15.071: The Analytics Edge, Sections A and B

Faculty: Robert Freund and John Silberholz

General Information and Syllabus

Spring 2017

Description
The amount of data available to organizations has been growing as never before, and companies and individuals who harness this data through the use of data analytics gain a critical edge in their business domain. In this class we examine how data analytics is used to transform businesses and industries, using examples and case studies in e-commerce, healthcare, social media, high technology, sports, the internet, and beyond. Through these examples and many more, we demonstrate the use of analytics methods such as linear regression, logistic regression, classification trees, random forests, text analytics, social network analysis, time series modeling, clustering, and optimization.

Prerequisites:
15.060: *Data, Models and Decisions,* or a basic statistics and a basic optimization course. Please contact the course instructors with questions about appropriate prerequisites.

Readings/Resources
All of the suggested readings are from the book *The Analytics Edge* by Dimitris Bertsimas, Allison O'Hair and William Pulleyblank, Dynamic Ideas LLC, 2016. We refer to the book below as the "AE book." The AE book is a good resource that complements the lecture material, and it is strongly recommended. Likewise, the readings are recommended but are not required.

We will also post a copy of the "Analytics Edge R Manual" on the Stellar course site.

Grading
Your course grade will be composed of the following:

1. Homework Assignments and Cases: 45%.
2. Final Course Project: 40%.
3. Class Participation: 15%.

By definition, class participation will be subjectively evaluated (see below).

Much of your education will take place outside the classroom, as you study, review, and apply the topics to which you are introduced in class.

Assignments:
There will be eight individual homework assignments, and a final project that should be done in teams of two. The following are tentative topics and due data for the homework assignments:

- Wednesday, February 22: Data analysis and linear regression in R
- Wednesday, March 1: Logistic Regression
- Wednesday, March 8: CART and Random Forests
- Wednesday, March 15: Internet Advertising and Optimization
- Wednesday, April 12: Demand Chain Management and Optimization
- Wednesday, April 19: Social Network Analysis
- Wednesday, April 26: Collaborative Filtering
- Wednedsay, May 3: Clustering, and Advanced Prediction Methods

All homework assignments are due at the start of class on the date assigned.

Final Project:
For the final project, by **March 16**, each team needs to submit a one-page proposal that outlines a plan to apply analytical methods to a problem the team has identified using some of the concepts and tools discussed in the course. The proposal should include a description of: (1) the problem, (2) the data that you have or plan to collect to solve the problem, (3) which analytic techniques you plan to use, and (4) the impact or overall goal of the project (say, if you could build a perfect model, what would it be able to do?). The teaching team will be available to answer questions over email, and will provide all students with electronic feedback by March 25.

In the week of **April 17**, each project team will set up a meeting with a member of the teaching team to show your progress in applying the analytical methods to your project topic. This meeting is intended to help you make progress on your project.

The final project submission will consist of a written report of at most 4 pages (not including appendices) that describes your analysis, as well as a 15-minute presentation (in powerpoint or pdf format) of your project. Unfortunately, due to time constraints, we will not be able to have all student teams present in class. However, ALL TEAMS will need to be *prepared* to give a 15-minute presentation on May 15 or May 17, and all teams are required to submit their presentation for a grade.

To determine who will present on May 15 and May 17, by midnight on Thursday **May 11,** each team will electronically submit (a) a 1-page abstract summarizing their project (including the scope and idea of the project, what analytical methods/models were used, and what results were obtained), and (b) the slide presentation. The abstracts will be uploaded to the class website. Students will vote by the end of the day on Sunday, May 14 about which projects they would like to see presented in class. The teaching team will vote as well (taking the abstracts and presentations into account), and the presenters will be notified in real-time during class on May 15 and May 17.

The four-page report (not including appendices) that describes your analysis is due on **May 17.**

INDIVIDUAL Work Assignments

All homework assignments are INDIVIDUAL work assignments. While you may find it useful to discuss broad conceptual issues and general solution procedures with others, the final product that you turn in must be done individually. What you turn in must be your own product, written in your own handwriting, or in a computer file of which you are the sole author. Copying another's work or electronic file is not acceptable. Although you may discuss your work with other students, what you turn in must represent your own work. You are expected to adhere to the following standards:

- Do not copy all or part of another student's work (with or without "permission").
- Do not allow another student to copy your work.
- Do not ask another person to write all or part of an assignment for you.
- Do not work together with another student in order to answer a question, or solve a problem, or write a computer program jointly.
- Do not consult or submit work (in whole or in part) that has been completed by other students in this or previous years for the same or substantially the same assignment.
- Do not use print or internet materials directly related to a case/problem set unless explicitly authorized by the instructor.
- Do not use print or internet materials without explicit quotation and/or citation.
- Do not submit the same, or similar, piece of work for two or more subjects without the explicit approval of the two or more instructors involved.

