignment 3** **(Due at 10am on October 16)**

		logistic regression – Support Vector Machines and Neural Networks	▪ https://github.com/caesar0301/awesome-public-datasets ▪ http://archive.ics.uci.edu/ml/datasets.html ▪ Start coordinating with your team.	
7	October 23	***The Science of Machine Learning 1*** ▪ Models, parameters and hyper-parameters ▪ Training, validation and test datasets ▪ K-fold cross validation ▪ Validation curves	▪ Read the lecture notes on non-linear logistic regression. ▪ Read Marc Andreesson's article, "This is Probably a Good Time to Tell You…" (http://blog.pmarca.com/2014/06/13/this-is-probably-a-good-time-to-say-that-i-dont-believe-robots-will-eat-all-the-jobs/) ▪ Read Kevin Kelly's article, "The Myth of Superhuman AI" (https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/)	**Homework Assignment 4** **(Due at 10am on October 23)**
8	October 30	***The Science of Machine Learning*** 2 ▪ Measuring model bias and variance ▪ Learning curves ▪ Measuring model performance	▪ Read the lecture notes on the science of machine learning, part 1. ▪ Peruse the list of AI tools for businesses available today (https://medium.com/imlyra/a-list-of-artificial-intelligence-tools-you-can-use-today-for-businesses-2-3-eea3ac374835). It might give you some ideas for your group project.	**Homework Assignment 5** **(Due at 10am on October 30)**
9	November 6	▪ Open time for each team to review progress on final project. Opportunity for early presentation feedback.	▪ Read the lecture notes on the science of machine learning, part 2. ▪ Select a dataset or two in coordination with your team. ▪ Determine the business problem(s) you want to investigate. ▪ Get a sense of how you'll use the data to solve the business problem(s).	
10	November 13	***The Art of Machine Learning*** ▪ Feature Engineering: Feature selection/extraction/representation ▪ Data compression ▪ Decision trees and random forests ▪ Ensembles and aggregation of results	▪ View and experiment with neural networks on the Tensorflow Playground (http://playground.tensorflow.org)	**Homework Assignment 6** **(Due at 10 am on November 13)**
11	November 20	***Select Topics in Machine Learning 1*** • Similarity and clustering • Profiling and anomaly detection	▪ Read the lecture notes on feature engineering. ▪ Read David Brook's article, "What Data Can't Do" (http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html) ▪ Read Brynjolffson and McAfee's article, "The Business of Artificial Intelligence" (https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence)	**Homework Assignment 7** **(Due at 10 am on November 20)**
12	November 27	***Select Topics in Machine Learning 2*** • Learning from Text	▪ Read the lecture notes on similarity and clustering. ▪ Read Richard Weiss' article, "Cargo Cult of Data Science" (http://blog.richardweiss.org/2017/07/25/data-science-in-organizations.html) ▪ Read Michelle Nijhuis' article, "How to Call BS on Big Data" (http://www.newyorker.com/tech/elements/how-to-call-bullshit-on-big-data-a-practical-guide)	

13	December 1, 2, 3 On Campus Weekend	**Final Project Working Sessions and Presentations** ***Select Topics in Machine Learning 3*** • Large scale machine learning • Building a machine learning system • Deep Learning	▪ Coordinate with your team ▪ Create your presentations ▪ Come prepared to present!	**Final Project Presentations**
14	December 4	***Select Topics in Machine Learning 4*** • Recommender systems	▪ Read the lecture notes on learning from text. ▪ View Frank Chen's video, "AI, Deep Learning, and Machine Learning" (https://vimeo.com/170189199) ▪ View Andrew Ng's video, "Nuts and Bolts of Applying Deep Learning" (https://www.youtube.com/watch?v=F1ka6a13S9I)	
15	Week of December 11	• Review and Wrap Up • Final Exam		**Final Exam**