Columbia Business School

Business Analytics 2: Foundations of AI
Business Analytics 3: Modern AI, Deep Learning, and Generative AI

Note that there is pre-work required before the first class – details are below, and on Canvas.

Business analytics 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 – are 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 based on the application of AI & analytics to large databases. With the increasing availability of broad and deep sources of information – so-called "Big Data" – business analytics are becoming an even more critical capability for enterprises of all types and all sizes.

AI is beginning to impact every dimension of business and society. In many industries, you will need to be literate in AI to be a successful business leader. The Business Analytics sequence is designed to prepare you to play an active role in shaping the future of AI and business. You will develop a critical understanding of modern analytics methodology, studying its foundations, potential applications, and – perhaps most importantly – limitations.

This sequence comprises two classes. **Business Analytics 2** covers the foundations of AI and analytics in business. We will primarily discuss structured or tabular datasets, and discuss how predictive analytics can be used with these datasets to drive business value. **Business Analytics 3** delves into more modern techniques such as regularized learning, deep learning, generative AI, and large language models. We will discuss how to handle enormous datasets with thousands of columns, and even datasets with no columns at all, such as images and text.

BA2 is a "soft" pre-requisite for BA3 in the sense that the classes are designed to be taken in sequence. However, in some circumstances, it may make sense for students to take BA3 only – for example, if a student has previous experience of AI. Feel free to reach out to the instructors for guidance in that respect.

Much as Business Analytics does, this sequence 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.

This sequence will not require any coding, and will not require any prior knowledge other than your core Business Analytics and Statistics classes.

Pre-work

Before class begins, you will be required to install an 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 will be removed from the class.

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.

Business Analytics 2: AI Foundations

- **Module 1 – Introduction**: in this module, we begin by asking the fundamental question "What is AI?" We discuss the history of the field, and discuss its relationship with machine learning, optimization and deep learning. We discuss the difference between structured and unstructured data, and give an overview of the techniques we will be covering in the class. Finally, we introduce xlkitlearn, an add-in which we will use throughout BA2 and BA3.
- **Module 2 – Overfitting Revisited**: this module reviews overfitting as covered in Business Analytics, and then delves deeper by introducing the bias-variance tradeoff, a fundamental concept in Business Analytics, and cross-validation, a key tool for model selection. We use k-nearest neighbors to illustrate the pitfalls of overfitting, and to illustrate the importance of cross validation.
- **Module 3 – Decision trees**: this module introduces decisions trees, a simple but powerful technique for predictive analytics. We will implement decision trees in the context of several business cases, and show how important overfitting and the bias-variance tradeoff is, even in the context of these simple models.
- **Module 4 – Boosted Trees**: this module introduces ensemble models, which comprise large collections of smaller models. We will begin with *boosted trees*, which are made up of many small decision trees that reinforce each other to produce highly predictive models.
- **Module 5 – Random Forests**: this module introduces a second kind of ensemble model called a *random forest*. These models are made up of many large decision trees that are highly overfit, but smooth each other out. The result is one of the most versatile and commonly used predictive models in industry. We also discuss model interpretability – the art of understanding how large, complex models produce their results.
- **Modules 6 – AI Implementation**: the final module of BA2 will discuss the many challenges that arise when AI is implemented in practice. We will develop a holistic understanding of AI as a complex engineering system borne out of economic, social, and political forces. Just like any engineering system, a predictive model builds on infrastructure such as human labor, computing servers, natural resources. They are also

highly dynamic systems, which evolve and need to be monitored over time. We will explore these complexities. Finally, we will turn out attention to the potential pitfalls of such systems – their potential unreliability, especially on tail events (black swans), their susceptibility to changes in user behavior and adversarial attacks, and their degraded performance in the context of underrepresented populations. We will discuss how to manage, communicate, and mitigate these limitations.

