



Harvard Extension School
HARVARD DIVISION OF CONTINUING EDUCATION

CSCI E-96
Data Mining for Business

Spring Term 2024

Course Information

CRN: 26599

Section Number: 1

Format: Flexible Attendance Web Conference

Credit Status: Graduate

Credit Hours: 4

Class Meetings: Mondays, January 22-May 11, 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: Students need to bring laptops to class to use for exercises.

Instructor Information & Office Hours

Edward Kwartler

Email: edwardkwartler@fas.harvard.edu

Kalle Georgiev

Preferred Pronouns: she/her

Email: kgeorgiev@g.harvard.edu

Section Meetings

Additional lab sessions to support the case study assignments will be hosted by the TAs.

The dates and times are announced within the class discussion forum "Ed Discussion" and are recorded.

Course Goals / Learning Outcomes

If you stay engaged in the course and complete the suggested code challenges 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 is a flexible attendance course, which means you can choose to: (1) attend class live over Zoom (synchronous option); or (2) watch the class recording afterward

(asynchronous option). 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 asynchronously:

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 and applicable homework assignments (undergraduate).

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 Level Assignments

Skills Assessment: 0%: Complete the provided but unfinished R script.

Case I 25%: Retail Store Transaction EDA

Case II 25%: Banking Case

Case III 40%: Hotel/Lodging Booking Case

Assessment 10%: Time Series Forecast TSA Passengers

Extra Credit: Homework II Visualization in R 1% of total grade

Undergraduate Level Assignments

Skills Assessment: 0%: Complete the provided but unfinished R script.

Case I 25%: Retail Store Transaction EDA

Case Case II 25%: Banking Case

Homework I: 10% Intro to R script - listed in class repository

Homework II: 20% Visualization in R script - listed in class repository

Homework III: 20% Obtain 2 AI/Data Ethics related articles (<https://incidentdatabase.ai> is a good resource). Perform the following tasks for each article.

- 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.
- Cite the gaps in understanding from the technology you note to provide a more nuanced understanding
- Write 1-3 paragraph(s) 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.

Case Work Product & Expectations

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 (provide the link in your Canvas submission), embedded as a voiceover in the slides or shared in a similarly appropriate manner to Canvas. 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.
- At least one case requires predictions be made on new data. For example, "identify the top 100 prospective customers". If mentioned in the specific case instructions, please score, identify the appropriate observations, save results in a CSV and submit that file as well.
- A Written supplemental: Submit something to represent the entirety of your presentation including the data, process, findings, and implications. Thus it's a professional report, anything less than a professionally written and organized report will be considered sub-optimal. Amazon for example doesn't use PowerPoint and instead uses "6 pagers" to make business recommendations, as such some organizations prefer written information over presentations. The use of external and verifiable sources is expected to add context and support any component of the paper. **The minimum is 2 pages maximum is 5. Double spaced and 12 point font.**

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

Case Evaluation

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?

Grade Definitions

Students registered for undergraduate or graduate credit who complete the requirements of a course may earn one of the following grades:

A Earned by work whose superior quality indicates a full mastery of the subject—and in the case of A, work of extraordinary distinction. There is no grade of A+.

B+, B Earned by work that indicates a strong comprehension of the course material, a good command of the skills needed to work with the course materials, and the student's full engagement with the course requirements and activities.

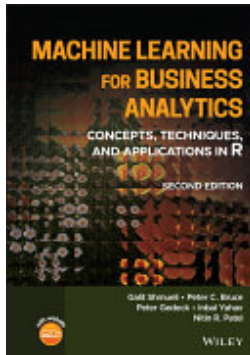
C+, C Earned by work that indicates an adequate and satisfactory comprehension of the course material and the skills needed to work with the course materials, and that indicates that the student has met the basic requirements for completing assigned work and participating in class activities.

D+, D Earned by work that is unsatisfactory but that indicates some minimal command of the course materials and some minimal participation in class activities that is worthy of course credit.

E Earned by work that is unsatisfactory and unworthy of course credit. This grade may also be assigned to students who do not submit required work in courses from which they have not officially withdrawn by the withdrawal deadline. Zero or E grades are assigned to students for missing work. These grades are included in the calculation of the final grade.

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 above.

Course Materials



Machine Learning for Business Analytics

ISBN: 9781119835172

Authors: Galit Shmueli, Peter C. Bruce, Peter Gedeck, Inbal Yahav, Nitin R. Patel

