COLUMBIA UNIVERSITY

IN THE CITY OF NEW YORK

Decision Risk and Operations

Advanced Business Analytics **Fall 2015**

Course Description

Business Analytics is about information– turning data into action. Its value derives fundamentally from information gaps in the economic choices of consumers and firms. Analytics unlocks this hidden value.

Consider consumers. In economic theory, they make choices to maximize their total utility. That's all fine, but to truly optimize one's choices - to discover what's best among the vast array of alternatives available is hard. Doing it well requires a lot of information and the ability to sort through that information to decide effectively. The result is that consumer decision-making is highly imperfect in practice and a great deal of potential value goes unrealized as a result. Business analytics helps capture this latent value by improving our choices as consumers: Google's search engine provides fast, highly relevant web content, enhancing the value of your time online.

Google Maps and Waze help you uncover a faster route home. Pandora's recommendation system makes it simple to find and enjoy the music you love; Netflix does the same for movies and television. Amazon makes it easy to discover great products to buy and streamlines the purchase transaction. Tripadvisor helps you create great vacations, while Okcupid finds you the perfect mate (well, at least a suitable one!). Each of these technologies enables us to capture value that would go unrealized without the help of data and analytics.

Firms benefit from analytics too. While economic theory suggests firms organize themselves to maximize the value they deliver, real companies frequently waste resources, fail to understand customer needs and hence engage in unproductive activities. Data and analytics help them overcome these inefficiencies. Capital One uses analytics to match credit card offers to customers more accurately than their competition. WalMart uses analytics to manage its inventory in a way that allows it to serve its customers reliably at exceptionally low cost. Cleveland Clinic uses analytics to provide more targeted treatment of patients and to fine-tune therapies that produce better health outcomes at lower costs. Axioma uses analytics to construct portfolios that provide better risk-reward trade-offs for their clients. In each case, data and analytics helps these firm uncover new opportunities to focus their efforts on value-adding activities and hence increase the gap between the value they deliver and their cost to serve.

This course is intended for GSB students that have taken the introductory Business Analytics course offered by the division of Decision Risk and Operations. The basic course presented predictive and prescriptive analytics tools in the context of business cases, with an emphasis on the challenges that can arise in implementing analytical approaches within an organization. This course goes beyond the basic analytic course in several ways. First, the course will cover the tools and theory at a deeper level. Second, the course will be based on R. R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners. In particular, we will use R-studio, a free and open source integrated development environment for R. Students should install R and R-studio prior to the beginning of the course.

Schapiro Room 822 Mail Code 4704 500 West 120th Street New York, NY 10027 212-854-2942 Fax 212-854-8103

The course emphasizes that business analytics is a practical discipline which requires mastery of both methodology and business applications. The concepts learned in this class should help you identify opportunities in which business analytics can be used to improve performance and support important decisions. It will teach you important tools that can be used to transform data into high-impact business decisions. Lastly, it should make you alert to the ways that analytics can be used — and misused — within an organization.

Course topics include a review of basis statistical ideas, numerical and graphical methods for summarizing data, linear regression, logistic regression, subset selection, cross-validation, classification, decision trees, factor models, clustering, support vector machines, and other emerging data analytics methods. The course presents real-world examples where a significant competitive advantage has been obtained through large-scale data analysis. Examples include advertising, eCommerce, finance, health care, marketing, and revenue management. The ultimate goal is, of course, help to make better business decisions using advanced analytics.

Instructor

Prof. Guillermo Gallego

Office Hours
Tuesdays: 3-4pm

Office: CEPSR 822 (8th floor)
Office Phone: 212 854 2935
Email: gmg2@columbia.edu

Textbooks and Learning Materials

There is *no required* textbook: all class materials will be available on the Canvas website. However, some books are very useful if you want to learn more and deeper about data analytics. The best way to *learn is by doing* (especially with programming). All of the books below have free on-line versions.

