Data Driven Decision Making

D3M Fall 2014

Instructor
Vishal Singh
911 Tisch Hall
212-998-0405 Email: vsingh@stern.nyu.edu

Contact

Office hours
Tuesday 1.00-3.00 or by appointment
Course Information

“Every two days we now create as much information as we did from the dawn of civilization up until 2003”
Eric Schmidt (CEO Google).

“Data are widely available; what is scarce is the ability to extract wisdom from them”
Hal Varian (UC Berkeley and Chief Economist, Google).

Motivation: The two quotes above summarize the main theme of this course. In every aspect of our daily lives, from the way we work, shop, communicate, or socialize; we are both consuming and creating vast amounts of information. More often than not, these daily activities create a trail of digitized data that is being stored, mined, and analyzed by firms hoping to create valuable business intelligence. For example, customer transaction databases provide vast amounts of high-quality data that can allow firms to understand customer behavior, and customize business tactics to increasingly fine segments or even segments of one. However, much of the promises of such data-driven policies have failed to materialize because managers find it difficult to translate data into actionable strategies. The general objective of this course is to fill this gap by training you with tools and techniques to analyze large databases and by instilling an intuition for DDDM.

Course Philosophy: Extracting useful insights from the vast amount of information involves a combination of analytical skills and intuition. It is both an art and science. The pedagogic philosophy in this course embraces the principle of learning-by-doing. Each concept that we cover has a software implementation and a problem or case whose resolution can be enhanced through use of the data. Statistical tools covered in the class will range from simple data analysis and visualization, to advanced regression and multivariate statistics. Our emphasis will be on applications and interpretation of the results for making real life business/policy decisions. Beyond what is necessary, we will focus less on the mathematical and statistical properties of the techniques used to produce these results.

In order to provide a broad intuition of the concepts and methods, we will use data/problems/example/case-studies from different fields such as Finance & Economics, Psychology & Sociology, Politics & Public Policy, and Medicine & Biology. Often this will involve using data and replicating results from major academic journals in these fields. However, since this is primarily a Marketing course, emphasis will be given to quantitative aspects of marketing decision making such as segmentation, estimating market potential and forecasting demand, developing optimal pricing policies, designing/positioning new products, data mining, and customer relationship management (CRM).

Objectives: Regardless of your chosen field or major, it is virtually impossible to survive in the professional world without a working knowledge of basic data analysis and use of some statistical software. The course is designed to train you in a wide spectrum of quantitative
problems that you are likely to encounter in your workplace. Some of the quantitative methods and concepts are fairly advanced and may seem intimidating at the beginning. Regardless of your prior background, an objective of the course is to remove any fear of data analysis, and provide you with the toolkits to become an accomplished empirical analyst. More generally, I hope to instill a general analytical intuition that enables you to analyze and comprehend contemporary issues such as the presidential elections, housing bubble and ensuing financial crises, debates on healthcare, government deficits, or climate change. In other words, become an educated consumer of news, issues, and challenges facing the society.

The specific objectives of this course are to:

1. Understand how analytical techniques and statistical models can enhance decision making by converting data to information and insights;
2. Provide intuition for data driven decision making by using practical examples from a wide spectrum of fields;
3. Provide insights on how to choose and use the most effective statistical tool based on the problem at hand;
4. Provide you with a software tool kit that will enable you to apply statistical models to real decision problems;
5. Most importantly, increase your comfort level with analyzing large databases to translate conceptual understanding into specific operational plans – a skill in increasing demand in the business world.

**Prerequisites:** Although there are no specific prerequisites for the class, an introductory class in statistics/regression and working knowledge of MS Excel would be helpful. However, the most important prerequisite for the class is a positive attitude towards learning.

**Course Organization**

**Textbook:**
There are no required text books. We will rely on:

- (1) Your old statistics/regression text (if you have one) +
- (2) Free online resources (I will point out where to find them for each topic) +
- (3) Extensive notes (Combination of ppt & pdf documents that I will provide)

**Software:**
A number of statistical packages have built in capacities to execute statistical procedures we need (and do so efficiently on large datasets): SAS, SPSS, STATA, R, Minitab, and so forth. In my own research I tend to use SAS (for data cleaning), R (for analysis/graphics), and
MATLAB (for advanced models—not required for most business applications you will encounter). The two most widely used in the business world are SAS and SPSS, although R is getting popularity. For this class, we will be using a combination of R and Tableau (BI Software). However, if you already know any particular software well, feel free to use that. I can provide help with SAS/R/Stata/SPSS. From your perspective, it is important to pick one statistical package and master that. This involves developing skills to load and manipulate large data files and run various statistical methods covered in the class. In addition, we will use a number of freebees from Google such as Fusion Tables & Google Trends. Finally, I will be introducing Tableau as a BI tool which provides ready access to terrific visualizations with virtually no coding requirement.

Course Website:
All relevant material related to the course will be posted on NYU Classes. Reading material, cases, and class notes will be made available at least a week before they are needed.

