Syllabus
Data Mining for Business Analytics - Managerial
INFO-GB.3336, Spring 2018

Course information

- When: Mondays and Wednesdays 3-4:20pm
- Where: KMEC 3-65

Professor

- Manuel Arriaga
- Email: marriaga@stern.nyu.edu
- Web: http://pages.stern.nyu.edu/~marriaga
- Office: KMEC 8-59
- Office Hours: By appointment

Teaching assistant

- Liam Greenamyre
- Email: ltg245@stern.nyu.edu
- Office hours: TBA

Course Overview

The goal of this course is to give you a solid understanding of the opportunities, techniques and critical challenges in using data mining and predictive modeling in a business setting.

This course will provide you with hands-on experience using a variety of real-world datasets. We will pay special attention to how we can best understand and translate business challenges into data mining problems. So that you can develop that ability, in our lectures we will cover the major issues involved in knowledge discovery and decision making as well as core technical concepts and machine learning methods. Our discussion of these more technical aspects will be carried out without getting into their mathematical underpinnings.

If you are interested in a deeper, more technical perspective and have some programming experience, consider taking Data Science for Business Analytics – Technical [INFO-GB.2336] instead.

This course doesn't promise to turn you into a data scientist (although this may happen anyway!). It is meant to make you **literate** in data science, which means you will be comfortable doing some hands-on work (albeit not at scale), interacting with and managing data scientists as well as evaluating data science proposals from a business standpoint.

Prerequisites

The course does not have any prerequisites.

Learning Goals

There are two primary and two secondary learning goals associated with this course:

(i) **Critical and Integrative Thinking:** specifically, how do you formulate business problems in terms that make them amenable to being solved through a systematic modeling approach. Formulation is key as is the construction and evaluation of the model. This skill is also essential as a manager tasked with evaluating the proposals, progress, and work outputs of data science teams.

(ii) **Modeling:** you should be competent in applying basic statistical and machine learning methods to data. Your modeling expertise should be sufficient for you to manage data science teams.

(iii) **Effective Oral Communication:** Each student shall be able to communicate verbally in an organized, clear, and persuasive manner, and be a responsive listener. You will have the chance to demonstrate communication skills via a presentation of your term project.

(iv) **Interpersonal Awareness and Working in Teams:** Students will submit a project which may entail working in a small group (2-4 people) and must apportion tasks appropriately and submit a quality product in a timely manner.

Self-learning is a particularly important part of this course. You will get the best value from this course if you experiment actively with ideas and explore ideas instead of just coming to class and expecting to be told what works and what doesn't. There's nothing like learning by doing. Accordingly, 35% of the grade is assigned to your project. So, **start early**. Exploratory work always takes longer than you think. Indeed, your very first assignment is to write a 1-2 page summary of what you might do as your project. Even if you end up changing topics, the exercise will help you get started in thinking about it seriously, before you get into the nitty-gritty of the quantitative exercises.

Reading materials

The textbook for this course is:

"Data Science for Business: What you need to know about data mining and data analytic thinking" by Provost & Fawcett (O'Reilly, 2013)

In the readings section of this syllabus, any reference to "chapters" without any additional information refers to chapters from our textbook.

We will also read some chapters of an old data mining book: **Seven Methods for Transforming Corporate data Into Business Intelligence, Vasant Dhar and Roger Stein, Prentice-Hall (1997)**. These chapters will be shared through NYU Classes. In the readings section of this syllabus, readings from this book can easily be identified by the prefix "DS".

Finally, additional reading materials will also be made available through NYU Classes.

Software

The key concepts and methods discussed in this course are not specific to any piece of software. However, for the assignments and hands-on practice we will use Weka, an open-source, multi-platform data mining toolkit:

https://www.cs.waikato.ac.nz/ml/weka/

Weka is a well-established, highly popular data mining application. For that reason, it has the added benefit of it being easy to find abundant documentation, how-to videos and Q&A threads online. The "official" go to source is known as the "Weka book:"

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

All individual assignments must be done in Weka.