The violation of the policy on individual work is a serious offense, and suitable consequences include grade reduction, an F grade, a transcript notation, delay of graduation, or expulsion from MIT Sloan.

The objective here is to learn. In our experience, the material of this class is best learned through individual practice and exposure to a variety of application contexts.

Class Participation and Conduct

Your class participation will be evaluated subjectively, but will rely upon measures of punctuality, attendance, and the relevance/insight of class participation. Your class participation will be judged by what you add to the class environment, regardless of your technical background. In general, questions and comments are encouraged.

Consistent with Classroom Values@MITSloan, we have the following policies:

- Students are expected to arrive promptly and be ready for class to start on time and to stay for the entire class.
- Laptops, e-pads, and smartphones are not to be open in the classroom.
- Cell phones and PDAs are not to be used or permitted to ring in the classroom.
- Students are expected to attend all classes.
- Maintenance of a professional atmosphere by using respectful comments and respectful humor.

- Refraining from distracting or disrespectful activities (e.g., avoiding side conversations and games).
- Courtesy towards all participants in the classroom.
- Observance of the most conservative standards when one is unsure about which norms apply.

Please refer to the Values@MITSloan materials for more details. Violations of Values@MITSloan policies will be marked. Three or more violations will result in an automatic penalty of a letter grade.

We ask that you use a name card for all class sessions.

Professors:

	Robert Freund	John Silberholz
Office:	E62-567	E62-569
Phone:	617 253-8997	617 324-4118
Email:	rfreund@mit.edu	josilber@mit.edu

Course Homepage:
The homepage for the course is accessible through the stellar class management system:

https://stellar.mit.edu/S/course/15/sp17/15.071AB/index.html

Lectures:

Section	**Lecture**		
Section A	MW	1:00-2:30	E51-345
Section B	MW	2:30-4:00	E51-345

Recitations:

Section	**Recitation**		
Section A	Friday	1:00-2:00	E51-345
Section B	Friday	2:00-3:00	E51-345
Advanced R	Friday	3:00-4:00	E51-149

Recitations will consist of interactive sessions that will cover additional examples of the analytics methods presented in the lectures, and -- most importantly – recitations will be used to show how to create models in R. Recitation attendance is highly recommended. All recitations are run by the Teaching Assistants.

For students who want to learn R at a more advanced level, we are offering this year an optional recitation (in addition to the regular session recitations) called the "Advanced R" recitation. This is in addition to the regular section recitation every week.

Teaching Assistants:

Section	TA	Email	Office hours (E51-242)
Section A	Chris McCord	mccord@mit.edu	Under construction
Section A	Yee Sian Ng	yeesian@mit.edu	Under construction
Section B	Colin Pawlowski	cpawlows@mit.edu	Under construction
Section B	Leon Valdes	lvaldes@mit.edu	Under construction
"Float"	Sebastian Cubela	scubela@mit.edu	Under construction

Course instructors and TAs are also available by appointment (and by email).

Tentative Syllabus for 15.071: The Analytics Edge, Sections A and B
Spring, 2017
(subject to changes as we refine the material)