Basis for Final Grade:
There are two components to the final grade: (1) Assignments/Case studies (75%), and (2) Take-home final exam (25%). All exercises for the assignments/cases/final are application-oriented and incorporate extensive use of R, Tableau, and EXCEL. They are drawn from three primary sources:

- Proprietary data from the business world (mostly used in my research),
- Publically available data from various businesses (e.g. Google), government (e.g. Census, BLS), international agencies (e.g. UN/World Bank)
- Published research in the top academic journals from various fields. Many (good) journals require authors to publish the data used in their research. We will replicate the findings from some of the most influential papers from various fields for which data are available.

All assignments/cases are open book, open notes, open internet. We will mimic the “real” world where your objective is to solve the problem at hand and make recommendations based on your analysis. I will NOT penalize you if you can find a solution to any exercise on the internet. On the contrary, I will point out (and provide links to) the original sources of data for all assignments/cases.

Assignments: There will be several Quantitative exercises during the course of the semester (approximately one every week). The objective of the assignments is to provide you with a working knowledge of the tools and techniques commonly used in the industry. We will
learn how to summarize and visualize data; execute advanced statistical models; and interpret how the output can be used for decision making. We should think of these assignments as learning a new language--they form the backbone for longer case studies.

**Case studies:** These go a step beyond simply executing models. Instead, they are designed to challenge you to (1) Understand the problem at an intuitive level, (2) Use simple data analysis and visualization to verify (or falsify) your intuition, (3) Use appropriate statistical analysis to present your arguments. In order to imitate the real life challenges, the case studies are fairly open-ended and provide little step-by-step instructions.

**Group Work:** For all assignments/cases, you are free to work individually or as a group. Please formulate groups yourself, with a maximum of 3 students per group. It is important to note that the material covered in the class can only be understood by hands-on work. Hence, working in a group should not imply division of labor--between say (analysis/writing) or (Q1/Q2). Assignment questions (or sub-parts of bigger case studies) are carefully chosen to illuminate broader conceptual issues that arise during empirical work. Hence, we should treat group work primarily as an opportunity to learn & teach each other. Once every individual in the group has understood the problem, please turn in a single document with names of all group members. Although groups in the class are self-selected, membership is ‘non-binding’. If at some point you feel that learning can be enhanced by working alone or group dynamics are not working out, feel free to ‘divorce’ the group.

**Due Date, Submission, and Grading:** All assignments/cases will be handed out on Wednesday and are to be turned in electronically via NYU Classes by Noon the following Tuesday. The objective is to have sufficient time to grade your assignments and point out means of improvement in class on Wednesday. From time to time, I will assign students to lead discussion of the assignments/cases on Wednesday (day after the due date).

All assignments should be submitted as a PPT converted to a PDF document. We will use minimal writing in the class, beyond bullet point arguments. Our job will be to illustrate the points using clean tables and high impact data visualization.

**Final Exam:** There will be a take-home Final Exam which will be handed to you well in advance. The exam will cover short exercises pertaining to each topic covered and will be similar in spirit to the case studies/exercises covered in the class. The final Exam is open book, open notes, open internet but “closed-friends”.

**Questions about the Assignments/Cases/Exam:** All questions pertaining to the assignments/exam need to be posted on the discussion board on BB rather than a personal
email to me (more on this in the class). This allows for smooth flow of information to everyone and ensures that no student is at a disadvantage. To encourage you get started early, the last chance to raise questions with me will be in class on Monday (day before the due date). You are free to post queries till the last minute but I will stop answering questions approximately 24 hours before the due date. After that, you will have to rely on the generosity of your class mates for hints. I encourage discussions on any topic (including assignments/cases) as long as it happens in an open forum.

**Class Participation:** This is a social component of the class and there are no grades allocated to class participation. However, a substantial part of the benefit that you will derive from this course is a function of your willingness to expose your viewpoints and conclusions to the critical judgment of the class, as well as your ability to build upon and critically evaluate the judgments of your classmates. You are strongly encouraged to share articles/videos on any topic that you find interesting and voice your opinions. We will use this as an opportunity to implement the scientific approach to decision making on contemporary issues in the media.

**Laptop Use:** Majority of the topics/methods will require use of laptop computers during the class sessions. It may not appear so but it is obvious to the instructor and your fellow students when your computer screen has Facebook rather than the class notes. Please be courteous to people around you and the educational institute by refraining from text messaging, FB, etc.

**Feedback:** Some of the material covered in the class is fairly advanced. Regardless of your current comfort level with data/technology/statistics, it is my objective to make sure that every student gets a good grasp of the concepts, methods, and implementation (literally ‘no child left behind’). If at any time you feel falling behind in class, please contact me. I am happy to work with you individually or in a group to catch up. However, please note that it is your responsibility to seek help.

It is my goal is to make this an excellent course. I encourage you to provide feedback on any issue that can enhance your learning and progress.
Methodological Topics

Topic 1: An intuitive introduction to data-driven decision making
We will begin the course with a general introduction on what we mean by data driven strategy and why it is important. We will use several examples and mini-case studies to illustrate the role of statistical analysis in decision making. These lectures will provide an intuition for ‘data linguistics’.

