Advanced Data Analytics

BUS 212A 2: Spring 2023
Tuesday, Thursday 3:55 pm – 5:15 pm

Instructor: Yeabin Moon, Ph.D. (yeabinmoon@brandeis.edu)
Subject to Change (January 9, 2023)

Office Hours: After class in my office (*Sachar International Center 209B*), or by appointment online through the following link.

TA: Yvonne Wang (yvonnewang@brandeis.edu). Please set an online meeting with her through the following link.

Course Description

The Advanced Data Analytics course is designed to provide students with the skills and knowledge needed to apply data analytics techniques to real-world business scenarios. The course covers a range of advanced data analytics methods, including machine learning, predictive modeling, and data visualization. Through a series of hands-on projects and case studies, students will learn how to use these techniques to solve complex business problems and make data-driven decisions.

Throughout the course, students will work with real-world data sets and tools, including Python and its corresponding libraries, to build and deploy predictive models. They will also learn how to communicate their findings effectively to stakeholders, including non-technical audiences. By the end of the course, students will have gained the ability to apply advanced data analytics techniques to a wide range of business scenarios and make informed, data-driven decisions. Success in this four-credit course is based on the expectation that students would need to study for about three hours for every hour of in-class time. Hence, students will spend a *minimum of 9 hours* of study time per week in preparation for this class.

Learning Goals

With this course, you will:

1. Understand the fundamentals of the Python programming language and be able to write code to perform data analysis and machine learning tasks.
2. Use key libraries such as NumPy, pandas, and scikit-learn to manipulate and visualize data in Python.
3. Understand statistical techniques for extracting insights from data and be able to apply them using Python.
4. Understand different types of machine learning algorithms and be able to apply them to real-world problems.
5. Work on hands-on projects to gain practical experience using Python for data analysis and machine learning.

6. Communicate your findings and results effectively using data visualization and other tools.
7. Have a strong foundation in data analytics and machine learning, and be able to continue learning and growing in these fields.

Main Reference

We will use ***Introduction to Machine Learning with Python*** by Andreas C. Müller, Sarah Guido (O'Reilly, 2016) as a main reference. There is a partial list of useful books that will be touched during the course.

- Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition*: More advanced treatment on practice
- Trevor Hastie, Jerome H. Friedman, Robert Tibshirani, *The Elements of Statistical Learning*: More rigorous treatment on mathematics

You have an **online access** for all the references listed above through Brandeis Library.

Prerequisites

1. College level Calculus, Linear Algebra, Probability, and Statistics (or **ECON 213a**)
2. Successful completion on BUS 211 or ample coding experience

Class Participation

There is no such thing as a stupid question. Dialogue is not only strongly encouraged, it is critical to your understanding of the material. Vocalizing your questions often helps you solidify what you do and do not understand. It also provides me important feedback on the areas in which we need to spend more time. During lectures, I will encourage questions, and I will solicit input. If I call on you, please relax, I am NOT trying to intimidate you or embarrass you in any way. I am trying to encourage active listening and keep you engaged in the course. This will greatly assist you in learning the material. If you do not know the answer, I will move on to another student. Hence, attendance is *mandatory* for this class.

Course Requirements

The grade consists of

1. 6 Assignments: 50 percent
2. 1 Exam: 30 percent
3. 1 Final Project: 15 percent
4. Attendance: 5 percent
5. Participation (bonus): 1 percent

First, six assignments will promote both your theoretical understanding and practical skills. All assignments contain both written parts and programming parts. You can drop

the lowest homework, and hence, your homework grade will be based on the resulting five assignments. You can find each deadline below. You can submit your late work with a 10 percent penalty if you can make it within ONE day after the deadline.

Second, there is one in-class exam scheduled on March 23rd. It is mandatory, and I will not provide any make-up exam for whatever reason. You should consult me if you have a time conflict by January 24th. You can bring your own 2-page cheat sheet.

The Final Project offers you the chance to apply your newly acquired skills to an in-depth application. You can do this alone or with a peer. If you want to work with a team of more than two people, you will have more stringent guidelines. Students can reuse the project with BUS 241, but you need to get permission from me in advance.

