# Course Syllabus: Data Mining for Business – Revised July 18, 2021

CSCI E-96 – *although this course includes business topics at its heart it is a computer science course. Please be mindful of this. Please take the skills assessment (ungraded) to ensure you have a basic knowledge foundation.*

Harvard Extension Fall 2021

Dates: Aug 30, 2021 – Dec 18, 2021

Time: Monday 8:10-10:10pm

Building: NA, Remote only.

Instructor: Ted Kwartler, MBA

Email:

[edwardkwartler@fas.harvard.edu](mailto:edwardkwartler@fas.harvard.edu)

Office Hrs: Available upon request

Optional Lab: TBD

## Important URLs:

**Canvas** (homework submissions and grading)  
TBD

**Class Question Forum TBD** (Piazza or similar)

TBD

This term we will be using an online forum for class discussion. The enables fast and efficient help from classmates and your teaching staff. Rather than emailing questions to the teaching staff directly, I encourage you to post your questions on the forum to benefit everyone when appropriate.

Keep in mind, the class forum is used to benefit distance learning, but it is never intended to be a customer service center, social networking website or the channel for sharing evaluations of teammates, staff or topics (save that for your course evaluations). Further, teaching staff responses are not expected within any given timeframe. In fact, your teaching staff may not respond at all for some topics so that students have to help each other, research and explore on their own and ultimately learn rather than be explicitly told. While some students find this problematic the course goal is to facilitate self-learning and improve confidence as much as direct instruction.

**Github** repository allows you to get all scripts, PowerPoints and data sets throughout the semester. For those not familiar with github, think of it like a shared drive similar to SharePoint or Dropbox but with added functionality for data and computer science.

*The following link has underscores not spaces (“\_”)*

<https://github.com/kwartler/Harvard_DataMining_Business_Student>

## Streaming & Video Information:

* Lectures will be streamed via zoom, with link in the canvas site.
* Students will be able to access the recordings to watch on demand. Recordings are usually posted within 24-48hours after the lecture.

**All lecture video links will be available from the Course Canvas site.**

## Prerequisites:

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

ISBN-10: 1118879368

<https://tinyurl.com/F21-CSCIE-96-1>

* Software: R & R-Studio
  1. If you are not familiar with R Studio please take a short introduction to R course at Lynda.com, DataQuest.com or DataCamp.com. *See the skillset assessment in the assignments table for a link.*
  2. *Be sure you have permissions to install R, & R-Studio on your laptop. It is mandatory.*
* Access to git software to download data sets and class material or ability to download directly from the Internet
  1. *Be sure you have permissions to install Git on your laptop. It is mandatory.*
* A webcam or other method to record case presentations & upload to the University’s approved site as well as interact with live class sessions if that is your mode of attendance. If you prefer you can share videos directly with teaching staff for privacy.
* Be prepared to obtain a free zoom account 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 understanding 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. Remote participation is measured by questions asked, posts, post views and responses to posts. 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](https://www.extension.harvard.edu/resources-policies/student-conduct). 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](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 case studies, a written assignment, and attendance and participation.

Case material is accepted up to 12 hours late. Any work submitted after the deadline but before 12 additional hours will be penalized 10%. After 12 hours no late submissions will be accepted under ANY circumstances. Pupils are expected to manage their own time and submit their work accordingly. 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.

* Class participation, and online forum participation 15% of final grade. The course is a collaborative learning environment. The expectation is that *all* students will view, comment, post interesting/current data science and business topics or news and ask questions. *Class participation is not free credit. If students do not contribute, they will not receive class participation credit.* Remote, asynchronous students are expected to contribute to the forum, by asking questions, posting relevant articles or answering others questions.

**Graduate Student Grading**

* Class Participation 20%
* Case I 20% of final grade: **OK Cupid Case**
* Case II 20% of the final grade: **Text Analysis & Document Classification**
* Case III 20% of the final grade: **Banking Case**
* Written assignment 20% of final grade

**Undergraduate Student Grading (no case II)**

* Class Participation 25%
* Case I 25% of final grade: **OK Cupid Case**
* Case III 25% of the final grade: **Banking Case Upload**
* Written assignment 25% of final grade

## Writing Assignment

Twenty-percent of the final grade (for graduate students) 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 data science ethics articles covered in class. The balance of the essay can incorporate new literary sources and 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, actions stakeholders can take to mitigate issues and the logical structure of your essay 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.

