Columbia Business School

# B8143: Foundations of AI for Business

Course information and syllabus

Prof. Hongseok Namkoong

- This course was previously circulated as Business Analytics II. **This is a standalone class with no prerequisites.**
- Due to recent curricular redesign in the program, if you have taken **Prof. Guetta**'s "IEOR E4650 Business Analytics" in Spring / Summer of 2021, you will not be able to enroll in this class. If you have taken the same course taught by other faculty (e.g., Prof. Elmachtoub), you may still enroll in this course.
- Need help? Email cbs.ba2@gmail.com. This email address will reach every course staff member. There is a system set up to ensure we don't miss your email—we get a lot of emails—and respond promptly, but it will only work through this email address.
- Note that there is pre-work required before the first class. Make sure to check Canvas a week before class begins.

## 1. Course Description

Business analytics (a.k.a. AI for business) refers to the ways in which enterprises such as businesses, non-profits, and governments use data to gain insights and make better decisions. Business analytics is applied in operations, marketing, finance, and strategic planning among other functions. Modern data collection methods——arising in bioinformatics, mobile platforms, and previously unanalyzable data like text and images——is leading an explosive growth in the volume of data available for decision-making. The ability to use data effectively to drive rapid, precise and profitable decisions has been a critical strategic advantage for companies as diverse as WalMart, Google, Capital One, and Disney. Many startups are tackling how to apply analytics on large databases. With the increasing availability of information—so-called "Big Data"—business analytics is a critical capability for enterprises of all types and all sizes.

Even though you may not program analytics solutions yourself, you will need to have a deep understanding of the technology in order to leverage it effectively. This course develops a critical understanding of modern analytics / AI methodology, studying its foundations, potential applications, and—perhaps most importantly—limitations. You will apply these methods to new cases in a broad range of industries.

- You will learn how to employ analytics methods in more unstructured and diverse situations, on complex real-life datasets, and on a broader range of structured and unstructured data (such as text data).

- We will cover substantially complex yet flexible methodologies for predictive analytics. For example, you will learn about random forests, which is one of the most powerful analytics method to date.
- The course introduces a framework that will help you translate business needs and priorities into analytic problems.
- You will learn the language required to communicate with data scientists and other engineers implementing these models.
- We will develop a holistic and critical perspective to analytics and AI systems. You will learn how to recognize and communicate the inherent limitations of a predictive model, and how to design more reliable analytics systems.

This course emphasizes that the discipline is not theoretical; we will apply these new methodologies in a number of cases, and use them to develop increasingly powerful insights and predictive capabilities. Many of the techniques we will be covering are now considered standard in industry, and developing a good understanding of them will deepen your ability to identify opportunities in which business analytics can be used to improve performance, drive value, and support important decisions. For those of you who will work closely with data science and product teams, the deep knowledge we will develop in this class will prove invaluable.

## 2. Pre-work; before class begins

Before class begins, you will be required to install a BA2 add-in for the class, prepare for our first case, and complete a short survey. Details will be posted on Canvas. Anyone who has not completed the pre-work at least three days before class begins may be removed from the class.

Attendance for the first class is *strongly recommended* as we will be familiarizing ourselves with the add-in which we will use during the rest of the class.

## 3. Detailed Class Plan

Due to the advanced nature of the material covered in this class, we will focus on quality over quantity, with a strong focus on making sure you understand the concepts in depth before we move on. The class will be divided into four modules. The first three modules focus on the foundational technical aspects of data analytics, studying how algorithms can leverage large datasets to impact decisions across medicine, banking, law, policy-making, and engineering applications (e.g., autonomous vehicles). In the final module, we develop a holistic and critical viewpoint to analytics and AI systems by studying the infrastructure they build on and their varied failure modes.

- **Module 1:** Introduction

  In this module, we introduce the BA2 Excel add-in. We review linear regression, including advanced topics including dummy variables for categorical data, interactions, and data

standardization. We introduce the bias-variance trade-off, a fundamental concept in Business Analytics, and cross-validation, a key tool for model selection. We finally apply K-fold cross-validation to linear regression

Detailed module plan:

- Introduction
- A quick review of linear regression, including dummy variables, interactions, and data standardization
- Causal inference vs. predictive analytics
- A review of overfitting
- The bias-variance tradeoff
- K-Fold cross-validation
- ***Case:*** Analyzing Performance in the New York City Public School System

- **Module 2:** Powerful Predictions; Boosted Trees and Random Forests

  In this module, we will introduce one of the most powerful, versatile, and popular predictive analytics tools used by businesses today —boosted trees and random forests. Both comprise of many smaller and simpler models called classification and regression trees, which are weak individually but reinforce each other to produce highly predictive models. These approaches are particularly well suited to problems with many variables. Along the way, we will also discuss decision trees, ensemble models, and model interpretability.

