15.071: The Analytics Edge

General Information and Syllabus

Fall 2020

Instructors:
Bart Van Parys

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 teach and 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, optimization, and machine learning.

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

There is no required textbook for the course. However, we have some suggested readings from *The Analytics Edge* by Dimitris Bertsimas, Allison O'Hair and William Pulleyblank, Dynamic Ideas LLC, 2016. You can order the book directly from the publisher at https://www.dynamic-ideas.com/books/kgsni67q285zmpdit17ix1cwbnuqac. We refer to this book as the "AE book" below.

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

Grading

Your course grade will be composed of the following:

1. Homework Assignments and Team Case:
   a. HW 1-4: 22.5%
   b. HW 5-7 and Team Case: 30%.
2. Final Course Project: 40%.
3. Class Participation: 7.5%.

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 seven individual homework assignments and a team case that should be done in teams of three. The following are tentative topics and relevant dates/times for the homework assignments:

Number	Topic	Publication Date	Due Date
HW0 (part 1)	Setting up R	Sept. 1	Sept. 4
HW1	Linear Regression	Sept. 4	Sept. 18
HW2	Logistic Regression	Sept. 12	Sept. 25
HW3	Generalized Linear Models/Non-linear Regression	Sept. 19	Oct. 2
HW4	CART, Random Forests	Sept. 26	Oct. 9
HW5	Boosting, Regularization, Clustering	Oct. 3	Oct. 16
HW6	Text Analytics, Collaborative Filtering	Oct. 10	Oct. 23
HW0 (part 2)	Setting up Julia	Oct. 27	Oct. 30
HW7	Optimization	Oct. 31	Nov. 13
Team Case	Dartboard	Nov. 7	Nov. 20

All homework assignments must be submitted to Canvas by the due date/time as a single pdf file. If you need to draw pictures or write equations by hand, you can include them in your report.

Final Project

Milestone	Activity/Deliverable
October 3	By this date, 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?).
October 16	The teaching team will be available to answer questions over email and will provide all students with electronic feedback by this date
Week of ***October 19***	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.
Thursday, ***November 27***	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) a 15-minute presentation (in PowerPoint or pdf format) of your project.
November 29	The project abstracts will be uploaded to the class website and each student will vote on which projects they would like to see presented in class (since, unfortunately, due to time constraints, we will not be able to have *all* student teams present in class). The teaching team will vote as well (taking the abstracts and presentations into account).
December 4	Each team will electronically submit their final project report. This will consist of at most 4 pages of analysis and conclusions presented as a managerial memo but you can include appendices with supporting information.
December 5&8	ALL TEAMS should come to class *prepared* to give their 15-minute project presentation. The "winning" presenters will be notified in real-time during class.

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.

Instructor:

	Bart Van Parys
Email:	vanparys@mit.edu
Office Hours:	Friday 4:00-5:00pm

Course Homepage:
The homepage for the course is accessible through Canvas.

Lectures:

Days	**Time**
TS	2:00-03:30pm

Recitations:

Days	**Time**
Th	6.30-7.30pm

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 and use models in R. Recitation attendance is not mandatory it is very highly recommended. All recitations are run by the Teaching Assistant.

Teaching Assistants:

TA	**Email**	**Office hours**
Ryan Cory-Wright	ryancw@mit.edu	Wednesday 4-6pm

Course instructors and TAs are also available by appointment.

Fall 2020
Outline and Readings
(tentative)

Note: readings are suggested but are not required. All text readings refer to 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."

All relevant materials will be posted on Canvas during the term.

1. September 1, 2020 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. September 5, 2020 Predicting Wine Quality

We will review linear regression and discuss how linear regression can be used to predict the quality of wine. We will also review categorical variables and see how they can be used to greatly improve the prediction of wine auction prices. 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. September 8, 2020 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. The suggested reading for this lecture is the second section of Chapter 21 of the AE book, titled "Logistic Regression".

4. September 12, 2020 Auto Insurance and Smart Cities

We will discuss extensions to the linear regression models. In particular, we will discuss the use of other link functions than the identity function used in linear regression and the popularity of generalized linear models in insurance. We will also explore the use of non-linear regression models in the context of smart cities, using data on NYC taxicab rides.

5. September 15, 2019 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 Trees) 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."