For your final project, you are welcome to either use Weka or explore other tools. The latter route will probably appeal to the more technically minded among you, in particular when considering tools such as R or Python's SciKitLearn library.

Requirements and grading

Given the nature of the material we will be covering, it is expected that you attend all sessions and do not arrive late. There is a strong "cumulative" aspect to the structure of this course, as is often the case when discussing more technical material.

There will be five assignments, each of which builds on a previous one. These will be "front loaded" so you get most of them over with in the first half of the semester which should give you time to spend on your term project**. Assignments will be due by the beginning of our Wednesday class (3pm)**. You must turn in all assignments on the dates they are due.

The project is the most important component of the course and gives you a chance to "do your own thing." **Start early**. You can do the project in groups of 2 to 4 people. Completing the project entails two deliverables – a project proposal and final report – as well as delivering an in-class presentation at the end of our course.

There is no final exam.

The grade breakdown is as follows.

Assignments: 55 points
Term project: 35 points
Class participation and attendance: 10 points

Term project

The term project should be a substantial piece of work that (i) involves the use and application of techniques learned in this course and, just as importantly, (ii) is of interest to you. Most projects fall in one of the following categories (these are just examples, not an exhaustive list of what is accepted):

a) **An original idea that you want to build on and test.** Examples:
   - Is it possible to extract useful "sentiment" information from news? If so, how?
   - Build and evaluate a machine learning-based trading strategy based on high frequency data.

b) **Replication/extension of an existing study or result.** Example:
   - Past research shows that boosting and bagging result in variance reduction: we compare these methods on 20 standard datasets from the UCI database and demonstrate under what conditions they work best.

c) **Extension of an assignment**. Example:
   - In Assignment 5 we considered an "imbalanced" class problem. We consider 20 imbalanced class problems and evaluate the impacts of oversampling the majority class.

d) **Applying a data-driven approach to a core business problem within your organization (must at a minimum include preliminary results and a detailed proposal for further analysis)**.

You will present your project in the last two sessions of the semester, so make sure you start on it early and give a polished presentation!

Timeline (subject to small revisions)

Please note: assignments are always due by the beginning of our second class of each week (i.e., Wednesday 3pm).

Week	Topic(s)	Readings	Assignments
Week 1 (starts Jan 29)	What is the course about? What is predictive analytics? The data mining process	Chap 1 & 2	Assignment 1 handed out
Week 2 (starts Feb 5)	Predictive modeling in action Introduction to Trees Software installation & demo	Chap 3 & 4	Assignment 1 due Assignment 2 handed out
Week 3 (starts Feb 12)	More trees; logistic regression and support vector machines Model performance analysis 1: evaluation and validation	Chap 5	Assignment 2 due Assignment 3 handed out
Week 4 (only Feb 21)	Overfitting and its avoidance Model performance analysis 2: ROC, lift, MSE, etc.	Chap 7 & 8	Assignment 3 due Assignment 4 handed out
Week 5 (starts Feb 26)	Text as data Bayesian modeling and the Naïve Bayes approach	Chap 9 & 10	-
Week 6 (starts Mar 5)	Connectionism: Neural networks and deep learning	DS Chapter 6	Assignment 4 due Project proposal due Assignment 5 handed out
SPRING BREAK			
Week 7 (starts Mar 19)	Similarity, clusters and neighbors	Chapter 6	Assignment 5 due
Week 8 (starts Mar 26)	Crowds of predictive models Boosting and Random Forests	Reading on website	
Week 9 (starts Apr 2)	Evolutionary approaches and genetic algorithms	DS Chapter 5	
Week 10 (starts apr 9)	Prediction and Noise revisited How to evaluate data science proposals	Chap 11 & 13	
Week 11 (starts Apr 16)	Topic TBD		
Week 12 (starts Apr 23)	Guest industry speakers		
Week 13 (starts Apr 30)	Term project presentations		Final project report due by May 7