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](mailto:mukul@reinventedu.com?subject=MGMT%20E5072%20Question) | Dr. Jitendra Subramanyam  Data Science Team Leader, *Synaptiq AI*  [jsub@synaptiq.ai](mailto:jsub@synaptiq.ai?subject=MGMT%20E5072%20Question) | Mr. Sasha Karimi |

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 abreast of 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. Listed below is the required book for the course – it can be purchased from many bookstores, including the Harvard Coop and online booksellers:

**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

In addition to the book we will use a variety of blog posts and other articles that can be accessed over the internet. You will also need to download and install some software on your local computer. These are indicated for each session on Canvas. Except for the book listed above, all other course materials are free.

***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-*](http://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-*](http://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-*](http://www.extension.harvard.edu/resources-policies/resources/disability-services-) *accessibility for more information, or contact* the Accessibility Services office at [Accessibility@dcemail.harvard.edu](mailto:Accessibility@tufts.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)> * Read the Information World article on what the top 5 companies are doing with machine learning today. (http://www.informationweek.com/elite100-2016.asp) | **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 | * [Read Provost and Fawcett, Chapter 1] * 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. | **Homework Assignment 1** **(Due at 10am on September 11)**  Visualize the telecom churn dataset on Orange. Write down a list of 5 visualizations you created. For each visualization, describe what you found. |
| 3 | September 18 | ***Predicting Numerical Values 1***   * Regression with a single feature | * [Read Provost and Fawcett, Chapter 2] |  |
| 4 | September 25 | ***Predicting Numerical Values 2***   * Regression with multiple features * Non-linear regression | * Read the Jupyter notebooks on single- feature regression. * Run a regression using Orange and a dataset of your choice. | **Homework Assignment 2**  **(Due at 10am on September 25)**  Use Orange to perform a single-variable regression on the restaurant profit data set. Alter the size of the data set and write down your observations on how predictions change. |
| 5 | October 2 | ***The Science of Machine Learning 1***   * Models, parameters and hyper-parameters * Training, validation and test datasets * K-fold cross validation * Validation curves | * Read the McKinsey article “The Simple Economics of Machine Learning” |  |
| 6 | October 16 | ***The Science of Machine Learning*** 2   * Measuring model bias and variance * Learning curves * Measuring model performance | * Read Provost and Fawcett, Chapter 7: Decision Analytic Thinking I: What is a Good Model? * Read the Jupyter notebook on logistic regression with 2 features. | **Homework Assignment 3** **(Due at 10am on October 16)**  Multivariate Regression. Use Orange to perform a regression on the house price data set. Describe your results. |
| 7 | October 23 | ***Predicting Categorical Values 1***   * Logistic regression with two features * Logistic regression with multiple features * Measuring the performance of logistic regression models | * Read the Jupyter notebook on logistic regression with 3 or more features. |  |
| 8 | October 30 | * Open time for each team to review progress on final project. Opportunity for early presentation feedback. | * Read Provost and Fawcett, Chapter 5: Overfitting and its Avoidance * Read the Jupyter notebook on Support Vector Machines * Read the Jupyter notebook on Neural Networks | **Homework Assignment 4** **(Due at 10 am on October 30)**  Multivariate Logistic Regression. Use Orange to perform a logistic regression on the entrance exam scores data set. |
| 9 | November 6 | ***Predicting Categorical Values 2***   * Non-linear logistic regression * Particularly useful models for non-linear logistic regression – Support Vector Machines and Neural Networks | * Read Provost and Fawcett, Chapter 8: Visualizing Model Performance | **Homework Assignment 5** **(Due at 10am on November 6)**  Submit your Final Project Proposal |
| 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 | * Read Provost and Fawcett, Chapter 11: Decision Analytic Thinking II: Toward Analytical Engineering | **Homework Assignment 6** **(Due at 10 am on November 13)**  Submit your answers to the quiz on measuring model performance. |
| 11 | November 20 | ***Select Topics in Machine Learning 1***   * Similarity and clustering * Profiling and anomaly detection | * Read the Jupyter notebook on Feature Selection * Read Provost and Fawcett, Chapter 6: Similarity, Neighbors, Clusters |  |
| 12 | November 27 | ***Select Topics in Machine Learning 2***   * Learning from Text | * Read Provost and Fawcett, Chapter 10: Representing and Mining Text | **Homework Assignment 7** **(Due at 10 am on November 27)**  Use Orange to cluster and visualize the whiskey dataset. |
| 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 * Choose your dataset(s) * Choose your methods of analysis * Create a “shell” presentation | **Final Project Presentations** |
| 14 | December 4 | ***Select Topics in Machine Learning 4***   * Recommender systems | * Read Edwin Chen’s “Winning the Netflix Prize: A Summary” (<http://blog.echen.me/2011/10/24/winning-the-netflix-prize-a-summary/)> | **Homework Assignment 8** **(Due at 10 am on December 4)**  Use Orange to classify the email dataset into spam and ham. |
| 15 | Week of December 11 | Review  Final Exam |  | **Final Exam** |

Some comments.

1. Syllabus:

- Dates are also listed at top of page 1 of syllabus. Need to fix.

- On campus weekend was scheduled for Dec 1-3. Have you spoken to Harvard admin about changing it? They have to find the space first.

- Did you want a final exam? If so we need to include it in weighting of grade.

2. Homeworks.

- Like the telco data set

- Did not find written instructions for homework beyond the one line in the syllabus. Did I miss the location?

3. Lecture Notes

- Like the first lecture. Have some thoughts on integrating with you for this first lecture. Lets discuss.

- I still it would be valuable to setup all the approaches you plan to cover across the semester in the first session.

- Regression lecture makes sense. Might make sense to add more notes for the example you will demo as well.

- What connection do you want to make with Prevost and Fawcett in your lecture notes? Helps them to see some return from having done the reading.

4. TA

Sasha can likely be helpful in building out the Homeworks and Final, if you need. I can join for part of your session with him on Tuesday to get us all aligned.

5. Pre-Course Survey & Other canvas materials -- on my list for early this week.