Topic 2: Basic Data Analysis & Intro to R/Tableau
In this session we will discuss various types of data that are commonly collected by firms. What methods to use and what inferences/insights can be obtained depend on the type of data that are available (stated versus revealed preference, level of aggregation, cross-sectional, time series, panel data and so forth). We will cover some of the nuts and bolts of preparing data for analysis, and use several mini-cases to review some basic yet extremely useful techniques such as data visualization, frequency distributions, mean comparisons, and cross tabulation. Statistical inferences using chi-square, t-test and ANOVA will be discussed. We will also look at the basics of the R & Tableau.

Topic 3: Experimental Design and Natural Experiments
Experimental designs are often regarded as the "gold standard" for making causal or cause-effect inferences. We will discuss the issues of design of experiments and internal and external validity. Several case studies in marketing, economics, and medicine that range from controlled lab and field experiments, to circumstances that provide us with “natural” experiments will be discussed using hands-on implementation.

Topic 4: Opinion Polls and Survey Based Analysis
Survey research is an important tool to assess attitudes and opinions on a wide range of issues. It is a fairly common form of data you will encounter in the industry. It is used in a wide array of fields such as politics, marketing research, and social psychology. We will briefly discuss issues of survey design and sampling, but focus primarily on analysis of survey data using examples from a variety of industries/topics such as customer satisfaction, debate on health care reform, and politics. Appropriate use of descriptive statistics (what's going on in our data) and inferential statistics (how to make inferences from our data to general population) will be discussed.

Topic 5: Regression Analysis
In this topic we will turn our attention to the relationships among variables. Regression is by far the most useful tool for analyzing relationships between a phenomenon of interest (independent variable) and one or more predictor variables. We will spend a fair amount of time on regression and its applications. Emphasis will be on use of regression output in forecasting, elasticity analysis, and various applications such as promotional planning and optimal pricing.

Topic 6: Advanced Regression Models (the dreaded log makes an appearance)
This session covers some important aspects of regression modeling including measures to control for seasonality and trend, capture non-linear effects, interactions, and use of appropriate functional forms (linear, semi-log, log-log).

Topic 7: Discrete Choice Models
Typical regression analysis is suitable when the dependent variable is continuous (e.g automobile sales, price of crude oil, stock prices). Often we encounter situations where the phenomenon of interest (i.e. your dependent variable) is discrete (e.g. vote or not, buy or don’t buy). In these circumstances, use of linear regression may be inappropriate. This class will discuss Logit models that are appropriate for discrete choice analysis.
Topic 8: Database Marketing/Data Mining/CRM
It is often thought that the value of a firm can be computed using the metric of life time value of its customer base. This topic will cover the important and growing area of CRM and customer equity. We will discuss various tools in database/direct marketing used to model customer acquisition and retention. Analytical tools to compute customer lifetime value (CLV) will be discussed. We will also cover few extremely useful techniques for data mining/reduction:

(a) **Step-wise** function in regression/logit: A useful (but somewhat dangerous) modification for regression analysis we would have already covered. Used as initial/quick analysis when faced with very large number of variables.

(b) **Cluster Analysis**: Hosts of algorithms that allow “grouping a set of objects in such a way that objects in the same group are more similar (in some metric) to each other than to those in other groups. It is used in many fields, including data mining, machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics” (from Wikipedia)

(c) **Factor analysis** is a “method used to describe variability among correlated variables in terms of a potentially lower number of latent factors. Factor analysis originated in psychometrics, and is used in behavioral sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data” (from Wikipedia)

Topic 9: Text Mining and Sentiment Analysis
Text analytics refers to the process of deriving insights from text. Perhaps the most widely known is Google’s n-gram (frequency of words and phrases) found in over 5.2 million books digitized by Google Inc. Several useful packages exist in R for text mining. We will use applications from several areas such as analyzing the media, using API from Facebook and Twitter to conduct sentiment analysis.
Examples of Topics for Assignments/Cases/Final

Brand Equity: How to quantify the value of a brand?

Optimal Pricing: I know! MR=MC. Ok but how do I go about actually doing this? Demand estimation, competitive reactions, and optimal pricing.

Presidential Election & Political Marketing: Can we do better than political pundits?

Population Growth: Forecasting the human population growth. Will it stabilize? What are the environmental implications?

Merger Analysis: Should FTC allow Continental and United to merge? Too late, but what are the implications for prices that consumers pay?

Google Trends and Web Traffic: Deep insights from seemingly trivial data

Budget deficits/ Social Security/ Medicare: Why are the projections so different between Congressional Budget Office (CBO) and the “think tanks”?

College Tuition and the Education system: Why you need to make the best of every moment spent at NYU.

Housing bubble & the Financial Meltdown: Hindsight 20/20, but how could anyone miss this?

Customer Lifetime Value (CLV) & Database Marketing: NPV analysis for millions of customers!

Development Economics (from Bottoms Up): Replicating cool field experiments

Fat Tax: Is taxing junk food a good public policy?

Geography: Of voting and brand preference

Sentiment in Tweets/FB Posts/& Blogs