NEW YORK UNIVERSITY STERN
Leonard N. Stern School of Business

DATA MINING FOR BUSINESS ANALYTICS INFO-GB.3336

SYLLABUS - Tentative

Professor	**Foster Provost,** Information, Operations & Management Sciences Department
Office; Hours	TBD
Email	fprovost@stern.nyu.edu **Begin subject: [DM GRAD] …** **← note!**
Telephone	Office 212-99-80806, Fax: 212-99-54228
Classroom	TBD
Class time	Tues 6pm-9pm
First : Last Class	TBD
Final Quiz	Take home after last class
Course Assistants CA Office Hours	TBD

1. Course Overview

This course will change the way you think about data and its role in business.

Businesses, governments, and individuals create massive collections of data as a by-product of their activity. Increasingly, decision-makers and systems rely on intelligent technology to analyze data systematically in order to improve decision-making. In many cases automating analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data analysis technologies can be used to improve decision-making. We will study the fundamental principles and techniques of data mining, and we will examine real-world examples and cases to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work "hands-on" with data mining software.

After taking this course you should:

1. *Approach business problems data-analytically.* Think carefully & systematically about whether & how data can improve business performance, to make better-informed decisions for management, marketing, investment, etc.

2. *Be able to interact competently on the topic of data mining for business analytics.* Know the fundamental principles of data science, that are the basis for data mining processes, algorithms, & systems. Understand these well enough to work on data science projects and interact with everyone involved. Envision new opportunities.

3. *Have had hands-on experience mining data.* Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies.

2. Focus and interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and some technical details of data mining and data science. The emphasis primarily is on understanding the fundamental concepts of data science and business applications of data mining. We will discuss the mechanics of how the methods work as is necessary to understand and illustrate the fundamental concepts and business applications. This is not an algorithms course. However, many techniques are the embodiment of one or more of the fundamental principles.

I will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures/discussions on the fundamental material, discussions of business applications of the ideas and techniques, case discussions, student exercises, and demos.

You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including (unless otherwise directed) having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class. In general, we will follow Stern default policies unless I state otherwise. I will assume that you have read them and agree to abide by them:

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

The NYU Classes site for this course will contain lecture notes, reading materials, assignments, and late-breaking news. You should check the site daily, and I will assume that you have read all announcements and class discussion.

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, come to office hours to see me or the TAs, or ask on the discussion board. The discussion board is much better than sending me email, which frankly I have a hard time keeping up with. Also, if you have the question, someone else may too and everyone may benefit from the answers being available on NYU Classes. Also, please try to answer your classmates' questions. In grading your class participation I will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

Worth repetition: It is your responsibility to check NYU Classes (and your email) at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent.

I will check my email at least once a day. ***Your email will get my priority if you include the special tag [DM Grad] in the email subject header***. I use this tag to make sure to process class email first. If you do not include the special tag, I may not read the email for a while (maybe a long while). If you forget and send without the tag and then remember, *just send it again* including the tag.

3. Lecture Notes and Readings

Book: The textbook for the class will be:

Data Science for Business: Fundamental principles of data mining and data analytic thinking Provost & Fawcett (O'Reilly, 2013).

This book covers the fundamental material that will provide the basis for you to think and communicate about data science and business analytics. We will complement the book with discussions of applications, cases, and demonstrations.

Lecture notes: For many classes I will hand out lecture notes. I expect you to ask questions about any material in the notes that is unclear after our class discussion and reading the book. Having the book allows frees up class time for more discussion of applications, cases, etc.—so many of your questions may be answered in the book. If not, please let me know! Depending on the direction our class discussion takes, we may not cover all material in the class notes for any particular session. If the notes and the book are not adequate to explain a topic we skip, you should ask about it on the discussion board. I will be happy to follow up.

I may hand out or post some additional required readings as we go along. *Note that some of these readings may be accessible for free only from an NYU computer. If you can't access a link from home, please try it from school.*

For those interested in going further, these following supplemental books give alternative perspectives on and additional details about the topics we cover. These are completely optional; you will not be required to know anything in these readings that are not in the primary materials or lectures. I have many other books that I can recommend, for example if you want a reference to a more mathematical treatment of the topics. Please don't hesitate to come and talk to me about what supplemental material might be best for you, if you want to go further.

- Supplemental book (optional):
  Data Mining Techniques, Second Edition
  by Michael Berry and Gordon Linoff, Wiley, 2004
  ISBN: 0-471-47064-3
  - *available as ebook for free: http://site.ebrary.com/lib/nyulibrary*
  - This book may give a different, useful perspective on many of the topics
  - The Third Edition is out. I have not read it yet. Berry says it has been improved substantially. I have a copy in my office if you want to talk a look at it before buying.
  - *available from Amazon*

"Weka Book" (optional):

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition
by Ian Witten, Eibe Frank, Mark Hall
ISBN-10: 0123748569

- available from Amazon
- This book provides much more technical details of the data mining techniques and is a very nice supplement for the student who wants to dig more deeply into the technical details. It also provides a comprehensive introduction to the Weka toolkit.

4. Requirements and Grading

The grade breakdown is as follows:

1. Homeworks: 20%
2. Term Project: 30%
3. Participation & Class Contribution: 20%
4. Final Quiz: 30%

At NYU Stern and the Center for Data Science we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving As and Bs, and only one or two perform only adequately or below and receive C's or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

Homework Assignments

The homework assignments are listed (by due date) in the class schedule below. Each homework comprises questions to be answered and/or hands-on tasks. Except as explicitly noted otherwise (see next paragraph), you are expected to complete your assignments on your own—without interacting with on the completion of your assignment. You are free of course to discuss the concepts with your classmates, and to discuss similar problems to the ones in the homeworks.

