Columbia Business School

B8101 ~ B7101

Business Analytics 2

Need help? Have any questions? **Email** **ba2@quetta.com** – this email address will reach the RoboTA for the class who will forward your email to me and every TA, and remind us every hour if we don't respond to your email promptly

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

Professor Daniel Guetta

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.

Real estate developers are unlikely to be skilled builders; but unless they understand *how* construction techniques work, they are unlikely to be able to do their job well. Similarly, modern executives are unlikely to need to code models themselves, but unless the understand *how* these techniques work, they are unlikely to be able to leverage them effectively.

In this class, you will extend the material you learned in your core Business Analytics class and apply these methods to new cases in a broad range of industries. In particular, you will

- Extend and deepen your study of the methods you learned 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 and image data).
- Learn more complex, powerful, and flexible methodologies for predictive analytics than those you covered in Business Analytics, such as random forests.
- Introduce a framework that will help you translate business needs and priorities into analytic problems.
- Learn the language you will need to communicate with data scientists and other engineers implementing these models.

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, 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.

Cases are listed in *red italics font* below.

- **Module 1:** Introduction
  - Introduction to the class – what this class is about, and not about
  - Using the XLKitLearn add-in
  - The three pillars of business analytics – predict, optimize, and explain
- **Module 2**: Regression & Uncertainty Revisited
  - *The New York City Public Schools case*
  - A review of basic regression
  - The nonparametric bootstrap
  - A deeper look at *p*-values and significance
  - A review of categorical variables
- **Module 3**: Ideas into Analytics
  - *The Fixed Income Trading case*
  - Going from a business problem to an analytics use case
  - The five ingredients of an analytics problem: the target variable, unit of analysis, features, evaluation metric, and the baseline
  - Incorporating time into analytics problems
  - Evaluation metrics: the mean squared error, the R-squared, the ROC curve, the ROC-AUC, the lift curve, calibration curves
  - Picking the correct evaluation metric
  - *Applying these techniques to the New York City Public Schools case*

- **Module 4**: Overfitting Revisited; the Magic of K-Fold Cross-Validation
  - A review of overfitting
  - The bias-variance tradeoff
  - *The "Cubic Model" of COVID-19*
  - Overfitting and K-NN
  - Diagnosing overfitting
  - Overfitting in the New York City Public Schools case
  - K-Fold cross-validation
  - Assessing model stability
- **Module 5**: Regression on Steroids: The Lasso
  - The bias-variance tradeoff in linear regression
  - Traditional techniques
  - *Come Rain or Shine: A Simple Weather Prediction Model*
  - Shrinkage estimators
  - From shrinkage estimators to the Lasso
  - Variable selection with the Lasso
  - *Predicting spending behavior with the Lasso*
- **Module 6**: The Case of Cambridge Analytica
  - *The Cambridge Analytica Case*
  - Storing BIG data: representing sparse datasets
  - Understanding the predictive analytic task
  - *From likes to traits: How Cambridge Analytica Profiled the World*
- **Module 7**: An Introduction to Decision Trees
  - *Autonomous Vehicles: The Analytics Behind the Hype*
  - Linear regression: Important Shortcomings
  - The magic of non-parametric models
  - An introduction to decision trees
  - Trees and the bias-variance tradeoff
  - Fitting decision trees: the CART algorithm
- **Module 8**: Boosted Trees
  - Ensembles: hacking the Bias-Variance Tradeoff
  - An introduction to boosting
  - *The Predictive Medicine Case*
- **Module 9**: Random Forests
  - *The Lending Club case*
  - Finding the right target variable
  - An introduction to bagging
  - Revisiting the marketing spend case
  - Random forests
  - Model interpretability and variable importance
- **Module 10**: The USPS Case
  - *The USPS Case*
  - Dealing with multi-class classification
  - An introduction to neural networks and deep learning
- **Module 11**: Data Visualization in Tableau
  - An introduction to data visualization
  - Visualization best practices
  - *The NYC Garbage Collection Case*
  - An introduction to Tableau

    - *The Citibike Case*
- **Module 12**: Text Analytics
    - Unstructured data: an introduction to text analytics
    - Evisort: An A.I.-Powered Startup Uses Text Mining to Become Google for Contracts
    - The bag of words representation
    - Text data for predictive analytics
    - Supervised vs. Unsupervised Learning
    - An introduction to Latent Dirichlet Allocation

Time permitting, we may cover one or more of the additional cases below to further illustrate the concepts in the class

- **Appendix A**: *Analyzing Racial Disparities in Vehicular Police Stops in Philladelphia*
- **Appendix B**: *Analytics at the NFL – Predicting Blitzes*
- **Appendix C**: *The Mathematics of Exponential Growth: Understanding COVID modelling, the SIR Model, and the R-0*
- **Appendix D**: *Advanced optimization: robust, automated class and exam scheduling systems*
- **Appendix E**: *Advanced simulation: extending limited COVID-19 testing capacity*

Requirements and Grading

Before class begins, you will be required to complete some pre-work – see Canvas 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 seven homeworks. 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 b learning in the class.

  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.

- **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 modelling *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*, or for which your absence is excused.
  - **Nameplate (25%)**: this part of your grade will be calculated by finding the fraction of classes your attend *with your nameplate clearly visible*, 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 (even late) or for which your absence is excused.
  - **Contributions in class (25%)**: this part of your grade will be calculated based on my impressions, an on your participation in ad-hoc assignments such as pre-class work.

  Please note that I am *very* generous with excusing absences – for any reason – provided you let me know at ba@guetta.com *at least an hour before class*.

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.

Software

This course will require the use of Excel – we will provide you with an add-in called XLKitLearn (http://guetta.org/xlkitlearn), 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.

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, which I use around 1-3 times a year to foster community among alumni of the class, 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…)

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.