DATE	TOPIC	SESSION
Wednesday, February 8	Introduction, Google's Search Engine, and R	1
Monday, February 13	Predicting Wine Quality	2
Wednesday, February 15	Improved Prediction of Wine Quality	3
Tuesday, February 21	Predicting Loan Defaults	4
Wednesday, February 22	Customer Retention and Churn Prediction	5
Monday, February 27	Predicting Medical Costs	6
Wednesday, March 1	Making Intelligent Parole Decisions	7
Monday, March 6	Predicting Click-Thru-Rates for Online Advertising	8
Wednesday, March 8	People Analytics at Google	9
Monday, March 13	Adwords and Internet Advertising	10
Wednesday, March 15	Twitter Sentiment Detection (Text Analytics)	11
Monday, March 20	NO CLASS DAY – SIP week	--
Wednesday, March 22	NO CLASS DAY – SIP week	--
Monday, March 27	NO CLASS DAY – Spring Vacation	--
Wednesday, March 29	NO CLASS DAY – Spring Vacation	--
Monday, April 3	Predicting Sales Volumes	12
Wednesday, April 5	Supply/Demand Chain Management	13
Monday, April 10	Social Networks Analysis	14
Wednesday, April 12	Community Detection in Networks	15
Monday, April 17	NO CLASS DAY—Student Holiday	--
Wednesday, April 19	Netflix and Collaborative Filtering	16
Monday, April 24	Clustering	17
Wednesday, April 26	Advanced Methods for Prediction/Analysis	18
Monday, May 1	Matching: from Medical Students to Organ Exchanges	19
Wednesday, May 3	Fantasy Sports	20
Monday, May 8	Analytics for Designing Drug Therapies, or Revenue Management with Demand Forecasting	21
Wednesday, May 10	IBM Watson	22
Monday, May 15	Student Project Presentations	23
Wednesday, May 17	Student Project Presentations	24

15.071: The Analytics Edge, Sections A and B
Spring 2017
Outline and Assignments
(subject to changes as we refine the material)

All readings are suggested but not required. Readings of the text refer to the textbook *The Analytics Edge* by Dimitris Bertsimas, Allison O'Hair, and William Pulleyblank, Dynamic Ideas LLC, 2016. We refer to the book below as the "AE book."

All materials are either in the text or will be posted on Stellar during the term.

1. February 8, 2017 Lecture 1 – Introduction, Google's Search Engine, and the software/program R

In the first lecture, we will illustrate the scope of modern business analytics by describing the ideas behind Google's search engine, and how these simple analytics ideas have enabled the worldwide web to revolutionize the way we do business, gather information, and interact with one another. We will also illustrate the versatility and power of the software R.

2. February 13, 2017 Lecture 2 – Predicting Wine Quality

We will review linear regression and discuss how linear regression can be used to predict the quality of wine. The suggested readings for this lecture are the first section of Chapter 1 of the AE book, titled "Predicting the Quality and Prices of Wine," and the first section of Chapter 21 of the AE book, titled "Linear Regression."

3. February 15, 2017 Lecture 3 – Improved Prediction of Wine Quality

We will learn about categorical variables and see how they can be used to greatly improve the prediction of wine auction prices.

4. February 21, 2017 Lecture 4 – Predicting Loan Defaults

(NOTE: This class is on Tuesday due to President's Day) We will discuss how analytics is used to model the probability that a loan will default. Through this example, we will introduce the methodology of logistic regression. The suggested reading for this lecture is the second section of Chapter 21 of the AE book, titled "Logistic Regression".

5. February 22, 2017 Lecture 5 – Customer Retention and Modeling of "Churn" in Telecom

Customer retention in the telecom industry is critical for revenue. "Churn" refers to customers who switch carriers to get better deals. We will show how analytics is used to

identify churn likelihoods for customers, and how firms use these analytics models to manage customer retention.

6. February 27, 2017 Lecture 6 – Predicting Medical Costs

We will explore how to predict medical expenses for millions of patients based on their previous years' expenditures, illnesses, medical conditions, and other patient data. We will introduce a new prediction tool called CART (Classification And Regression Tree) and we will compare and contrast its capabilities with linear regression. The suggested readings for this lecture are the third section of Chapter 21 of the AE book, titled "CART and Random Forests."

7. March 1, 2017 Lecture 7 – Making Intelligent Parole Decisions

We discuss how analytics is used to make more informed, objective, and defensible prisoner parole decisions. Through this example, we will discuss how CART can be extended for the task of classification (as opposed to regression). We will use data for parole cases to build CART models for predicting parole violations.

8. March 6, 2017 Lecture 8 – Predicting Click-Thru Rates for Online Advertising

We discuss the critical role of analytics in predicting Click-Thru Rates (CTRs) for online advertising. Through this example we will discuss the prediction method Random Forest, which is an extension of CART. We will then use data from sponsored search ads to build Random Forest models that predict ad CTRs. We will also compare and contrast the various prediction methods we have been using thus far: linear regression, logistic regression, CART, and Random Forests.