For the hands-on parts of the assignments (with Weka or Python), I encourage you to work with your group members and other classmates to understand how to get Weka or Python to do what you need to do, and then to complete your assignment on your own. So, for example, you could have a classmate help you do something similar, such that then you would be able to complete the assignment.

I hope with the support of me, the TAs, and your classmates, we operate under a "diligent attempt but limited frustration" policy: (1) If you get stuck on something, spend some time Googling to try to find the answer. If you seem to be moving forward, keep going. That search and discovery will pay off, both in terms of the direct learning about how to do what you need to do, and also in terms of your learning *how to find* such things out. (E.g., if you don't know what stackoverflow is, you will learn!). BUT, (2) limit frustration—start your assignments early enough that if you run into a wall, you can just stop searching and ask about it. Let's say, if you feel like you have not moved forward after 15 minutes of being stuck, just stop and ask: your classmates, on the discussion board, to the TAs. If you don't get a solution, escalate it to me.

Completed assignments must be handed on blackboard at least one hour prior to the start of class on the due date (that is, by 5pm), unless otherwise indicated. Assignments will be graded and returned promptly. Answers to homework questions should be well thought out and communicated precisely, avoiding sloppy language, poor diagrams, and irrelevant discussion.

The hands-on tasks in the homeworks will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class. You will use these data to build and evaluate predictive models.

For the hands-on assignments you will use either the (award-winning) toolkit Weka or Python and its data science/analytics/visualization libraries.

http://www.cs.waikato.ac.nz/ml/weka/ download the "latest stable" version (3.6.10) (which is the version associated with the 3rd edition of the Weka Book)

For Python we will provide installation instructions to make sure that you have all the required libraries.

IMPORTANT: ***You must have access to a computer on which you can install software. If you do not have such a computer, please see me immediately so we can make alternative arrangements.*** *You should bring your computer to class.* During class we will have a "lab session" during which we will aid anyone who needs help with installing and configuring the software, getting it running, and dealing with the inevitable glitches that a few of you might experience. If you need additional help with using the data mining software, please see the Course Assistant(s).

Generally the Course Assistants should be the first point of contact for questions about and issues with the homeworks. The primary course assistant (see first page) will have the responsibility to make sure that all questions are answered in a timely fashion, but please make use of both, as we have staggered the office hours to provide broader coverage. *If they cannot help you to your satisfaction, please do not hesitate to come see me.*

Late Assignments
As stated above, assignments are to be submitted on NYU Classes at least *one hour prior* to the start of the class on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time.

Term Project
A term project report will be prepared by student teams. We will give you the instructions on how to form your teams. Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. You will submit various milestone deliverables through the course. We will discuss the project requirements in class.

Final Quiz
The final quiz will be a take-home to be completed during the days following the last class. The subject matter covered and the exact dates will be discussed in class.

Participation/Contribution/Attendance/Punctuality
Please see Section 2.

Regrading
If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., pages in the book, a copy of class notes, etc.). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go "in your favor" and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

***FOR STUDENTS WITH DISABILITIES*:** If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

Please read the policies for Stern courses

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

Please keep in mind the Stern Honor Code

http://www.stern.nyu.edu/mba/studact/mjc/hc.html

Example Class Schedule from Fall 2013

Class Number	**Date**	**Topics** *(subject to change as class progresses)* **Most classes will also include a case study/guest/lab/demo/etc.**	**Readings**	**Deliverables**
1		Introduction to the Course Introduction to Predictive Modeling **Case:** **Data science for managing churn in wireless telecom**	Ch. 1 & 2	**Info Sheet (online)**
2		Predictive Modeling (cont.) Supervised Segmentation **Discussion: Target – predicting pregnancy** **Installation "lab"** **Data Mining Demo**	Ch. 3	**HW#1 due**
3		Predictive Modeling (cont.) Model performance analytics I: Fitting the data and overfitting the data, holdout testing, cross-validation, learning curves **Case Study: Data Mining for Operations Support**	Ch. 4 & 5	**Due: Team Choices and initial project ideas**
4		Model performance analytics II: Ranking, Profit, Lift ROC analysis, expected value framework, domain knowledge validation **Case Study: Modeling consumer behavior for marketing (banking, online advertising)**	Ch. 7 & 8	**HW#2 due**
5		Similarity, Distance, Nearest Neighbors **Case Study: IBM Salesforce Optimization**	Ch. 6	**Project Proposal due**
6		Mining fine-grained data, Prediction via evidence combination, Bayesian reasoning, text classification, "Naïve" Bayes	Ch. 9 & 10	**HW#3 due**

Class Number	Date	Topics	Readings	Deliverables
7		Descriptive data mining, unsupervised methods, associations, clustering **Case Study: TBD**	Ch. 6 (revisit clustering) Ch. 12	**Project Update Due**
8		Data Visualization **Guest: Prof. Kristen Sosulski**	TBD	
9		Explanatory Data Science **Guest: Ori Stitelman** **Director of Data Science at Millennial Media** **(Formerly: at Wells Fargo)**	TBD	**HW#4 due**
		No Class (Thanksgiving Break)		
10		Issues in Deployment Data Science Team Development **Guest: Prof. Claudia Perlich** **Chief Scientist, Dstillery**	Ch 13	**HW#5 due**
11		Toward Analytical Engineering Data Science and Business Strategy Wrap Up	Ch. 11 (Revisit 13)	
12		Project Presentations		**Project report**
Final Quiz: *Taken online*				