BSAD 8700 -- Business Analytics

Instructor: Dustin White
Days and Times: Wednesdays, 6:00 to 9:35 PM
Classroom: Mammel Hall 115

This course is designed to help you learn about the most important methods underlying modern data analytics, how to use data tools to improve your understanding of business decisions, and how to communicate clearly with both data analysts about the goals of your team as well as with managers about results from data analysis projects.

We will have eight classes during the semester. The first hour (or 90 minutes, depending on the week), will be focused on teaching you about the tools available for data analysts. The rest of the class will be a REQUIRED lab in which you will use data analytics to create useful information for class projects.

The goal of this course is to improve your ability to take advantage of data analytics in your career. If there are things that we are not covering that you feel are critical, *please* bring them up to me so that I can help you with the topics you care most about.

Office Hours

I will hold office hours just before class (Wednesdays from 4:45 PM to 5:45 PM), or by appointment. I would love to get to know you and your goals, so please come by and talk to me.

Grading

This course will be graded as follows:

- 70% of your grade will be based on performance on the two projects that make up lab and take the place of exams. Each project will receive a weight of 35%.
- 15% will be based on in-class participation. This involves coming prepared to class so that you can discuss the topics we cover each week, and being ready to ask questions. Asking questions is one of the best ways to learn!
- 15% will be based on completing the reading assignment for the week.

Final grades will be based on the weighted average of your work, and distributed according to the following scale.

Letter	Percent	Letter	Percent	Letter	Percent
A	94-100	C+	77-79.9	F	< 60
A-	90-93.9	C	74-76.9		
B+	87 - 89.9	C-	70-73.9		
B	84 - 86.9	D+	66-69.9		

B-	80-83.9	D	60-65.9

Reading Assignments

The assignments for this class will be made up of reading assignments from two books: Predictive Analytics, by Eric Siegel (ISBN:978-1119145677) and Data Science for Business by Foster Provost and Tom Fawcett (ISBN:978-1449361327). Each week, you need to write a one to two page summary of the applications from the writing that you feel best apply to your job, and how they could be implemented by you or your team at work.

Projects

In place of exams, this course will have two projects, which we will work on during each class period. The best way to learn is to do, and so we will focus on *doing* data analytics. I don't expect you to know how to code, but you will need to understand some code during the course of the semester. I will help you do so, and will make the process as painless as possible. The primary goal is to help you *experience* data analysis. These projects will be the largest contributors to your grade, so please make sure that you schedule time to remain for *all* of class each week. These projects may be done as part of a group of up to three class members (*recommended*), or alone.

Project 1 - Due at the start of class, Week 5

This project is designed to help you become acquainted with data analysis, and will require elements of each of the first four lectures and labs to complete. You must submit all of the code that you use during the process of completing this project. Aside from any visuals or code you submit, the project should be approximately 2-3 pages long.

1. Using the ACS data provided, begin by choosing three plots that you made in PowerBI that you find most interesting, and explain what they tell us about some subset of ACS respondents.
    - Why is this group interesting to you?
    - What do you hope to learn by examining this group more closely?
2. Using SQL, isolate your population(s) of interest from Step 1. Within the new population(s) that you have chosen, create summary statistics of the variables that you decide to investigate as well as any useful demographic information about respondents. Make sure to select at least 10 variables for use in this project.
    - What are the means and standard deviations of the variables you are interested in?
    - Looking at the minimum and maximum values, do you see anything that is worth noting? Do you think you should restrict your sample further to eliminate outliers?
    - Why is it so important to look at summary statistics before plunging into the data?

3. Based on your new sample from last class, create two visualizations of your data that could be used to inform policy or decisions. Explain why you chose those visuals, and what they suggest about your population of interest.
    - What are the advantages of using data visualization to make decisions?
    - Are there any disadvantages?
    - How do you think that we should use data visualizations in decision-making?
4. Continuing to use the population you selected in Week 2, run two separate linear regressions (using different dependent variables) in order to evaluate the impact of at least three parameters on each outcome (dependent variable).
    - What do your regressions tell you about the population?
    - Do you have any reason to believe that there is a causal relationship between your parameters and outcomes?
    - What makes regression analysis so powerful?

Project 2 - Due at the start of class, Week 8

This time around, you will be making use of a dataset voted on by the class. You will focus in this project on making the most accurate out-of-sample prediction possible. I will introduce new techniques each week in order to help you develop more accurate predictions of the outcomes of interest. Once again, all code used in the course of the project must be submitted. The writeup for this project should be 1-2 pages long, not including any visuals that you choose to utilize.

