# Course Syllabus: Data Mining for Business

CSCI S-96

**Revised 7-15-2018**

Harvard Extension School Summer 2018 Session II

Dates: July 16-August 2, 2018

Time: Mon-Thurs; 6:30-9:30pm

Building: Maxwell Dworkin Room: G125

Instructor: Ted Kwartler, MBA

Email: [ehk116@gmail.com](mailto:ehk116@gmail.com); 330-780-5037

Office Hrs: Available upon request

Piazza Forum URL: [**https://piazza.com/extension.harvard/summer2018/cscis9634203/home**](https://piazza.com/extension.harvard/summer2018/cscis9634203/home)

## Prerequisites:

* Textbook: Data Mining for Business Analytics: Concepts, Techniques, and Applications in R

ISBN-10: 1118879368

Harvard Coop Bookstore link for the book: [**https://tinyurl.com/A18-CSCI-S-96-1**](https://tinyurl.com/A18-CSCI-S-96-1)

* Software: R & R-Studio

## Course Learning Objectives:

If you stay engaged in the course and complete the suggested readings and assignments:

You will be able to think systematically about how data is used to make business decisions. This objective will be accomplished through the use of ideas from statistics, economics and computer technology and using business related case studies.

Students will learn how to implement a variety of popular data mining algorithms in R (a free and open-source software) to tackle business problems and identify opportunities. This course will help introduce the basics of R in data mining.

**As a business leader, you will acquire the skill of applying data science concepts within business domains to improve decisions and learn how data scientists approach projects.**

## Attendance:

Regular attendance is essential to the successful completion of this course. Attendance will be taken regularly and you are responsible for material covered in class whether or not you have attended class. Given the short summer session missing more than 1 class session for any reason will result in an automatic reduction in course grade. Unsatisfactory attendance may result in a failing grade. You should plan on spending at least three hours of independent study for each hour of class attendance.

## Code of conduct:

This course expects you to uphold and report violations of the Extension School code of conduct found [here](https://www.extension.harvard.edu/resources-policies/student-conduct).

You are responsible for understanding Harvard Extension School policies on academic integrity ([www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity](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-plagiarism](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.  
  
Accessibility  
The Extension School is committed to providing an accessible academic community. The Disability Services 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](http://www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility) for more information.

## Grading:

A course grade will be assigned on the basis of student performance on two examinations, coding assignments, a written assignment, attendance and participation and group work.

* Class presentation by group 15% of final grade
* Attendance 10% of final grade
* Midterm Exam 25% of final grade
* Final Exam 25% of final grade
* Written assignment 10% of final grade
* Coding Assignments 15% of final grade

## Writing Assignment

Ten percent of the final grade will be determined by the quality and completeness of a 900 to 1500 word essay concerning business implications and ethics of data mining. Approximately, no more than 25% of the essay should comprise a summary and synthesis of the assigned data science ethics articles. The balance of the essay can incorporate new literary sources and/or student reflections for how business is affected by the rise of cheap computing, large scale creation and storage of data and development of new algorithms. Example questions to spur creative reflection include (but are not limited to):

* Is it ok to have a “black box” algorithm where users do not know how it functions?
* Is there an ethical duty to tell users you are collecting information and reselling it or simply bury it in a terms of service agreement? Does anyone really read the agreements?
* Are algorithmic traders crowding out less sophisticated retail investors? Does the market have a duty to train others, disclose code or report market manipulation?

## Group Case Presentations

The cohort will be broken up into groups of ~4 each working on a business case that will use data to affect the outcome. Each group will have ~15 minutes for presentations including limited time allotted for optional q/a. Every individual in a group is expected to present a portion of the group’s effort. Presentations will be graded on their use of data, code demonstration (if applicable), strategic business thinking, qualitative understanding of the business objective and overall persuasiveness/presentation skills.

Groups may choose different cases, or expand on the problem presented below.