MACHINE LEARNING FOR BUSINESS ANALYTICS Machine learning —also known as data mining or data analytics— is a fundamental part of data science. It is used by organizations in a wide variety of arenas to turn raw data into actionable information. Machine Learning for Business Analytics: Concepts, Techniques, and Applications in R provides a comprehensive introduction and an overview of this methodology. This best-selling textbook covers both statistical and machine learning algorithms for prediction, classification, visualization, dimension reduction, rule mining, recommendations, clustering, text mining, experimentation, and network analytics. Along with hands-on exercises and real-life case studies, it also discusses managerial and ethical issues for responsible use of machine learning techniques. This is the second R edition of Machine Learning for Business Analytics. This edition also includes: A new co-author, Peter Gedeck, who brings over 20 years of experience in machine learning using R An expanded chapter focused on discussion of deep learning techniques A new chapter on experimental feedback techniques including A/B testing, uplift modeling, and reinforcement learning A new chapter on responsible data science Updates and new material based on feedback from instructors teaching MBA, Masters in Business Analytics and related programs, undergraduate, diploma and executive courses, and from their

students A full chapter devoted to relevant case studies with more than a dozen cases demonstrating applications for the machine learning techniques End-of-chapter exercises that help readers gauge and expand their comprehension and competency of the material presented A companion website with more than two dozen data sets, and instructor materials including exercise solutions, slides, and case solutions This textbook is an ideal resource for upper-level undergraduate and graduate level courses in data science, predictive analytics, and business analytics. It is also an excellent reference for analysts, researchers, and data science practitioners working with quantitative data in management, finance, marketing, operations management, information systems, computer science, and information technology.

Publisher: John Wiley & Sons

Publication Date: 2023-03-08

Edition: 2nd

THIS BOOK IS OPTIONAL.

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.

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.

AI Technologies. The Extension School's [Academic Integrity Policy](#). prohibits students from representing work as their own that they did not write, code, or create. It is never permissible to submit work generated by machine learning and AI technologies (such as ChatGPT) without proper attribution.

In our course we allow for the use and support of the BETA software, www.llmaid.com, where students have the option of signing up (instructions given in class). This

technology uses chatGPT but is adjusted based on our course material. Thus chatbot answers are returned along with references to specific lecture material and students are encouraged to use this resource if desired.

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

January 22: Introduction to R, R-studio & Git software.

We will set up our development environment on your local laptops & explain how each of the software components are used. The class will execute some basic code examples.

January 29: Introduction to Data Mining & Exploratory Data Analysis (EDA)

We introduce more coding concepts including data types & functions. Then we start to get summary statistics and build visualizations based on data.

February 5: More R Practice & EDA

The class will join data sets, and build more robust and interesting visualizations as part of more sophisticated EDA.

February 12: Data Mining Workflows

As we increase our sophistication, before applying modeling and machine learning methods in a business use case we learn the basic steps to prepare our data set including dummy variables, feature engineering and imputation. By the end of class we build our first simple machine learning model.

February 19: NO CLASS (University holiday)

February 26: Regression & Logistic Regression Modeling

The class will learn the basics of linear regression and how this traditional model can still be useful for solving business problems. The class will follow code and execute the steps of a model fit including future predictions after cleaning up realistic business data.

March 4: Decision Tree & Random Forest

Students will conceptually learn how a decision tree and random forest learn patterns within a data set. The class will build and evaluate real machine learning or "AI" models in a business context using all the code, data cleaning, EDA and machine learning methods learned to date.

March 11: *Spring Break* - NO CLASS

March 18: Time Series Data

The class will learn about a new data class called time series which is useful in demand forecasting. Code and explanations are given for predicting quarterly revenue and other time series so students learn the basics of this robust field and how to identify when data has a temporal element.

March 25: Equities

Students will use an API to obtain stock trading data and learn basic methods of algorithmic trading. The goal is not to make day-traders of the participants but to showcase how markets are highly efficient and technically supported despite retail investors being naive to technical trading rules and methods.

April 1: Predicting Risk (Credit Default) & Non-Traditional Market Simulation/Investing

The class covers building models to predict individual investment outcomes then selecting optimal investments based on the relationship between risk & reward. Next, a simple simulation is constructed to mimic a non-traditional investment market investing in collectible trading cards so an investor can understand the outcomes before risking capital.

April 8: Text Analysis & Natural Language Processing

As a new data set to the class, we explore how to extract insights and frequent phrases from text. As a basic primer on natural language processing we discuss and demonstrate code for "bag of words" methods.

April 15: Text Analysis & Natural Language Processing , Continued

We expand on the previous NLP lesson to include sentiment analysis & unsupervised methods to identify topics using spherical K-Means and LDA topic modeling machine learning methods. Time permitting we can introduce a document classification system using elastic-net regression.

April 22: Data Sources, cURL & APIs

In this lesson the class will learn how to perform limited web-scraping, and obtain data from unsecured and secured application program interfaces (APIs). Time permitting we will demonstrate R performing a "cURL" request mimicking a web browser to obtain data.

April 29: Responsible AI & Technology Ethics

Technology ethics, and how to think about the responsible use of technology are paramount in this AI driven world. As practitioners and consumers affected by AI, it is important to have a grasp of the technology, its limitations and how to ethically apply these technologies. This lesson covers some basic tenets of responsible AI and students will purposefully build a sexist resume tool then apply methods to identify this unjust (and illegal) model behavior.

May 6: Class Lab

The course material is very wide ranging yet dense. This sessions allows for flexibility to cover topics that need additional attention, expand on use cases discussed, help with the final case assignment. In the past students have come to class to ask about approaching problems using the course techniques but applied to their own professional process.