6. September 19, 2020 Making Intelligent Parole Decisions

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 in order to make more informed, objective decisions in the criminal justice system. The suggested readings for this lecture are the third section of Chapter 21 of the AE book, titled "CART and Random Forests."

7. September 22, 2020 Predicting Click-Thru Rates for Online Advertising

We will 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.

8. September 26, 2020 Regularization

We will discuss how to handle datasets with a large number of variables (also called dimensions), how to systematically select variables for the purpose of avoiding overfitting and creating simpler, more interpretable models. We will use the example of diabetes prediction to motivate the discussion

9. September 9, 2020 Ensemble Learning and Boosting

We will discuss the use of many (an ensemble of) models to improve model performance without overfitting. Random Forests (ensemble of trees) were an example of the concept we already covered. We will discuss boosted trees and the XGBoost algorithm.

10. October 3, 2020 Clustering for Customer Segmentation

We will discuss the importance of customer segmentation in a variety of 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."

11. October 6, 2020 Text Analytics and Sentiment Detection

We will 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 analyze Donald Trump's tweets.

12. October 10, 2020 Netflix and Collaborative Filtering

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

October 13, 2020 Columbus Day (Monday's Schedule)

13. October 17, 2020 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 both externally and internally. We will present important social network concepts including centrality and closeness. We will demonstrate how these concepts are used to better understand customers (as well as employees).

14. October 20, 2020 Introduction to Neural Nets

We will discuss the basic notions involved in the design of neural networks, including input layers, hidden layers, output layers, activation functions, and training the network. These will be illustrated online auctions.

15. October 24, 2020 Using X-Ray Images to Improve Medical Diagnoses

The development of image acquisition devices is one of the key ongoing transformations in the health care industry. The rapid growth in medical images provides opportunities to supplement analysis by trained professionals with data-driven algorithms to identify diseases. Medical image processing has benefitted greatly from recent developments in deep learning, a machine learning method based on neural networks. In this lecture, we will cover the main deep learning algorithms, and illustrate how they can provide an edge in medical image processing.

16. October 27, 2020 Optimizing World Food Program Operations

We will discuss an important type of optimization called network flow optimization. This will be illustrated using the example of optimizing the operations of the World Food Program. The suggested reading for this lecture is Chapter 12 of the AE book.C

17. October 31, 2020 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.

18. November 3, 2020 Fairness and Bias in Prediction and Machine Learning

We will present the challenges of identifying bias and unfairness in prediction and machine learning models. We will discuss several ways that model and/or data bias can be reduced.

19. November 7, 2020 Optimizing School Bus Routes and School Start Times

Large public-school systems like Boston's use fleets of hundreds of buses to transport thousands of students to hundreds of schools on a daily basis. We will first discuss how to optimize these operations, which led to $5 million in yearly savings for Boston public schools. We will then discuss how the problem could be relaxed in order to find the "best" school starting times. We will introduce the framework of multi-objective optimization to aid decision-makers by eliciting the complex trade-offs.

November 10, 2020 No Class

20. November 14, 2020 Capacity Planning in Electricity Markets

Electricity markets require adequate generation capacity to meet customer demand at all times. Capacity planning involves constructing capacity given a number of physical and regulatory constraints. A critical complexity, however, is that future customer demand cannot be forecasted with perfect accuracy. We will introduce methods of stochastic optimization to solve this problem of decision-making under uncertainty in a way that that balances the considerations arising in high-demand scenarios and low-demand scenarios.

21. November 17, 2020 Experiments/Course Wrap-Up and Lessons Learned

- In many real-world settings, historical data (to feed the model-building process) may not be available or it may be of inadequate quality. In these situations, conducting experiments can generate datasets that are ideal for modeling, and for directly answering important business questions. We will discuss how to design, execute and analyze experiments.
- We review the techniques we learned over the course of the semester and offer guidelines and recommendations for successfully putting them to work in the real world.

November 21/24, 2020 Thanksgiving Holidays

22. December 1, 2020 Dartboard Case

This lecture will feature a case study on demand-supply management at a large consumer package goods retail company. It will illustrate how to bring together predictive and prescriptive principles in order to improve decision-making and solve complex business problems. We will also discuss how to communicate insights derived from analytics methods and how to work collaboratively with different stakeholders to implement analytics-based solutions.

23. December 5/8, 2020 Student Project Presentations

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