Harvard Extension School
HARVARD DIVISION OF CONTINUING EDUCATION

CSCI E-96
Data Mining for Business

Fall Term 2023

Course Information

CRN: 15736
Section Number: 1
Format: Flexible Attendance Web Conference
Credit Status: Graduate
Credit Hours: 4

Class Meetings: Mondays, September 11-December 21, 8:10pm-10:10pm

Course Description: This course introduces non-mathematical business professionals to data science principles widely used in today's corporations. Quantitative methods affect many of today's interactions for business leaders, students, and consumers. Emphasis is placed on practical uses and case studies utilizing data to inform business decisions rather than theoretical or complex mathematics. Case study topics include understanding customer demand, marketing, new market forecasting, revenue projections, and data mining to improve decisions. Learning goals include quantitative business application, basic programming, algorithm development, and process workflow. The course highlights methods that business leaders and data scientists have found to be the most useful. It introduces the basic concepts of R for data mining. This course is for students who want an introduction to how data science improves business outcomes.

Prerequisites: Since this course utilizes R throughout the semester students should complete the 4-hour free online course *Introduction to R* at DataCamp.com found here: https://www.datacamp.com/courses/free-introduction-to-r.

Instructor Information & Office Hours

Ted Kwartler
Email: edwardkwartler@fas.harvard.edu

Kalle Georgiev
Email: kgeorgiev@g.harvard.edu

Course Goals / Learning Outcomes

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 understand what business stakeholders expect of data scientists.

Mode of Attendance & Participation Policy

This class offers a live or on-demand option, which means you can choose to attend the class live over Zoom or watch the class recording afterward. You do not need to commit to the same mode of attendance for the whole semester.

If you are attending live over Zoom:

Class meetings take place over Zoom. Because they involve active participation, discussion, and dialogue, you are expected to attend all class meetings. Please arrive on

time. You should attend Zoom meetings with a functional web-camera and microphone, prepared with materials needed, to engage thoughtfully, and with your camera on. You may turn off your camera for occasional interruptions or momentarily for privacy.

You will also need the most up-to-date Zoom client installed on your computer to join class. Please participate from a safe and appropriate environment with appropriate clothing for class. Participating while traveling or in a car is not permitted. In addition, please do not join class via mobile phone or web browser.

If you are participating on demand:

You are expected to watch the class recording and complete any assignments before the next live class meets.

Please be sure to review important information on **Student Policies and Conduct**.

Grading & Grade Definitions

Grading

A course grade will be assigned based on student performance on case studies, applicable homework assignments, and a written assignment.

Assignments are accepted up to 12 hours late. Any work submitted after the deadline but before 12 additional hours will be penalized 10% of the total weight of the assignment. 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.

Graduate Student Grading (4 cases)

- Skills Assessment: 0%: Complete the provided but unfinished R script.
- Case I 25% of final grade: EDA Case
- Case II 25% of the final grade: Banking Case
- AI Ethics Case 25%: Build a model evaluating it for both accuracy and unfair bias of a protected feature
- Extra Credit: Homework II Visualization in R 1% of total grade

Undergraduate Student Grading (only 3 cases)

- Skills Assessment: 0%: Complete the provided but unfinished R script.
- Case I 25% of final grade: EDA Case
- Case II 25% of the final grade: Banking Case
- Homework I: 10% Intro to R script - more info to come
- Homework II: 20% Visualization in R script - more info to come
- Homework III: 20% Obtain 2 AI/Data Ethics related articles (https://incidentdatabase.ai is a good resource):
    - Use ChatGPT to summarize the article
    - Critique the summarization as appropriate, inappropriate, missing relevant facts, creating or citing information from outside the article etc. in a single paragraph
    - Write 1 paragraph WITHOUT GPT with your personal reflection on the use or misuse of the technology cited in the article. In the paragraph suggest ways to mitigate or monitor to protect against the issue within the article.

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

Case Work Product

Each case will have a description and specific instructions provided through the course github repository.

Each student will work on case studies individually. Each case will have the following work artifacts:

- Maximum 10min recorded slide presentation uploaded to youtube, embedded as a voiceover in the slides or shared in a similarly appropriate manner.
    1. The presentation will outline the business problem, the insights identified, describe the data and the outcomes/recommendations satisfying the case
- Slide presentation uploaded to canvas (pptx file for example)
- R Script(s) supporting the creation of any visuals, models or recommendations made during the presentation.
- Written supplemental describing problem, data, and specific recommendations.

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

The presentation will be evaluated on an equal-weighted scale with the following criteria.

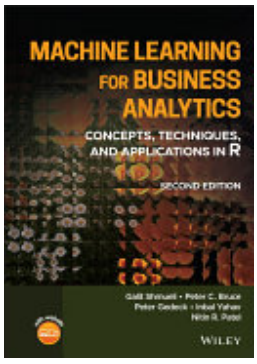
- Organization – Was the presentation well organized?
- Delivery – Was the content delivered clearly and persuasively with the audience in mind?
- Code Documentation – Was the data mined to support the conclusion?
- Written Supplemental – Is the information clear and supported in narration and code? Did the information satisfy the case problem?
- Data Mining & Modeling Process – Overall, as a complete portfolio of work, is the topic interesting, organized, researched, supported and delivered effectively? Was CRISP-DM, SEMMA, or a similar workflow followed to organize the work?

