BSDS 200
Applied Techniques in Data Science
Spring 2020

Instructor: Nicholas Ross

Contact email: ncross@usfca.edu

Office Hours: Thursday after class in HR 107B (James Wilson's office)

Textbooks: No textbook is required, the following books are provided as potential references.

- *Data Management* – Details to be given on the first day of class.

Other References. We will be covering PostgreSQL and Pandas, so here are some references to help you:

- *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* – Wes McKinney. This is the *the* book about Pandas, written by Wes McKinney – the author of Pandas!
- *PostgreSQL Documentation* which can be found at `https://www.postgresql.org/docs/`. This is the PostgreSQL's documentation and is incredibly detailed and easy to follow.

1 Introduction

A key component of modern data science is using Python and SQL to manipulate and analyze data. In this course, we build upon knowledge from BSDS 100 and computer science courses to provide you these applied skills. In particular, the primary focus of this class is the manipulation of datasets with Python and SQL with a secondary focus on applying data science techniques to this data.

This course contains three streams of content. The first stream is SQL, the most common tool used to manipulate structured data sets. The variant of SQL used in this class, PostgreSQL is popular in the data science field and is, arguably, the most modern of all relational databases systems. The second stream of content is using Python (and Jupyter notebooks) to manipulate data with NumPy and Pandas, two libraries that are used to reshape, manipulate and "wrangle" data.

We will also learn how to, using Jupyter notebooks, connect Python and SQL and apply data science techniques (hypothesis testing, regression, etc.). This may be the most common data science work flow used today.

The above two streams are both foundational and vocational data science skills; they are required to do anything with data science in the same way that using a hammer is necessary to build a house. After we are comfortable with the techniques of these two streams we can move onto the last stream of content. In this section we will apply data science techniques to a number of common, customer-level, business problems – in particular, churn prediction and the estimation of customer long-term value ("LTV").

At the end of this course, I expect students to easily manipulate data within Python and SQL as well as understand the basic trade-offs involved in moving between these tools and how to leverage that knowledge when applying data science techniques.

The challenge of this class is that we are learning a new language and a new way of thinking about data manipulation. Just like learning any new language, the most efficient way to learn is via strategic repetition, which this class will provide plenty of. I expect students to spend 10-20 hours per week outside of class on this course. If you are uncomfortable with that level of commitment, this may not be the best use of your time. This course demands *a lot*, but from it you will gain the skills necessary to be a data scientist.

Course Learning Objectives

By the end of this course, you will be able to:

1. Use SQL to extract, transform and manipulate large data sets in preparation for analysis. This includes using common clauses, such as SELECT, FROM, WHERE, GROUP BY, and ORDER BY as well as more modern analytic/window/partition functions.
2. Communicate results in an efficient manner, suitable for business settings. This includes basic data visualization with Python and Plotly.
3. Use Python Notebooks (Jupyter) to complete exploratory data analysis.
4. Use NumPy and Pandas to manipulate data within Python. This includes data wrangling, reshaping, manipulating strings, numbers and dates.
5. Normalize data in order to efficiently use it in different data science tools (long vs. wide, encoding, etc.).
6. Undertake statistical analysis using Python and its associated libraries (stats, sci-kit, etc).
7. Utilize common data science techniques to answer applied business focused questions, including churn prediction and estimation of customer LTV.

Requirements

As an advanced class within the BSDS major this course has a number of requirements:

- **Python (CS 110):** Students should be familiar with Python. This includes:
    1. Loops (for, while)
    2. Conditionals (if)
    3. Data objects (lists and dictionaries)

4. List comprehensions
5. Defining functions (def)
6. File handling (open)

- **General Computing:** Students should have a computer and be able to install Anaconda and Jupyter on it.
- **Statistics (BSDS 100):** Students should understand basic statistical concepts including:
  1. Basic Hypothesis Testing
  2. Linear Regression
- **Mathematics (Math 110, 230):** Students should be familiar with:
  1. Calculus I (integration and derivation)
  2. Linear Algebra (manipulation of matrices)

To be clear – the concepts mentioned above are going to be assumed during this class and will be not be explained. If you know you are weak in one of the areas above it would behoove you to refresh your knowledge.

Assessment

This class uses what is called a more "continuous" feedback strategy. Under this system, students are provided feedback in a more continuous manner. In particular, this course puts weight on frequent quizzes, rather than relying on a few large exams.

This class is fast and difficult. You will NOT be able to catch up.

In terms of overall grades, the breakdown can be found below.

Type	Percent of Grade
Quizes	20%
Homework	10%
Exam #1	15%
Exam #2	15%
Final Exam	30%
Professionalism	10%
Total	100%

Quizes

Quizzes are administered at the start of class and will last approximately 20 minutes. Each quiz will ask a few questions about the preceding material and is graded aggressively; the material should be fresh in your mind. Quizzes are done individually and there no make-ups. If a student is late to class they will not be given additional time. Quizzes are done each Thursday and cover the material up to and including the Tuesday beforehand.