Business Analytics 3: Modern AI, Deep Learning, and Generative AI

- **Module 1 – The Lasso**: Business Analytics 3 focuses on unwieldly datasets. This first module looks at datasets with an enormous number of columns – so many, in fact, that classical models fail on them. In a world of big data, these datasets are everywhere – from social media data to internet logs. We introduce a modern technique called the Lasso, which modifies linear regression to enable it to handle these datasets. We begin with the concept of "skill versus luck", which you covered in Business Analytics, and explain how it can be applied to linear regression to produce a regression model that automatically selects the best variables to use.
- **Module 2 – The Case of Cambridge Analytica**: this module explores a scandal that still makes headlines today – the case of Cambridge Analytica. Using real but anonymized social media data, we replicate some of the analyses Cambridge Analytica claimed to carry out, understand what they were trying to achieve, and discuss whether they were successful. The Lasso will be essential in solving this problem, because of the enormous dimensionality of the data concerned. We will also discuss some of the consequences of the data explosion we live in, and the resulting impact on privacy.
- **Module 3 – The Deep Learning Revolution**: this module begins our exploration of the most modern and exciting models around today – deep neural networks. We will discuss the need for these models, their basic architectures, and using the case of a company revolutionizing the recycling industry, we will demonstrate the way they can automatically extract meaning from largely unstructured datasets. We will discuss the training of such models using gradient descent. We will build a mini neural network in Excel, and then use a technique called transfer learning to create our own mini-AI app, which will automatically be able to detect whether a person is wearing a mask or not. We will discuss the importance of architecture, and discuss convolutional neural networks. Finally, we will discuss diffusion models, self-supervised learning, model hubs, and other timely topics relating to neural networks.
- **Module 4 – Text Data and Large Language Models**: the latest and most exciting developments in deep learning are in the area of text analytics. In this module, we explore the cutting-edge research that has enabled computers to understand human text, ultimately culminating in ChatGPT. We will discuss circumstances in which these models are truly necessary, and show that in some cases, old models work just as well. Finally, we will explore transformers, the neural network architecture that made these developments possible, and discuss the importance of data scale in the training of these models.
- **Module 5 – Topics in AI**: in this last module, time allowing, we will cover recent, cutting-edge AI innovations, and other related topics. Examples might include: new techniques for model interpretation and reliability, specific predictive modelling approaches, a discussion of modelling (the art of translating real world problems to predictive analytics problems), visualization tools, and more.

Requirements and Grading

Before class begins, you will be required to complete some pre-work – see Canvas for details.

Each of the two classes in the sequence will be graded separately, 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 (45%)**

- **Homeworks (30%)**: 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 the class.

  There will be a menu of homework assignments (which will change periodically) and you'll have to pick a subset of those assignments to complete, depending on the iteration of the class you take; details will be given in the first class and on Canvas.

  Data science is difficult, and I would be doing a disservice if I made the homeworks easy. As such, be warned – *these homeworks are designed to be difficult*. To make things fair, 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 (but perhaps not correctly) and submitted in a clear, well-presented, and easy-to-follow report clearly explaining the logic behind the steps you took.

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

- **Attendance and participation (25%)**

Course Materials

There is no required textbook for the class. There will be cases and slides that will be posted on Canvas.

For those of you looking for additional reading, I have found the following three resources to be excellent:

- *Data Science for Business*, by Foster Provost and Tom Fawcett. This book is pitched at the MBA level, and covers many of the topics 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, very 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* greatest depth into any of the topics in this class, this is the class to go.
- *Business Data Science*, by Matt Taddy. This book is also pitched a more advanced level, and requires some knowledge of statistics, probability, and calculus. For those with that background, it covers many (but not all) of the topics we will be discussing in our class, and includes excellent examples.
- *Deep Learning with Python*, by François Chollet. Also pitched at a more advanced leve, this book does a great job at *really* teaching the fundamentals of data science by showing you how to fit these models in Python. It also does a great job of avoiding complex mathematics, but it does require a level of coding and numerical sophistication.

Software

This course will require the use of Excel – we will provide you with an add-in called XLKitLearn (www.xlkitlearn.com), which will extend the functionality of Excel to cover the topics in this follow-up elective. You will be asked to install this add-in as part of the pre-work for this 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.

Absolutely no Python or coding is required to complete this class.

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.