Optional Textbook 1 (*highly recommend*, easy following with many examples and data sets): **Data Mining and Business Analytics with R**, by Johannes Ledolter; Publisher: Wiley (2013), ISBN-13: 978-1118447147. A pdf file of the book can be downloaded by following the link: http://www.nataraz.in/data/ebook/hadoop/Data_Mining_and_Business_Analytics_with_R__Johannes_Ledolter.pdf

Optional Textbook 2 (solid primer, with theory and explanation): **An Introduction to Statistical Learning with Application in R**, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; Publisher: Springer (2013); ISBN-13: 978-1461471370. The book can be downloaded from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

Optional Textbook 3 (a great advanced text): **Elements of Statistical Learning: Data Mining, Inference, and Prediction,** by Trevor Hastie, Robert Tibshirani and Jerome Friedman (2009); ISBN 978-0-387-84858-7. This book requires some mathematical sophistication and goes beyond the material we will be covering. The book is **free** at **http://statweb.stanford.edu/~tibs/ElemStatLearn/index.html**

Software:

- We require the **R** Statistical Software, which is **powerful** and **free**. R can be downloaded at the link below: http://www.cran.r-project.org/
- **Rstudio** is a free platform for both writing and running R, available at www.rstudio.org. Some students find it friendlier than basic R.
- We do not assume that you have used R in a previous class. I will provide limited software instruction, in-class demonstration, and code to accompany lectures and assignments. However, **this is not a class on R**. Like any language, **R is only learned by doing**. You should install R as soon as possible and familiarize yourself with basic operations.
- Students can become proficient in a few weeks. Some manuals are very helpful to learn R, e.g., http://cran.r-project.org/manuals.html
- Additional resources: (a) Tutorials at data.princeton.edu/R are fantastic (and there are many others out there). (b) Youtube intros to R, e.g. the series from Google Developers.

Canvas Site

A Canvas site is set up for this course. Each student is expected to check the site throughout the semester as Canvas will be the primary venue for outside classroom communications between the instructors and the students.

Attendance Policy

Attendance and class participation are part of each student's course grade. Students are expected to attend all scheduled class sessions. Failure to attend class will result in an inability to achieve the objectives of the course. Regular attendance and active participation are required for students to successfully complete the course.

You are expected to come to class prepared, and ready to discuss the pre-class reading, case or assignment questions. More details on the assignments will be provided on Canvas. Class participation is an important part of learning. If you have a question, it's likely that others do as well. I encourage active participation, and course grades will take into account students who make particularly strong contributions.

Assignments

There will be three *group assignments* (1-4 students). All homework assignment should be submitted through the Canvas links.

- Assignment 1. Due session A5
- Assignment 2. Due session A8
- Assignment 3. Due session A12

Group Project: 3-5 students form a group and work on the projects as a team. Students can identify a company or a scenario, collect data, use techniques taught in class to study the data patterns or to

predict future outcomes. Students are required to write a 4-6 page project report, and present in class using Power Point slides.

- Project proposal due session A6
- Project presentations due session A-makeup

Final Exam: the final exam is a *take home individual written exam.* It will be given on session A12. The final exam is due by midnight December 20.

Late submission including assignments, projects and exams will *not* be accepted.

Study Group (not required, but highly recommend)

Many students learn better and faster when working in a group, so I encourage collaborative learning. The study groups can be different from your project groups.

Evaluation and Grading

Assignment		**Weight**
Attendance and participation in class discussion		10%
Homework		30%
Project		20%
Final Exam		40%
Total		100%

Syllabus at a Glance

- Session A1 (September 11): Motivation, course description, Introduction to R
- Session A2 (September 12): Statistical learning, bias-variance tradeoff; logistic regression
- Session A3 (September 25): Linear and quadratic discrimination analysis; knn
- Session A4 (September 26): Multiple linear regression
- Session A5 (October 9): Cross-validation and the bootstrap
- Session A6 (October 23): Subset selection, ridge and lasso regressions
- Session A7 (October 24): Principal component analysis and partial least squares.
- Session A8 (November 6): Class canceled. Will make up class on December 18.
- Session A9 (November 21): Difference-in-difference. Propensity score matching
- Session A10 (December 4): Tree-based methods
- Session A11 (December 5): Support vector machines
- Session A12 (December 12): Simulation and Optimization in R
- Session A-makeup (December 18): Project presentations

COLUMBIA UNIVERSITY

IN THE CITY OF NEW YORK

Decision Risk and Operations

Schapiro Room 822 Mail Code 4704 500 West 120th Street New York, NY 10027 212-854-2942 Fax 212-854-8103