9. March 8, 2017 Lecture 9 – Guest lecturer Prasad Setty: People Analytics at Google

We will have a guest lecture by Prasad Setty, Vice President, People Analytics and Compensation at Google, who will discuss analytics for human resource management at Google.

10. March 13, 2017 Lecture 10 – Adwords and Internet Advertising

We will discuss how analytics is used to choose online advertising impressions on websites (as well as how analytics is used by companies to develop their online advertising strategies). The suggested reading for this lecture is Chapter 12 of the AE book.

11. March 15, 2017 Lecture 11 – Twitter Sentiment Detection (Text Analytics)

We discuss how tweets on the social networking site Twitter can be used to understand public perception and analyze sentiment. We will use this example to introduce the method of text analytics, which we will use to predict the sentiment of tweets about Apple.

NO CLASS from March 20 – March 31 due to SIP week and Spring Break.

12. April 3, 2017 Lecture 12 – Predicting Sales Volumes

We will discuss how analytics is used to predict retail sales, which is a critical step in the inventory/ordering decision-making process used by retailers. Through this example, we will discuss time series modeling, which we will then use to predicting sales volumes in a practical setting.

13. April 5, 2017 Lecture 13 – Supply/Demand Chain Management

We will discuss how to combine forecasting and optimization modeling for supply/demand chain management in a large consumer package goods retail company.

14. April 10, 2017 Lecture 14 –Social Network Analysis

We will discuss how analytics is used to evaluate the structure of social networks, which is an important task for many companies. We will present many important social network concepts including centrality and closeness. We will demonstrate how these concepts are used to better understand customers (as well as employees).

15. April 12, 2017 Lecture 15 – Community Detection in Networks

We discuss how to discover communities with network data, which then provides actionable information about the structure of a network. We show how these concepts are used to segment customers and to cluster products, with obvious implications for targeted advertising.

16. April 19, 2017 Lecture 16 – Netflix and Collaborative Filtering

We discuss recommendation systems and the Netflix prize competition. We discuss collaborative filtering as a method for estimating user preferences based on the preferences of many other users. We apply collaborative filtering and other prediction methods to the problem of estimating the preferences for movies among users.

17. April 24, 2017 Lecture 17 – Clustering for Customer Segmentation

We discuss the importance of customer segmentation in a variety of data-mining settings. We present two types of clustering methods – k-mean clustering and hierarchical clustering – for customer segmentation. The suggested reading for this lecture is the fourth section of Chapter 21 of the AE book, titled “Clustering.”

18. April 26, 2017 Lecture 18 – Advanced Methods for Prediction, Forecasting and Digital Recognition

We will discuss advanced methods for prediction, forecasting, and digital recognition.

19. May 1, 2017 Lecture 19 – Matching: from Medical Students to Organ Exchanges

We will discuss two matching problems. The first is the National Resident Matching Program, wherein medical students are efficiently assigned to residency programs. The second is organ patient/donor matching in organ exchanges, with a particular emphasis on paired kidney exchanges.

20. May 3, 2017 Lecture 20 – Fantasy Sports Gambling

We discuss how to gain an edge in fantasy sports gambling using analytics. In particular we show how an MIT Sloan research team has been using analytics modeling to win in fantasy sports leagues (and donate their winnings to charity).

21. May 8, 2017 Lecture 21 – Analytics for Designing Drug Therapies, or possibly Revenue Management with Demand Forecasting

We will discuss how the prediction and optimization models covered thus far can be combined to aid in the design of clinical trials for drug therapies. We will show how databases of the results of clinical trials for drug therapies for advanced cancers can be used to predict the outcomes of proposed clinical trials before they are run and to design new therapies that combine multiple drugs. Alternatively, we will discuss how optimization is used in airline revenue management, and how airlines harness the power of analytics to create a competitive edge. The suggested reading for this lecture is Chapter 17 of the AE book.

22. May 10, 2017 Lecture 22 – IBM Watson

We will discuss how IBM built a computer that could beat the best human players at Jeopardy, a game known for testing human knowledge and reasoning. The suggested reading for this lecture is Chapter 3 of the AE book.

23. May 15, 2017 Lecture 23 – Student Project Presentations

During this class, selected students will give 15-minute presentations of their projects.

24. May 17, 2017 Lecture 24 – Student Project Presentations

During this class, selected students will give 15-minute presentations of their projects.