1. Course Description

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. 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. In addition, many current and recent startups are based on the application of 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.

You were introduced to the fundamentals of business analytics in your core ‘Business Analytics’ class. In this class, you will continue your study of Business Analytics, and apply these methods to new cases in a broad range of industries. In particular, we will

- Extend and deepen the methods your learnt in Business Analytics. You will learn how to use these 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).
- Introduce more complex, powerful, and flexible methodologies for predictive analytics than those you covered in Business Analytics, such as random forests.
- Introduce new frameworks such as visualization (in Tableau) and regularization that will supplement any analytics work you do.

Much as Business Analytics does, 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.
This course will not require any coding or prior knowledge other than your core Business Analytics and Statistics classes. However, the material presented will require more mathematical sophistication than your core classes.

2. **1.5 vs. 3 credits**

This course is offered in two formats - 1.5 credits and 3 credits. The arc of the class will be the same in both formats, but the 3 credit course will include more material and additional cases to illustrate the topics we will cover. The 3 credit course will also include far more time for in-class work so that you can get hands-on experience with the topics in 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. Items in red below will only be covered in the 3 credit version of the class.

- **Pre-work:** Before class begins, you will be required to install the BA2 add-in, 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.

  Attendance at the first class is compulsory, because we will be familiarizing ourselves with the add-in which we will use during the rest of the class.

- **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 deeper look at p-values and the nonparametric bootstrap
  - A review of overfitting
  - The bias-variance tradeoff
  - K-Fold cross-validation
  - Best-subset selection for linear regression
– The Lasso and the power of regularization
– Come rain or shine: A simple weather prediction model
– Storing BIG data: sparse and dense datasets
– **Case:** Analyzing Performance in the New York City Public School System
– **Case:** Exploring the Efficacy of the NCC Jail & Prison Ministry
– **Case:** Analyzing Racial Disparities in Police Stops in Philadelphia
– **Case:** From Likes to Traits: How Cambridge Analytica Profiled the World

**Module 2: Powerful Predictions; Regression 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 – the random forest. Random forests comprise many smaller and simpler models called classification and regression trees, which are weak individually but reinforce each other to produce highly predictive models. Random forests 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:

– The magic of non-parametric models
– Decision trees – from scratch
– **Case:** Elections R Us and Ecolights LLC
– Trees and the bias-variance tradeoff
– Ensemble models – boosting and bagging
– Boosted decision trees
– Bagged decision trees and random forests
– **Case:** Data Driven Investment Strategies for Peer-to-Peer Lending – the Case of Lending Club
– Model interpretability and variable importance; permutation and gini variable importance
– **Case:** Automated Mail Sorting at the USPS
– **Case:** Should Analytics be Feared or Loved in the American Football Universe?

**Module 3: Data Visualization in Tableau**

Many of the cases we have discussed thus far have featured companies with very specific problems, and the way they have used analytical techniques to solve these problems.

In real life, things are rarely this clean. Companies are often faced with ill-defined problems that have no single, obvious solution, and a complex data landscape that does not immediately lend itself to easy analysis. In those situations, businesses need to engage in exploratory data analysis to narrow the scope of their problem, and when datasets are large enough, even the simplest of exploratory tasks can be difficult.

In this module, we will discuss the art and science of data visualization using a tool called Tableau, and show how companies can use this tool to leverage their data against their most pressing problems.

Detailed module plan:
An introduction to data visualization
Combining datasets – left, inner, right, and outer joins
An introduction to Tableau
Case: The Global Superstore case
Case: Understanding Citibike: Data Visualization and Exploration in Tableau
Case: “Everyone’s Gone!” – using NYC Open Data to understand the flight from New York as part of the COVID-19 pandemic

Module 4: 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

Additional material. In the first iteration of the 3 credit course, we might have time leftover at the end for some additional material. I’ll take a vote from the class to decide which of the following three topics we’ll cover, time permitting

- Epidemic Analytics – understandings the SIR model and $R_0$
- Advanced Optimization: robust automated class and exam scheduling systems
- Advanced Simulation: extending limited COVID-19 testing capacity

4. Course Materials

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

For those of you looking for additional reading, I have found the following two 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 we will be covering in this class. It is excellent, but does not go into quite as much depth as we will.
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 four homeworks, one for each module. 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.

For the 1.5 credit format, your homework grade will be calculated as

\[
\frac{\text{HW1 grade} + \text{HW2 grade} + \max(\text{HW3 grade}, \text{HW4 grade})}{3}
\]

For the 3 credit format, you homework grade will be calculated as

\[
\frac{\text{HW1 grade} + \text{HW2 grade} + \text{HW3 grade} + \text{HW4 grade}}{4}
\]

The homeworks for the 3 credit version will also include additional questions pertaining to the additional material; these will be clearly indicated.

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.

**Group project (extra credit)**: Those interested will have the option of completing a group project for extra credit in groups of up to four students. For this project, you will need to identify a business problem that might be tackled using the techniques discussed in this class (predictive modeling, text analytics, or visualization) and a dataset on which you can apply those techniques. This dataset must be obtained from a company that you are currently working with, or have worked with in the past. You will then put together a report listing your findings and conclusions (maximum two pages of single-spaced letter paper, plus unlimited technical appendices, tables, diagrams, and spreadsheets/code. You will only Think of this group project as a mini-version of the cases we cover in this class, and use these cases for inspiration. Indeed, to get a 10 on your project, it will need to be of sufficient quality to be used in a later iteration of this course as a case or a homework (and I might indeed use it in that way, with attribution!)

The project will be graded out of 10 as follows:

- **0 points**: no attempt at a project.
- **2 points**: some attempt at identifying a problem that could be tackled with the techniques discussed in the class, and an associated dataset, presented in a compelling, entertaining report.
- **4 points**: all the requirements for 2 points, *and* some attempt at implementing the techniques discussed in the class on the dataset in question.
- **6 points**: all the requirements for 4 points, *and* a comprehensive attempt at implementing at least one technique (eg: not just trying one decision tree, but tuning for the best depth).
• **8 points**: all the requirements for 6 points, *and* a clear understanding of what techniques are appropriate and what techniques are not. An attempt to implement a range of appropriate modeling and visualization tools to build a convincing, end-to-end analysis.

• **10 points**: outstanding project, of comparable quality to the homeworks and case studies in the class. I expect between 0 and 1 projects in the class to get this score.

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

### 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, and **absolutely no Python is required to completed this class**.

### 7. The BA$^2$ Community

I maintain a Business Analytics$^2$ mailing list for all alumni of the class. When you complete the class, you will automatically be added to this list. I will email the list 1-3 times a year to foster community among BA$^2$ alumni, update you on the latest and greatest changes to XLKitLearn, and announce one-off lectures I will be hosting for alumni on topical analytics-related subjects. (You are, of course, welcome to unsubscribe at any time, though I’ll question your life choices...)

### 8. 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](http://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](http://sexualrespect.columbia.edu/gender-based-misconduct-policy-students).