  Detailed module plan:

  - Introduction to decision trees: the shortcomings of linear regression and magic of non-parametric models
  - Trees and the bias-variance tradeoff
  - Ensemble models—boosting and bagging
  - Boosted decision trees
  - Random forests
  - Model interpretability and variable importance
  - ***Case:*** Data Driven Investment Strategies for Peer-to-Peer Lending (Lending Club)

- **Module 3:** Text Analytics

  One of the most impactful ways the data landscape has changed over the last decade is the availability of large-scale unstructured data as well as structured data. Chief among these are textual data. From financial disclosure statements to tweets and news articles, there is an enormous amount of text data now available electronically, and many companies are realizing there are valuable insights to be gleaned from this mass of data.

  Unfortunately, valuable as these data might be, they are more difficult to analyze than structured data. In this module, we will study techniques that can be used to extract meaning and value from textual data.

  Detailed module plan:

  - Unstructured data and text analytics

  - ***Case:*** Evisort: An A.I.-Powered Startup Uses Text Mining to Become Google for Contracts
  - The bag of words representation
  - Text data for predictive analytics
  - Topics modeling: Latent Dirichlet Allocation

- **Module 4:** A Critical Look at Analytics and AI

  In the final module, we develop a holistic understanding of analytics and AI technology as an engineering system borne out of economic, social, and political forces. Just like any engineering system, a predictive model builds on intangible and material—often capital-intensive—infrastructure such as human labor, computing servers, organization, and natural resources. We begin by studying how analytics and AI systems are made end-to-end, from organization and data-collection to deployment and operation. Using the technological foundations developed earlier in class, we put into context how the AI revolution in the past decade enabled such efforts.

  Then, we turn our attention to the potential pitfalls of such systems. Technologically, analytics and AI methods can be highly unreliable. Performance of predictive models severely degrade on tail events (black swans), changes in user behavior, adversarial attacks, strategic behavior, changes in the environment, and underrepresented populations. As datasets embody interests of capital and social relations of the world, predictive models built on this infrastructure produce and replicate power structures and associated inequities. We will discuss how to manage, communicate, and mitigate these limitations.

  Detailed module plan:

  - A holistic perspective to the modern analytics pipeline
  - The AI revolution: an infrastructural view of analytics and AI systems
  - Failure modalities
  - Embodiment of social, economic, and political power structures in analytics models
  - Quantitative methods for addressing reliability issues in analytics
  - Qualitative approaches to recognizing the values and interests replicated in analytics systems and preventing their harms

## 4. Course Materials

There is no required textbook for the class. There will be cases and slides, which will all be posted on canvas. For those of you looking for additional reading, the following two resources are particularly good.

- *Data Science for Business*, by Foster Provost and Tom Fawcett. This book is pitched at the MBA level, and covers many of the topics we will be covering in this class. It is excellent, but does not go into quite as much depth as we will.
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. This is the *bible* of machine learning,

written by some of the greatest innovators in the field over the last 20 years or so. It is, however, extremely mathematical, and therefore will be out of reach to most MBAs. That said, if you have a particularly quantitative background and want to dive in *much* greater depth into any of the topics in this class, this is the place to go.

## 5. Requirements and Grading

Before class begins, you will be required complete some pre-work; see section 3 for details.

The class itself will be graded as follows. **Please see Canvas for the due dates for each of these components, and for due dates for the pre-class work required for each module.**

**Final exam (40%)** : The final exam will be multiple choice. It will *not* require the use of Excel or of a computer. A practice final will be provided on canvas.

**Homeworks (35%)** : There will be three homeworks, one for the first three modules. Each homework will be based on a real-world application of the techniques in this class, and will require you to use the tools we will be learning in this class.

