Course Syllabus: Data Mining for Business

CSCI E-96
Revised 9-9-2018; Final

Harvard Extension Fall 2018
Dates: Sept 10-Dec 17, 2018
Time: Mon 8-10pm
Building: Maxwell-Dworkin G115
Instructor: Ted Kwartler, MBA
Email/Phone: ehk116@gmail.com; 330-780-5037
Office Hrs: Available upon request

Important URLs:

Piazza URLs:
https://piazza.com/class/jlr83pl8h293n
http://piazza.com/extension.harvard/fall2018/cscie9615736
Access Code (if needed):
dataMiningHarvard

Canvas URL:
https://canvas.harvard.edu/courses/53027

Streaming Information:
Maxwell Dworkin G115 does not use Zoom to stream lectures, rather it is streamed via Opencast. Students will be able to access the live stream via a link on the Lecture Video page, which will be posted by Monday morning and becomes live a minute before the start of class. The lecture video will then be posted to the same page within 24 hours of the lecture start time.
https://canvas.harvard.edu/courses/53027/external_tools/22940
Week 1 lecture will also be posted here:
https://matterhorn.dce.harvard.edu/engage/ui/index.html#/2019/01/15736

Github Repo:
https://github.com/kwartler/HarvardFallStudent2018

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/300-CSCI-E-96-F18-1

- Software: R & R-Studio

- Access to git software to download data sets and class material or ability to download directly from the Internet
- A webcam or other method to record case presentations & upload to the University's approved site
- Be prepared to obtain a free zoom account as each group will need a single zoom participant to record case presentations

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.

As a data scientist, you will acquire practical applications of data mining methods that are used in many of today's most successful organizations as well as being to understand what business stakeholders expect of data scientists.

Attendance:

Regular attendance and remote participation on the class forum is essential to the successful completion of this course. Attendance will be taken regularly for on campus sessions and forum participation will be monitored for remote participants. You are responsible for material covered in class even if you have not attended class or watched the recorded lectures. Given the amount of information covered, missing more than 1 class session for any reason may 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. Further, all assignments are the responsibility of each *individual* pupil unless assigned as a group assignment. Utilizing the class forum, online resources, teaching assistants, and the class professor to ask questions is (of course) acceptable but copying another peer's work is considered a violation of the University code of conduct.

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.

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 for more information.

Grading:

A course grade will be assigned on the basis of student performance on examinations, homework assignments, a written assignment, attendance and participation and group work. Remote students will take their final exam online which proctors through a webcam. More details will be shared during class. On campus pupils will attend a class session for in person proctoring.

No late homework will be accepted under ANY circumstances. Failure to submit submissions through the University approved portal by the assignment deadline will be considered late and not accepted. Submissions to any other location will not be accepted. During exams, no phones, tablets or computers should be used even as calculators. If you need a calculator you must bring one to your examination period. A student may prepare a single, double sided 3inch by 5inch, *handwritten* index card for use during any examination. Cards that are larger, typed or multiple cards will constitute cheating according to Harvard's academic integrity policies.

- Class participation, attendance, and online forum participation 10% of final grade
- Case I 15% of final grade
- Case II 20% of final grade
- Final Exam 20% of final grade
- Written assignment 15% of final grade
- Homework Assignments 20% of final grade

Writing Assignment

Fifteen percent of the final grade will be determined by the quality and completeness of a 900 to 1200 word ***essay concerning ethical implications of data mining within a business context***. 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 based on open source licenses or report market manipulation?

While defining an ethical framework can be a personal matter, the organization and robustness of your argument along with supporting statements to the argument are subject to evaluation. It is not the case that all ethical actions are relative or that ethical considerations are incapable of objective evaluation. Further the level of sophistication you demonstrate in understanding the issue discussed, addressing applicable opposing viewpoints and the logical structure of your tenets will impact your grade. Lastly, primary source philosophical paradigms, not mere opinions should be used as a foundation for your logical construction of what is ethical in a data mining and business context.

Each page should have a header with a clear label including the author, date, page number and title. As a personal reflection paper concerning ethics, APA or similar citation method is *not* necessary.