* The head of okcupid.com’s marketing department wants to identify novel customer segments. As part of the business intelligence team, your group has been asked to review user profiles and identify 4 or more interesting insights from user profile data. Your team will present exploratory data analysis findings to the head of marketing.
* As part of a world-wide capacity planning team you have been asked to forecast customer demand for 4 products over 12 months. You will present your findings to the head of sales which is skeptical of analytical forecasting techniques. As a result, presentation must earn trust by explaining how a forecasting technique works in the context of the products the team chooses as well as the product forecasts themselves.
* At National City Bank you are part of the research and modeling team. The bank has come up with a new loan product and wants it to be a success to drive non-traditional revenue. As a new pilot project, the bank is offering the product only to a select group of customers. Your team has been asked to model customer propensity to improve the pilot’s acceptance rate.

## Classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | DOW | Topics Covered | Reading  Due | Assignments  Due |
| July 16 | Mon | * Introduction to R * Prediction & Classification * Supervised & Unsupervised * R Visualizations | NA |  |
| July 17 | Tues | * R Data Structures * Data Mining workflow * EDA * Feature Cleanup, engineering & enrichment | Chapter 1  Chapter 2 | 1 . Day1\_Homework\_v2.R  2. C2.1 Data Mining Techniques  3. C2.2 Data Partition  4. C2.3 Data Sample  5. C2.4 Modeling Steps  6. C2.5 Overfitting  7. C2.6 Data Leakage  8. C2.11 ToyotaCorolla.csv *only “a.”* |
| July 18 | Wed | * Regressions for Predictions & Binary Classification * Basic model evaluation | Chapter 6  Chapter 10 | 4. day2\_airfareEDA.R  5. day2\_diabetesPrep.R |
| July 19 | Thurs | * KNN * Trees & Random Forests | Chapter 7  Chapter 9 | 6. C6.1 Predicting Boston Housing Prices  7. C6.2 Predicting Software Reselling Profits  8. C6.3 Predicting Airfare on routes  9. C10.4 Competitive Auctions on ebay |
|  |  |  |  |  |
| July 23 | Mon | * Time Series Forecasting | Chapter 16  Chapter 17  Chapter 18 | 10. C7.1 Calculating Distance  11. C7.2 Personal Loan Acceptance  12. C9.3 Predicting Prices of Used Cars |
| July 24 | Tues | * Association Rules * Collaborative Filtering * Ethics | Ethics Articles  Chapter 14 | 13. C17.3 Toys R Us Revenue  14. C16.1 Impact of 9/11 on Air Travel  Sales customers case  15. C17.1 9/11 Impact pt2  16. C18.1 9/11 Impact pt2  17. C18.9 Australia Wine Sales |
| July 25 | Wed | * Securities * Financial Risk Modeling * Non Traditional Markets |  | C14.1  C14.4 |
| July 26 | Thurs | Mid-Term Exam |  | 19. Turn in a script using TTR for any stock’s SMA, MACD, historical returns & make a buy/sell recommendation at current prices.  20. Prepare & Score new lending club loans in file  primaryMarketNotes\_browseNotes\_July16.csv  Turn in the script used for loading, preparing and scoring the new loans. Recommend 1 note from grade A and another for B.  21. Perform a non-traditional mkt analysis on Magic’s Core 2019 set. What is the avg return? Turn in the script with avg return for 100 simulated boxes. |
|  |  |  |  |  |
| July 30 | Mon | * Text Mining Basics * Report automation | Chapter 20 |  |
| July 31 | Tues | * Cluster Analysis * Guest Speaker, TBD * Group Case Work w/professor consultancy as needed | Chapter 15 | 22. .Using sampled AirBnB Reviews calculate polarity on each comment then divide the comments into positive and negative reviews. Perform a frequency analysis on each of the corpora to understand the features renters associate with good and bad stays. |
| Aug 1 | Wed | * Group presentations | Various articles | 24. C15.1 University Rankings  25. C15.4 Mktg Freq Fliers |
| Aug 2 | Thurs | Comprehensive Final Exam |  | Writing Assignment |