MIS 373 04045 Advanced Analytics Programming

Schedule:	MW 2:00-3:30 pm in CBA 5.330
Instructor:	Deepayan Chakrabarti
Email:	deepayan.chakrabarti@mccombs.utexas.edu
Teaching Assistant:	Rohit Arora (arorarohit@utexas.edu)
Instructor office hours:	By appointment (CBA 6.462)
TA office hours:	Monday 3:15-4:30 pm or by appointment (CBA 3.332L)
Pre-requisites:	MIS 304 (Intro to Programming)

Course Overview

Should I lend to this borrower? Can I detect fraudulent credit card transactions? What are the main types of complaints of my customers? Did my new website design significantly change sales?

Data-driven analysis has wrought a quiet revolution in business. As disk storage and computing power have become cheaper, companies have started maintaining detailed logs of inventories, sales, and customer activity, among others. Yet, this is only half the job; the real need is for *insights*, and this course teaches you the tools for that.

We will learn data analysis in Python, a general-purpose language that lies at the intersection of (a) easy enough to learn, (b) fast enough to scale, and (c) endowed with a wide range of powerful libraries that make data cleaning, visualization, and many common data analysis tasks a cinch. The course is split into five parts:
(1) **Introductory Python**, where we learn the basic language syntax, and gain familiarity with general-purpose tools such as string manipulation,
(2) **Pandas**, which is a powerful data analysis toolkit (similar to R) that makes it easy to explore and visualize data,
(3) **Classification**, where we develop an understanding of how to make predictions,
(4) **Clustering**, where we learn how to discover the major groups or components of a given dataset, and
(5) **Other Topics**, including regression and hypothesis testing.

Course Materials

Books. For the first two parts of the course, we will use *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, by Wes McKinney. For the remainder, we will use material from a variety of sources, one of which is *Applied Predictive Modeling*, by Max Kuhn and Kjell Johnson. For introductory Python, an additional good reference is *Think Python*, by Downey, available here.

Software. Python, including the following packages:

- Jupyter
- Numpy
- Pandas
- Matplotlib
- Sklearn
- Statsmodels

The preferred distribution is Anaconda (Python 3.x version).

Grading Policy

The course grade will be calculated as follows.

Work item	Weight
3 Group Assignments	$3 \times 10 = 30$
Midterm	20
Group Project	20
Final exam	30

Groups. Towards the beginning of the class, students will be randomly assigned to groups of 4 students each. All homework assignments and the project must be done in groups.

Homework assignments. There will be three homework assignments. These must be completed by each group, and must be handed in electronically via Canvas before the beginning of the class.

Project. You must develop a group project on any topic that interests you. The project should have the following characteristics:

- There should be a clear motivation. You should be trying to solve a question that matters, either to society or business.

- The dataset should be "large enough." It should have contain enough information to get reliable conclusions. A rule of thumb is that the dataset should have over 10,000 rows and 20 columns, but smaller datasets are acceptable if the motivation is strong and larger datasets are not publicly available.
- The project should present insights. The goal is not to just report classification accuracies. The goal is to find interesting insights about the data, e.g., which features are most important, what aspects are abnormal, what is surprising?

You will write up a project report, which you must submit via Canvas. Each group must also present your findings in class. The presentations will be held over two days, and the ordering of groups will be chosen by lottery.

Exams. There will be a midterm and a final exam. They will be open-laptop and open-notes.

Statement on Students with Disabilities

Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259, http://www.utexas.edu/diversity/ddce/ssd/.

Religious Holy Days

By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

Policy on Scholastic Dishonesty

The McCombs School of Business has no tolerance for acts of scholastic dishonesty. The responsibilities of both students and faculty with regard to scholastic dishonesty are described in detail in the BBA Program's Statement on Scholastic Dishonesty at http://www.mccombs.utexas.edu/BBA/Code-of-Ethics.aspx. By teaching this course, I have agreed to observe all faculty responsibilities described in that document. By enrolling in this class, you have agreed to observe all student responsibilities described in that document. If the application of the Statement on Scholastic Dishonesty to this class or its assignments is unclear in any way, it is your responsibility to ask me for clarification. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure

in the course and/or dismissal from the University. Since dishonesty harms the individual, all students, the integrity of the University, and the value of our academic brand, policies on scholastic dishonesty will be strictly enforced. You should refer to the Student Judicial Services website at http://deanofstudents.utexas.edu/sjs/ to access the official University policies and procedures on scholastic dishonesty as well as further elaboration on what constitutes scholastic dishonesty.

Campus Safety

Please note the following recommendations regarding emergency evacuation, provided by the Office of Campus Safety and Security, 512-471-5767, http://www.utexas.edu/safety:

- Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when a fire alarm is activated. Alarm activation or announcement requires exiting and assembling outside.
- Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building.
- Students requiring assistance in evacuation should inform the instructor in writing during the first week of class.
- In the event of an evacuation, follow the instruction of faculty or class instructors.
- Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
- Behavior Concerns Advice Line (BCAL): 512-232-5050
- Further information regarding emergency evacuation routes and emergency procedures can be found at: http://www.utexas.edu/emergency.

Table 1: *Tentative schedule*

Date	Topic	Details
01/22	Introduction	
Introduction to Python		
01/27	Python I	Values and variables, control flow, and functions (Think Python chapters 2, 3, and 5)
01/29	Python II	Data Structures (Think Python chapters 10, 11, and 12)
02/03	Python III	Data structures detailed example
02/05	Python IV	Strings and regular expressions Files (Think Python chapters 8 and 14)
Using Pandas		
02/10	Series and DataFrames I	McKinney chapters 5 and 6 **Deadline for group formation** **First assignment released**
02/12	Series and DataFrames II	Examples using the NYC Complaints dataset
02/17	Data wrangling	Merging, concatenation Reshaping, pivoting (McKinney chapter 7) **First assignment due**
02/19	Visualization	Plotting Histograms (McKinney chapter 8)

Continued on next page

Table 1 – continued from previous page

Date	Topic	Details
02/24	Grouping data I	McKinney chapter 9 **Second assignment released**
02/26	Grouping data II	
03/02	Time Series	McKinney chapter 10 **Second assignment due**
03/04	Statistics	Means and standard deviations Medians and quantiles Correlations
(midterms)		
03/09	Review session	
03/11	**Midterm**	
03/16-21	*Spring break*	
Regression		
03/23	Regression I	R-square Degrees of freedom Relation to correlation
03/25	Regression II	Logistic regression Regularization
Classification		
03/30	Intro to classification	Examples Loss function Class imbalance Train/test split Holdout set, cross-validation Accuracy measures

Continued on next page

Table 1 – continued from previous page

Date	Topic	Details
04/01	Nearest Neighbors	Distance metrics
04/06	Naive Bayes	Probability Conditional Probability The Naive Bayes Algorithm
04/08	Logistic Regression	
04/13	Decision Trees	Basic methodology Information gain and Entropy **Third assignment released**
04/15	Ensemble Methods	
Clustering		
04/20	Intro to clustering	Examples Uses of clusters as features in classification Clustering quality via RAND scores **Third assignment due**
04/22	K-Means	Distance metric Examples Selecting the number of clusters
(finals)		
04/27	**Project Presentations**	
04/29	**Project Presentations**	
05/04	Finals Review	
05/06	No class	
05/14	**Final exam: 2:00pm-5:00pm**	