Homework Assignments

This class has a ton of homework, all of which is completed in groups. Homework is graded **lightly**. Homework needs to be uploaded by 9:55AM the day that it is due. **Late work is not accepted.**

Exams

There are three exams in this course: two in-class and a final. Exams are closed-book, closed-notes and no calculator is required. Tentative dates for the exam are:

- **Exam #1:** 2/27/2020
- **Exam #2:** 4/9/2020
- **Final Exam:** 5/12/2020 10-12

Professionalism

As you can see, quite a bit of the grade involves being a professional. In this class, that means:

- Showing up on time
- Showing up prepared
- Contributing to class
- Being responsible to your group when doing group work
- Doing required readings

Students who show up on time, work through the example problems and pay attention in class have a tendency to do extremely well in this class. The opposite if also true; students who fail to make an effort, fail to contribute in class and fail to pay attention have a tendency to do poorly in this class. An approximation of your participation grade may be made by following the rubric below:

Grade	Attendance / Promptness	Participation / Professionalism
A	Prompt and Complete attendance each day	Full attention (e.g. no playing on the phone) and **frequent, informed** contribution to class discussion
B	Prompt and Complete attendance each day	Full attention (e.g. no playing on the phone) and **regular** contribution to class discussion
C	Prompt and Complete attendance each day	General attention and occasional contribution to class
D/F	Frequently late / tardy	Marginal Attention (texting, playing on a computer) and little or no classroom contribution

Class Topics and schedule

Since this is the first time that this course has been taught at this University, I'm leaving some leeway in the speed that we will cover material. I'll be publishing the next few weeks of the course on canvas. That being said, here are some important dates:

Odds and Ends

- Cheating is not tolerated. At all. Unless an assignment is clearly designated as group work, I expect it to be done alone. USF's honor code can be found online and I expect it to be followed. Disciplinary action will be taken against any student found violating this code.

- As a Jesuit institution committed to cura personalis - the care and education of the whole person – USF has an obligation to embody and foster the values of honesty and integrity. USF upholds the standards of honesty and integrity from all members of the academic community. All students are expected to know and adhere to the University's Honor Code. You can find the full text of the code online at www.usfca.edu/academic_integrity. The policy covers:

  - Plagiarism – intentionally or unintentionally representing the words or ideas of another person as your own; failure to properly cite references; manufacturing references.
  - Working with another person when independent work is required.
  - Submission of the same paper in more than one course without the specific permission of each instructor.
  - Submitting a paper written by another person or obtained from the internet.
  - The penalties for violation of the policy may include a failing grade on the assignment, a failing grade in the course, and/or a referral to the Academic Integrity Committee.

- Attendance is required. Unless a student gives me prior warning, all absences are considered unexcused.

- The quickest way to reach me is via email.
- Class participation is required. This is a big part of your grade. I expect class to include discussions. Just to make this clear and give you an even bigger incentive to pay attention and ask questions, if you catch me making a mistake, I will give you a chocolate (if you are allergic to chocolate, please let me know).
- If you are a student with a disability or disabling condition, or if you think you may have a disability, please contact USF Student Disability Services (SDS) at 415 422-2613 within the first week of class, or immediately upon onset of disability, to speak with a disability specialist.

  If you are determined eligible for reasonable accommodations, please meet with your disability specialist so they can arrange to have your accommodation letter sent to me, and we will discuss your needs for this course. For more information, please visit: http://www.usfca.edu/sds or call (415) 422-2613.
- All students are expected to behave in accordance with the Student Conduct Code and other University policies (see http://www.usfca.edu/fogcutter/). Open discussion and disagreement is encouraged when done respectfully and in the spirit of academic discourse. There are also a variety of behaviors that, while not against a specific University policy, may create disruption in this course. Students whose behavior is disruptive or who fail to comply with the instructor may be dismissed from the class for the remainder of the class period and may need to meet with the instructor or Dean prior to returning to the next class period. If necessary, referrals may also be made to the Student Conduct process for violations of the Student Conduct Code.
- As an instructor, one of my responsibilities is to help create a safe learning environment on our campus. I also have a mandatory reporting responsibility related to my role as a faculty member. I am required to share information regarding sexual misconduct or information about a crime that may have occurred on USFs campus with the University. Here are other resources:
  - To report any sexual misconduct, students may visit Anna Bartkowski (UC 5th floor) or see many other options by visiting our website: www.usfca.edu/student_life/safer.
  - Students may speak to someone confidentially, or report a sexual assault confidentially by contacting Counseling and Psychological Services at 415-422-6352.
  - To find out more about reporting a sexual assault at USF, visit USFs Callisto website at: www.usfca.callistocampus.org.
  - For an off-campus resource, contact San Francisco Women Against Rape (SFWAR) (415) 647-7273 (www.sfwar.org).