Group Case Presentations

The cohort will be broken up into groups of ~4. Each group will working on a two business cases that use data to affect the outcome. Each group will create and upload verbal presentations for review and grading. During the recorded presentation, each 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, succinctness, persuasiveness, qualitative understanding of the business objective, and overall presentation skills. Each group presentation is to be no more than 15minutes in length. All supporting material including scripts, visuals and or presentation slides will also need to be turned in for review.

Classes

Date	8-9pm	9-10pm	Reading Due	Assignments Due
9-10	Introduction & Administrative	Intro to Data Mining	NA	
9-17* *Taught by the teaching assistant.*	Intro to R		Chapter 1 Chapter 2	1. Piazza introduction post 2. C2.1 Data Mining Techniques 3. C2.2 Data Partition 4. C2.3 Data Sample 5. C2.4 Modeling Steps
9-24	Data Mining in a Business Workflow	Data Preprocessing Donor Bureau Case	Chapter 3	6. C2.5 Overfitting 7. C2.6 Data Leakage 8. C2.11 ToyotaCorolla.csv *only "a."* 9. Day2_Homework_v2.R

10-1	Regression	Logistic Regression	Chapter 6 Chapter 10	**CASE I. OK Cupid Case Upload**
10-8	HES Holiday Indigenous Peoples' Day			
10-15	Model Evaluation	KNN	Chapter 7	10. C6.1 Predicting Boston Housing Prices 11. C6.2 Predicting Software Reselling Profits 12. C6.3 Predicting Airfare on routes 13. C10.4 Competitive Auctions on ebay
10-22	Decision Tree	Random Forest	Chapter 9	14. C7.1 Calculating Distance 15. C7.2 Personal Loan Acceptance
10-29	Time Series Forecasting	Equity Trading	Chapter 16 Chapter 17	16. C9.3 Predicting Prices of Used Cars
11-5	Financial Risk Modeling	Non-Traditional Investment Modeling • Possible Guest Speaker TBD	Chapter 18	17. C17.3 Toys R Us Revenue 18. C16.1 Impact of 9/11 on Air Travel Sales customers case 19. C17.1 9/11 Impact pt2
11-12	Data Sources with R - APIs	Reporting Automation	NA	20. C18.9 Australia Wine Sales 21. Turn in a script using TTR for any stock not covered in class with SMA, MACD, historical returns & make a buy/sell recommendation at current prices. 22. 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 for investment.
11-19	Collaborative Filtering	Association Rules	Chapter 14	Using an API: 23. Create a script to construct a powerpoint with lib(officer) 24. Create a script to construct a flexdashboard
11-26	Text Mining	Text Mining	Chapter 20	25. C14.1 26. C14.4
12-3	Ethical considerations of Data mining	Ethical considerations of Data mining	Ethics Articles	27. .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.
12-10	Guest Speakers		NA	**CASE II. Banking Case Upload**

	• Greg Cochara, VP Fusion Media Group (The Onion, Clickhole etc.) • Victor Arias, Senior Account Supervisor, Edelman (Advertising) • Ross Leav, Snr Dir, Presidio Ventures (Venture Capital) - *possibly*		
12-17	Comprehensive Final	NA	**Writing Assignment**

Graduate Credit Students

This course is open to non-credit, graduate and undergraduate students. As a result, the course experience will vary for each cohort.

Noncredit students may submit presentations, homework, and the ethics paper. Your assignments will receive feedback to improve your acumen. However noncredit student may not take exams or receive letter grades.

Graduate credit students are expected to do more work and perform at higher standards than undergraduate credit students. On the midterm and the final, there will be additional knowledge tested for graduate credit students. These may include but are not limited to additional multiple choice questions, short form answers or coding sections. Further, a graduate credit student's ethics paper should incorporate an additional 3 sources of information beyond the assigned reading.

Grading Scale

You earn the grade based on assignments according to the scale below. Grades are not curved to fit a predetermined distribution. A student's degree or certificate candidacy will not have any impact on a course grade. Note there are no "minus" grades given in the course. It is the belief of the instructor that minus grades constitute a false precision in many academic courses and further penalize frequent "A-" students since there is no way to obtain an "A+" to rebalance a GPA.

Max	Min	Grade
100	90	A
89.9	87	B+
86.9	80	B
79.9	77	C+
76.9	70	C
69.9	67	D+
66.9	60	D
59.9	0	F