NEW YORK UNIVERSITY
STERN
Leonard N. Stern School of Business

INFO-GB.3336.xx
Data Mining for Business (Managerial)

Professor	**Claudia Perlich, PhD** Adjunct Professor, Stern School of Business Email: cperlich@stern.nyu.edu please put “NYU-DM” in the subject Phone: 914 4095609

Course Description

The goal of this course is NOT to turn you into a data scientist, but rather to give you a deep enough understanding of the opportunities, techniques and critical challenges of using data mining and predictive modeling in a business setting and managing data science teams. It will be a hands-on experience, comparatively light on programming requirements. The focus is on the ability to understand and translate business challenges into data mining problems and cover in depth the process of thinking about data sciences with respect to a specific problem that needs solving. A core requirement is the understanding of core technical concepts and the personal experience in homework and projects of building data mining solutions. We will examine how data mining and predictive modeling 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. A large part of the class is the discussion of ‘cases’ as experiences by the teacher over her 15+ years of applying data science in a number of industries.

The course is a combination of lecture, case studies, homework and a final project.

Course Objective

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 intelligence.* Know the basics of data mining processes, algorithms, & systems well enough to interact with CTOs, expert data miners, consultants, etc. Envision 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.

Focus and interaction

The course will explain through lectures and real-world examples explore uses and some technical details of data mining techniques. The emphasis primarily is on understanding the business application of data mining techniques, and secondarily on the variety and technical details of techniques. We will discuss the mechanics of how the methods work as is necessary to understand the fundamental concepts and business application. This is not an algorithms or a programming course. We will cover a number of cases during the course of the semester. Their goal is for you to understand how certain modeling decisions were made in the context of the business decision. Part of the homework’s will be a review of cases and your understanding. I will expect you to participate in class discussions and be proactive about ensuring that you understand what we have done in the prior classes. The assigned readings will cover the fundamental material and I am always available to discuss.

Class Room Attendance and Equipment

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 having all electronic devices turned off and put away for the duration of the class (this is

Stern policy, see below) unless stated otherwise. 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://www.stern.nyu.edu/AcademicAffairs/Policies/GeneralPolicies/DefaultPoliciesforSternCourses/index.htm

Office Hours and Email/Phone

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 talk to me after class or see me during the office hours. You may also send emails to ask questions or set up appointments outside of office hours. Please type "NYU-DM" in the subject line of every email that you send. I will check my email at least once a day during the week (M-F). If no reply is received within 48 hours, your email may have been overlooked and please feel free to send another email. I strongly encourage you to seem me at least once in this semester in my office hours. For urgent matters including setting up additional meeting times, you are welcome to call me or send me a text.

Course Homepage

The NYU Classes site for this course will contain lecture notes, reading materials, assignments, extra-class discussions, and late-breaking news. You should check the NYU Classes site regularly. If you do miss one of the classes it is your responsibility to acquire the content. In particular, all classes are taped and you can watch the last session on Echo 360. Important information and due dates will be communicated via announcements.

Lecture Notes and Readings

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).

All slides will be posted on NYClasses shortly prior to the lecture. I may hand out or post some additional required readings as we go along. For those interested in going further, these following supplemental books give alternative perspectives on and additional details about the topics we cover – some of them are located in the Resources/Books folder on NYClasses. 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.

- "Weka Book":Data Mining: Practical Machine Learning Tools and Techniques by Ian Witten, Eibe Frank, Mark Hall ISBN-10: 0123748569
- Elements of Statistical Learning by Hasties, Tibshirany and Friedman ISBN-10: 0387848576
- Introduction to Statistical Learning with code examples in R
- Data Science at the commandline

Cases

This class does not cover cases in the classical sense but I will on post a business problem during class and expect you to talk about it. Data Science is a craft that requires doing in order to learn. Unfortunately we have online limited time and I cannot expose you in depth to more than one project. Instead we will talk about these cases during class and I will share my experience in solving them. Some cases are full fledges publications and as the course progresses you should be able to understand nearly all of the technical details. Your participation in these discussions is crucial to your learning experience.