Suggested Tasks & Assignments

Please note that suggested tasks and assignments may be expanded upon within the class repository.

https://github.com/kwartler/Harvard_DataMining_Business_Student

Course Materials



Machine Learning for Business Analytics
ISBN: 9781119835172
Authors: Galit Shmueli, Peter C. Bruce, Peter Gedeck, Inbal Yahav, Nitin R. Patel
Publisher: John Wiley & Sons
Publication Date: 2023-03-08
This textbook has some overlap with lessons and should be purchased by students wishing to expand beyond lessons to add additional fluency and technical knowledge. Each week has suggested readings and exercises from this textbook to reinforce and expand topics covered in class.

Academic Integrity Policy

You are responsible for understanding Harvard Extension School policies on Academic Integrity and how to use sources responsibly. Violations of academic integrity are taken very seriously. Visit Using Sources Effectively and Responsibly and the Harvard Guide to Using Sources to review important information on academic citation rules.

Writing Code. While it may be common practice in non-academic settings to adapt code examples found online or in texts, this is not the case in academia. In particular, you should never copy code produced as coursework by other students, whether in the current term or a previous term; nor may you provide work for other students to use. Copying code from another student or any other source is a form of academic dishonesty, as is deriving a program substantially from the work of another.

Writing code is similar to academic writing in that when you use or adapt code developed by someone else as part of your assigned coursework, you must cite your source. Paraphrasing without proper citation is just as dishonest with programming as it is with prose. A program can be considered plagiarized even though no single line is identical to any line of the source.

Accessibility Services Policy

The Division of Continuing Education (DCE) is committed to providing an accessible academic community. The Accessibility Services Office (ASO) is responsible for providing accommodations to students with disabilities. Students must request accommodations or adjustments through the ASO. Instructors cannot grant accommodation requests without prior ASO approval. It is imperative to be in touch with the ASO as soon as possible to avoid delays in the provision of accommodation.

DCE takes student privacy seriously. Any medical documentation should be provided directly to the ASO if a substantial accommodation is required. If you miss class due to a short-term illness, notify your instructor and/or TA but do not include a doctor's note. Course staff will not request, accept, or review doctor's notes or other medical documentation. For more information, email accessibility@extension.harvard.edu.

Publishing or Distributing Course Materials Policy

Students may not post, publish, sell, or otherwise publicly distribute course materials without the written permission of the course instructor. Such materials include, but are not limited to, the following: lecture notes, lecture slides, video, or audio recordings, assignments, problem sets, examinations, other students' work, and answer keys. Students who sell, post, publish, or distribute course materials without written permission, whether for the purposes of soliciting answers or otherwise, may be subject to disciplinary action, up to and including requirement to withdraw. Further, students may not make video or audio recordings of class sessions for their own use without written permission of the instructor.

Class Meeting Schedule

Adjustments will be made to the lessons based on the learning rate and priorities of the class.

September 4: NO CLASS (University holiday)

September 11: Class 1 - Introduction & Administrative, Introduction to R

September 18: Class 2 - Introduction to Data Mining, Basic EDA

Assignments: Forum Introduction post (assuming we can add a forum to Canvas) & Finish Skills Assessment

Suggested Reading: Chapters 1,2

September 25: Class 3 More R Practice: Visualization and more EDA

Assignments DUE (undergraduate only): HW1 Intro_To_R_Homework.R

Suggested Reading: Chapter 3

October 2: Class 4 Data Mining in a business workflow, data preprocessing, Donor Bureau Case

Suggested Reading: Chapter 6

October 9: NO CLASS (University holiday)

October 16: Class 5 Regression & Logistic Regression

Assignments DUE: Case I EDA Case

Suggested Reading: Chapters 6 & 10

October 23: Class 6 Decision Tree & Random Forest

Assignments DUE (mandatory for undergraduate & extra credit for graduate students): Visualization R Script

Suggested Reading: Chapter 9

October 30: Class 7 Time Series & Equity Trading

Suggested Reading: Chapters 17,18,19

November 6: Class 8 Consumer Credit Risk Modeling & Non-traditional Investment Modeling

November 13: Class 9 Natural Language Processing (NLP) - Bag of Words

November 20: Class 10 Natural Language Processing (NLP) - How does chatGPT work & Document Classification

Assignment DUE: Case II Banking Case

November 27: Class 11 *Possible Prerecorded Session* Getting Data for your own projects: APIs & Webscraping

December 4: Class 12 Unsupervised Clustering Analysis & Discriminant Analysis

Suggested Reading: Chapters 12,16

December 11: Class 13 Hearing from industry professionals in the data space

Guest Speakers, awaiting confirmation

- Ross Leav, Presidio Ventures
- Rachel Switchenko, VP Customer Care Plymouth Rock Assurance
- James Liu, Product Manager Amazon Web Services

December 18: (Final Exam or final class meeting) - Class Lab session for final Case help

December 20: (NO CLASS) - Class Lab session for final Case help

Assignments DUE (undergraduate only): Homework III: AI Ethics reflection assisted with GPT

Assignments DUE (graduate only): Build and Evaluate a model for accuracy & bias