1. Course Overview

The goal of this course is to give you a deep understanding of the opportunities, techniques and critical challenges of using data mining and predictive modeling in a business setting. Such a skill is valuable analytic skill in its own right, but also useful in managing data science teams.

The course be a hands-on experience using a variety of real-world datasets, with an emphasis on problem formulation. This includes the ability to understand and translate business challenges into data mining problems and cover in depth the major issues involved in knowledge discovery and decision making. Formulation also requires a core understanding of core technical concepts and machine learning methods which are covered in this course without getting into their mathematical details.

I created this course in 1994. I had just formed a new “Data Mining” group at a major Wall Street bank with the “mission” of extracting business value through Machine Learning methods using the large amounts of transaction and market data being collected by the bank. The obvious applications were in the customer and product arenas and in improving operations. These have become mainstream applications in data science along with others where predictive analytics have become possible due to the increasing volume and variety of data. Data mining is now core to every business. Data Science is the new discipline that has emerged through our experiences with data mining over the last two decades. “Big Data” is the new challenge.

What has not changed is the need to formulate problems intelligently and to be able to design the right solutions to business problems quickly. There is still some art involved in this exercise that we shall also consider in the course.

---

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 broadly speaking, means you should be comfortable with doing hands-on work albeit not at scale, and comfortable interacting with data scientists and getting the most out of them. You should also be capable of evaluating data science proposals from a business standpoint.

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

2. 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 solution via a systematic modeling approach. Formulation is key as is the construction and evaluation of the model. This thinking is also essential as a manager, to evaluate proposals, progress, and work outputs of data science teams.

II. **Modeling:** you should be competent in applying basic statistical and machine learning approaches 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 (two 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 actively 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, 40% 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.

3. Teaching Materials
The textbook for the course is:

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

http://data-science-for-biz.com/

I will also post a few chapters from an old data mining book by Dhar and Stein (referred to as DS in the course schedule):
Seven Methods for Transforming Corporate data Into Business Intelligence, Vasant Dhar and Roger Stein, Prentice-Hall (1997).

In addition, readings are posted on the website that should be read prior to the session for which they are assigned.

4. Software
This course is software agnostic. For the assignments we will use Weka, which is an open source data mining toolkit. The documentation for this software is online and also in the "Weka book:"


You will also have the option of using Salford Predictive Miner (SPM), which provides the basic data mining methods with an easy to use interface. Documentation on SPM is available online.

You should also feel free to use open source software libraries on your own, especially for the project. The SciKitLearn library has a powerful set of tools in Python as does the R language. However, no formal support is provided for these software tools although I’ll try my best to help if you go down the Python route.

5. Requirements and Grading
It is imperative that you attend all sessions, especially since the class meets infrequently, and the sessions build on previous discussion. Please do not arrive late.

This is the most important component of the course and gives you a chance to “do your own thing.” But start early. The assignments are “front loaded” and largely done midway through the course which should give you time to focus on your term project. You can do the project individually or jointly with another person. You must hand in a formal project proposal during the fifth week of class.

There will be a total of five assignments. You must turn in all assignments on the dates they are due.

There is no final exam. The grade breakdown is as follows.

i. Assignments: 50 points
ii. Term paper on a trading strategy: 40 points
iii. Class participation and attendance: 10 points

6. Focus and interaction
The course will explain through lectures, discussions, 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. Also, if you have the question, someone else may too and everyone may benefit from the answers being available on NYU Classes. Please try to answer your classmates’ questions as appropriate.

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.
### 6. Timetable

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Reading/Preparation (posted on BB)</th>
<th>Submission/Handout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb 6</td>
<td>Introduction and Objectives&lt;br&gt;What is predictive analytics?&lt;br&gt;The data mining process</td>
<td>Dhar, V., Data Science and Prediction, CACM December 2013.&lt;br&gt;Chap 1 &amp; 2</td>
<td>Assignment 1 handed out</td>
</tr>
<tr>
<td>Feb 13</td>
<td>Predictive modeling in action&lt;br&gt;Introduction to Trees</td>
<td>Chapter 3 and 4&lt;br&gt;DS Chapter 10</td>
<td>Assignment 1 due&lt;br&gt;Assignment 2 handed out</td>
</tr>
<tr>
<td>Feb 20</td>
<td>PRESIDENT’S DAY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 27</td>
<td>More trees and logistic regression&lt;br&gt;Model performance analysis 1: evaluation and validation</td>
<td>Chapter 5</td>
<td>Assignment 2 due&lt;br&gt;Assignment 3 handed out</td>
</tr>
<tr>
<td>Mar 06</td>
<td>Overfitting and its avoidance&lt;br&gt;Model performance analysis 2: ROC, lift, MSE, etc</td>
<td>Chapter 8</td>
<td>Assignment 3 due&lt;br&gt;Assignment 4 handed out</td>
</tr>
<tr>
<td>Mar 13</td>
<td>MIDTERM BREAK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar 20</td>
<td>Crowds of predictive models: Are they better than individuals?&lt;br&gt;Boosting and Random Forests</td>
<td>Reading on website</td>
<td>Assignment 4 due</td>
</tr>
<tr>
<td>Mar 27</td>
<td>Connectionism: Neural networks and deep learning&lt;br&gt;Clustering and nearest neighbor</td>
<td>DS Chapter 6 Reading on website</td>
<td>Formal project proposal due Assignment 5 handed out</td>
</tr>
<tr>
<td>Apr 03</td>
<td>Bayesian modeling and the Naïve Bayes approach</td>
<td></td>
<td>Assignment 5 due</td>
</tr>
<tr>
<td>Apr 10</td>
<td>Text as Data</td>
<td>Chapter 10 Reading on website</td>
<td>Assignment 6 handed out</td>
</tr>
<tr>
<td>Apr 10</td>
<td>Evolutionary approaches and Genetic Algorithms</td>
<td>DS Chapter 5</td>
<td>Assignment 6 due</td>
</tr>
<tr>
<td>Apr 17</td>
<td>Prediction and Noise revisited&lt;br&gt;How to evaluate data science proposals</td>
<td>Dhar HBR article. Chapters 11 and 13</td>
<td></td>
</tr>
<tr>
<td>Apr 25</td>
<td>Industry invited speaker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 01</td>
<td>Student project presentations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 08</td>
<td>Student project presentations</td>
<td></td>
<td>Final projects are due within one week of this session</td>
</tr>
</tbody>
</table>