## Case Presentations

Each student will work on four case studies individually (or 3 as an undergraduate). Cases will involve business related data driven decision making. Each case will have the following work artifacts:

* Maximum 10min recorded slide presentation uploaded to youtube, embedded as a voiceover in the slide or shared in a similarly appropriate manner.
  + The presentation will outline the business problem, the insights identified and the outcomes/recommendations satisfying the case
* Slide presentation uploaded to canvas
* R Script(s) supporting the creation of any visuals, models or recommendations made during the presentation.

Essentially all supporting material including scripts, visuals and/or presentation slides will need to be turned in for review.

## Classes

**Tasks in BOLD are considered assignments and/or extra credit.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | 8-9pm | 9-10pm | Reading Due | **Tasks**  Suggested Book Exercises & Tasks |
| Aug 30 | Introduction & Administrative | Intro to R | NA | 1. Skillset Assessment:    1. Create R Studio Cloud Account    2. Connect to Git Student Repository    3. Chapter1: [Free Intro to R Course](https://campus.datacamp.com/courses/free-introduction-to-r/chapter-1-intro-to-basics-1?ex=1) (1-2hr for chapter1) |
| Sept 6 | **Labor Day – US Holiday No Class** | | | |
| Sept 13 | Intro to Data Mining | EDA | Chapter 1  Chapter 2 | **1.** **Forum introduction post**  -C2.1 Data Mining Techniques  -C2.2 Data Partition  -C2.3 Data Sample  -C2.4 Modeling Steps |
| Sept 20 | Data Mining in a Business Workflow | Data Preprocessing  Donor Bureau Case | Chapter 3 | **Extra Credit**  **2. Intro\_To\_R\_Homework.R** |
| Sept 27 | Regression | | Chapter 6 |  |
| Oct 4 | Logistic Regression | | Chapter 7  Chapter 10 | **3. CASE I. OK Cupid Case Upload**  -C6.1 Predicting Boston Housing Prices  Only do a, & b |
| Oct 11 | **Indigenous Peoples’ Day – US Holiday No Class** | | | |
| Oct 18 | KNN | Decision Tree | Chapter 9 | -C10.3 Sales of Riding Mowers  Only do a, b,c & d  -C7.2 Personal Loan Acceptance |
| Oct 25 | Random Forest | Time Series Forecasting | Chapter 16, 17, & 18 | -C9.3 Predicting Prices of Used Cars  Only do “a” |
| Nov 1 | Natural Language Processing (NLP) | | 20 |  |
| Nov 8 | NLP Document Classification | | NA |  |
| Nov 15 | Equity Trading | | NA | -C16.1 Impact of 9/11 on Air Travel  \*use Sept11Travel\_REVISED.csv  -C18.9 Australia Wine Sales  Only do “a”, & “b” |
| Nov 22 | Consumer Credit Risk Modeling | Non-Traditional Investment Modeling | NA | **4. CASE II Text Analysis & Document Classification** |
| Nov 29 | Data Sources with R – APIs & Webscraping | Reporting Automation w/ Flexdashboard | NA |  |
| Dec 6 | Collaborative Filtering | Ethics | Chapter 14 | **-** C14.2 Identifying Course Combinations |
| Dec 13 | Guest Speakers, *awaiting confirmation or Growth Modeling as a backup*   * Angela Chow Manager Toast * Gaurav Rao, Head of Product, Neural Magic * Rachel Switchenko, Dir Customer Care Plymouth Rock Assurance * James Liu, Product Manager Amazon Web Services * Ross Leav, Partner, Presidio Ventures | | |  |
| Dec 17 (Friday) | NA- not a class session | | | **5. Case III Banking Case Upload**  **6. 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 case presentations, extra credit homework, and the ethics paper. Your assignments will receive feedback to improve your acumen. However noncredit student may not receive letter grades.

Graduate credit students are expected to do more work and perform at higher standards than undergraduate credit students. As a result, a graduate credit student’s written assignment should incorporate an additional 3 sources of information beyond the covered text. Similarly a graduate writing paper needs to demonstrate nuanced sophistication of the ethical considerations presented along with appropriate counter arguments using a philosophical paradigm not opinion. Graduate students must complete all 3 case studies while undergraduate credit students are assigned 3.

## 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, certificate candidacy, or funding status will not have any impact on a course grade. “Needing an A” for any reason is not sufficient to earn an A 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. To the student’s benefit, one can still earn a “plus” on their final grade according to the scale below.

|  |  |  |
| --- | --- | --- |
| 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 |