Software

During class we will use primarily WEKA and occasionally Excel (or another tool of your choice). WEKA is a GUI driven package with most major algorithms available. Instructions are on NYClasses.
http://www.cs.waikato.ac.nz/ml/weka/

Please download the "latest stable" version (3.6) (which is the version associated with the 3rd edition of the Weka Book)
You should bring your computer to class. During class we will have on occasion a "lab session" during which we will install and configure the software, get it running, and deal with the inevitable glitches that a few of you might experience. If you need additional help with using the data mining software, please reach out to me and plan on staying after class during one of the initial weeks. WEKA will be the tool you have to use for all homework assignments. For the project you are free to use ANY tool/software you wish. Recommended are Perl, Python, R, UNIX command line tools etc. WEKA is limited in its support of data formatting and is also unsuitable to deal with large datasets.

Homework Assignments

The homework assignments are listed in the class schedule below and are due as indicated on the class calendar on NYClasses. Final information will be communicated with announcements. 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. For the hands-on parts of the assignments (with Weka), I encourage you to initially work with your group members and other classmates to understand how to get Weka 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. Completed assignments must be handed on NYClasses unless otherwise indicated. The hands-on tasks will be based on data that we will provide. You will analyze 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. Plan on running into issues and start your homework early to be able to ask me questions PRIOR to the due date. Homework are a core learning experience - I prefer you to ask questions and get the homework right.

Late Assignments

I resume the right to reduce the grade for late assignments unless you have been in direct communication with me PRIOR to the due date - 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.

Term Project

The most important learning experience is the project. Please consider this a huge opportunity, not only to get a good grade in this course, but to shape your career as many students have done. It is an opportunity learn how to work with a team on data science problems, have a great entry on your resume, to potentially solve a cool problem and make a difference, and have me 'consult' on your problem for free.

The project will be executed by student teams of 3-4 students. *We will help seeing the group and you should decide on your teams by the end of the second week, and submit them to me – see the deadlines in the syllabus.* Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. Each team will present its project at the end of the semester. We will discuss the project requirements and presentations in class. Deadlines for project deliveries will be in the class calendar on NYClasses and the syllabus. If the two disagree – please follow the class calendar.

I strongly advocate that you start your project early – you will have to find data as well as formulate the problem. The most time consuming by far will be managing the data. You are free to use any tool at your disposal and it is highly recommended that you form diverse teams with at least one person who can program and one with a strong business background.

Final Quiz

The final quiz will be a take-home to be completed during the week following the last class. The subject matter covered and the exact dates will be discussed in class. The quiz will test your understanding of the subject matter and your ability to express the learned concepts in a concise way. You will be able to access all course material during the quiz.

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., a photocopy of class notes). 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.
betray

Requirements and Grading

The grade breakdown is as follows:

Homework	20%
Term Project (presentation and report)	40%
Final Quiz	30%
Class Participation/Case Discussion	10%

Stern grading policy requires specific percentage in the different grade levels and the total number of points will be rescaled accordingly. Please understand that I am not at liberty to make exceptions and that your grade unless you can show evidence of factual error in the grading is final.

Generic Class Schedule

Class	Topic	Readings/ Preparation	Deliverables (Preliminary)
1	Introduction Terminology		
2	Trees, Evaluation Basics WEKA 101	Read Chapter 1&2&3	Install WEKA Look for data
3	Optimization methods Classification Evaluation	Read Chapter 4&5	Homework 1 Project groups
4	Evaluation in Depth Server/Medical	Read Chapter 7&8,11	Homework 2
5	Naïve Bayes Language Watson/Banter	Read Chapter 9 & 10	Project Proposal
6	Recommender Knn Wallet	Read Chapter 6	Homework 3
7	Causal Modeling Chobani Guest: Ori Stitelman		Homework 4
8	Data Mining Applications Guest: F. Provost & S. Hill		Project Update
9	Unsupervised: associations Clustering Bidding/Feightliner	Chapter 12 & 6	Homework 5
10	Feature selection & creation Leakage Click / Netflix		
11	Deployment & Decomposition Managing & Hiring Mailing	Chapter 13 & 14	
12	Project Presentation		Project Writeup