1. Using the training data, come up with your best explanation of how to predict the outcome of interest using visualizations of the data. Write down a procedure for deciding how to classify new observations.
   - How accurate are your predictions? (I will test the procedure out for you over the week after lab.)
   - What was the biggest problem for you in making accurate predictions?
   - What would you change if you made predictions like this again?
2. Implement two different decision tree classifiers to make the same predictions that you tried to make in Step 1. Use different attributes in each implementation, and compare the results. How well do decision trees do in predicting out-of-sample observations?
   - What was the accuracy reported by each of your decision tree classifiers?
   - What was (were) the hardest part(s) of implementing the decision trees?
   - What strengths (and shortcomings) have you observed while using decision trees?
3. Create a random forest using one (or both!) of your decision trees from Step 2. Generate your forest using 100 and 1000 trees. Is a random forest an improvement from your decision tree classifier? Is it superior to your procedure from Step 1?

- What do you think might make a random forest more powerful than a single decision tree?
- Does a random forest solve any of the shortcomings of decision trees?

Course Schedule

Week 1

Introduction to Data Analytics -- This week we will discuss what data analytics is and what it is not. Data analytics is kind of a hot thing right now, and we want to make sure that we understand some limitations of using data. Topics will include: defining data analytics, flaws in data, omitted variables, and asking the right questions. In lab, we will get acquainted with data by using Microsoft's PowerBI software to visualize and get to know some interesting data.
Read PA: Intro and DSB: Ch. 1

Week 2

SQL and Databases -- This week we take a critical step in making use of very large datasets: we need to learn how to access small segments of them quickly, and to construct basic explanations of that data, so that we can understand what the data might be used for. I will walk you through what SQL is, and how to use it to *query* a database. In lab, we will use our time to find an interesting subset

of a database for our first project, and construct summary statistics for that data.
Read PA: Ch. 1 and DSB: Ch. 2

Week 3

Data Visualization -- Now that we have done some basic work in both creating smaller, more tractable data, and in basic data visualization, we are ready to talk about the importance of visualizing data, and using that to inform our analysis. We will talk about good and bad visualizations, and how to tell the story of your data using visuals. In lab, we will prepare high-quality visualizations that we could use to present and explain important aspects of the data that we explored in Week 2.
Read PA: Ch. 2 and DSB: Ch. 3

Week 4

Correlations and Regressions -- After spending the past few weeks exploring our data, and getting our heads around the questions that we might want to ask of our data, we will talk about classical methods of data analysis. We will spend some time talking about what correlations and regressions are, and about when they are the best tools for the job. In lab, we will take our data, and explore the effects of variables that we choose on some outcome that we care about. We will be able to combine this information with our visualizations and summary statistics from

previous weeks into a nice briefing for our "boss" about what we learned from our data, and why he/she should believe our discovery.
Read PA: Ch. 3 and DSB: Ch. 4

Turn in First Project At Start of Class, Week 5

Week 5

Classification, Supervised and Unsupervised Learning -- For the second half of the course, we will focus primarily on classification techniques. This week, we will discuss what classification is, and why it is so popular in data analytics. We will also define supervised and unsupervised learning, and look at an example of each. Lab this week will be focused on trying to visually classify individuals using plotting and other visualizations of our data.
Read PA: Ch. 4 and DSB: Ch. 5

Week 6

Decision Trees -- Decision trees are one of the simplest and most powerful analytic tools in use today. We will talk about what a decision tree is, what makes it different from a histogram classifier, how it is "grown", and how it is "pruned" in order to come to the best tree for making out-of-sample classifications of observed data. Lab will consist of preparing data for use in decision tree algorithms, as well as training a decision tree in order evaluate its

accuracy in out-of-sample predictions.
Read PA: Ch. 5 and DSB: Ch. 6

Week 7

Ensemble Methods -- Just like people, an individual algorithm might be biased. By combining many decision trees or other machine learning algorithms, we can come to a more accurate estimate based on the available data. We will discuss how ensemble methods work, their advantages, and their use in industry. In lab, we will generate random forests of decision trees in order to improve the accuracy of our predictions from last lab.
Read PA: Ch. 6 and DSB: Ch. 7

Turn in Second Project At Start of Class, Week 8

Week 8

Cutting Edge: Neural Networks and Other Advanced Algorithms -- The tools we have learned so far have been easy to apply, and they work in ways that are easy to visualize. More advanced techniques such as neural networks perform the same tasks, but can be applied to many more difficult problems such as natural language processing and image processing. We will go over the basic concepts of neural networks in class, and explore a simple neural net in lab in order to compare it to the other techniques we have learned.
Read PA: Ch. 7 and DSB: Ch. 8