Your homework grade will be calculated as

$$\frac{\text{HW1 grade} + \text{HW2 grade} + \text{HW3 grade}}{3}$$

Data science is difficult, and I would be doing you a disservice if I made the homeworks easy. As such, be warned – *these homeworks are designed to be difficult.* To make things fair, therefore, I will *not* grade these homeworks based on correctness – instead, I will grade them based on effort, understanding, and execution on a scale of 1 to 6 using the following rubric:

- **0 points**: no significant effort.
- **2 points**: some questions tackled; evidence some analysis was carried out on the data, but perhaps not correctly.
- **4 points**: all questions tackled; evidence some analysis was carried out on the data, but perhaps not correctly.
- **6 points**: all questions tackled and submitted in a clear, well-presented, and easy-to-follow report clearly explaining the logic behind the steps you took.
- **8 points (extra credit)**: outstanding work, not only answering the questions in the homework and meeting the requirements for 6 points, but also carrying out *further* investigations based on the data given. Homeworks completed correctly in Python would merit this grade.

Note that each of these rubric descriptions require excellence in modeling *and* exposition/ presentation.

**Attendance and participation (25%)** : Your attendance and participation score will be calculated as follows:

**Punctuality (25%)** : this part of your grade will be calculated by finding the fraction of classes you arrive at *exactly on time* and *with your nameplate* (if in person), or for which your absence is excused.

**Attendance (25%)** : this part of your grade will be calculated by finding the fraction of classes you attend or for which your absence is excused. You can get these points even if you show up slightly late, or without your name plates.

**PollEverywhere (25%)** : this part of your grade will be calculated by finding the fraction of PollEverywhere questions you participate in (note: you do *not* need to answer these correctly to score these points – just to participate).

**Contributions in class (25%)** : this part of your grade will be calculated based on my impressions, and on your participation in ad-hoc assignments such as the pre-class work.

For classes your attend on zoom, the first two parts of your grade will be self-reported.

Please note that I am *very* generous with excusing absences – for any reason – provided you let me and the TAs know *at least an hour before* class.

## 6. Software

This course will require the use of Excel, and we will provide an add-in called XLKitLearn, which we have developed to extend the functionality of Excel to cover the topics in this follow-up elective. This add-in should work on a Mac natively, without the need for a virtual machine. You will be installing this add in on your computer as part of the pre-work for the class.

Even though this course only requires you to use Excel, the add-in itself will be powered by Python code. Python has quickly become the lingua franca of business analytics, and those hoping to enter analytics-related industries will likely carry out further study to deepen their knowledge of this programming language. Every run of the add-in will output the equivalent Python code you would need to run to get the same result so you can implement these methods in Python if you like.

Solutions to all the Homeworks and in-class cases will be provided in the add-in *and* in Python. You are welcome to complete the homeworks using *either* tool, but **absolutely no Python is required to completed this class**.

## 7. Inclusion, Accommodations, and Support for Students

At Columbia Business School, we believe that diversity strengthens any community or business model and brings it greater success. Columbia Business School is committed to providing all students with the equal opportunity to thrive in the classroom by providing a learning, living, and working environment free from discrimination, harassment, and bias on the basis of gender, sexual orientation, race, ethnicity, socioeconomic status, or ability.

Columbia Business School will make reasonable accommodations for persons with documented disabilities. Students are encouraged to contact the Columbia University's Office of Disability Services for information about registration. Students seeking accommodation in the classroom may obtain

information on the services offered by Columbia University's Office of Disability Services online at www.health.columbia.edu/docs/services/ods/index.html or by contacting (212) 854-2388.

Columbia Business School is committed to maintaining a safe environment for students, staff and faculty. Because of this commitment and because of federal and state regulations, we must advise you that if you tell any of your instructors about sexual harassment or gender-based misconduct involving a member of the campus community, your instructor is required to report this information to a Title IX Coordinator. They will treat this information as private, but will need to follow up with you and possibly look into the matter. Counseling and Psychological Services, the Office of the University Chaplain, and the Ombuds Office for Gender-Based Misconduct are confidential resources available for students, staff and faculty. "Gender-based misconduct" includes sexual assault, stalking, sexual harassment, dating violence, domestic violence, sexual exploitation, and gender-based harassment. For more information, see http://sexualrespect.columbia.edu/gender-based-misconduct-policy-students.