Students have two options: the final default project or a final custom project. For both options, credit for the project is broken down as follows:

- Project Proposal (30 percent)
- Project Report (70 percent)

Each deadline is listed below.

Again, attendance is mandatory. However, if you have an emergent situation, please let me know. We can make you join the class online (or through recordings).

I will give you a bonus credit for those who participate in the class actively. The credit is designed to promote a discussion-oriented environment, not a grading curve. This is solely at my discretion.

Course Plan

The class covers the major applications of data analytics. The foundation relies on the machine learning literature, but the focus is more on its practice. We will mainly examine the practical use cases and delve into theories where necessary. The following outline provides a high-level overview of the course.

1. **Introduction (January 17)**
   (a) Orientation
      - Logistics
      - Course requirements overview

   (b) End-to-End Machine Learning Project

2. **What is Machine Learning? (January 19)**
   (a) Finding a function
   (b) Learning problems
   (c) Model Accuracy assessment
      - Bias-variance Trade-Off

   (d) Decision boundary
   (e) Interview question reviews

3. **Python review (January 24 and January 26)**
   (a) Coding Style in Python: PEP8
   (b) Object-oriented programming
   (c) Introduction to Scientific Computing
   (d) NumPy, Matplotlib, Pandas

Assignment 1 deadline (11:59 AM, January 28)

4. **Linear model (January 31, February 2)**
   (a) Linear Regression set up
      - Model assumptions
      - Hypothesis testing
      - Assess the accuracy

   (b) Multiple Linear Regression
      - Interpretation in practice

   (c) Discussion on the code

5. **Introduction to Optimization Algorithm (February 7)**
   (a) Gradient Descent
      - Discussion: do we need this it ever?

   (b) Other gradient methods
      - Batch, Stochastic gradient, mini-batch

   (c) Extensions of the linear model
      - Saturated model
      - Polynomial regression

6. **Discussion on the linear regressions (February 9)**
   (a) Potential concerns
      - Theory and practice

   (b) Comparison of K-Nearest Neighbors

Assignment 2 deadline (11:59 AM, February 11)

7. **Introduction to Classification (February 14, February 16)**
   (a) Classification problem
   (b) Important assumptions
      - Discrete choice set up

   (c) Logistic Regression
   (d) Training a Binary Classifier
   (e) KNN
   (f) Performance measure

February break: no Classes Feb 20 – 24

8. **Code review on Classification (February 28)**
   (a) Discussion on the code

9. Resampling Methods (March 2)

(a) Cross-Validation

- Discussion on the validation set
- K-fold Cross-validation approach

(b) Practice

Assignment 3 deadline (11:59 AM, March 4)

10. Model selection and Regularization (March 7, March 9)

(a) Shrinkage Methods

- Ridge
- Lasso

(b) Discussion on the hyperparameter

11. Dimensionality Reduction (March 14, March 16)

(a) Curse of Dimensionality

(b) Principal Components Regression

Assignment 4 deadline (11:59 AM, March 18)

Midterm Reading day: no Classes March 21
Midterm exam (March 23)

12. Algorithm Chains and Pipelines (March 28)

(a) Building pipelines

(b) Pipeline algorithm in practice

13. Introduction to Decision (March 30, April 4)

(a) Tree-based methods

- Prediction analysis
- Regression trees
- Classification trees

(b) Discussion on the tree-based methods

(c) Decision tree in practice

Project Proposal deadline (11:59 PM, April 4)

14. Ensemble learning methods (April 18, April 20)

(a) Tree-based methods

- Prediction analysis
- Regression trees
- Classification trees

(b) Discussion on the tree-based methods

15. **Decision tree in practice (April 25)**
    (a) Code discussion

Assignment 5 deadline (11:59 AM, April 29)

16. **Unsupervised Learning (April 27, May 2)**
    (a) Introduction to Unsupervised Learning
        - Challenge
        - Discussion: popularity on supervised learning
    (b) PCA revisit
    (c) Introduction to Clustering
        - K-Means
        - Other methods
    (d) Code discussion

Assignment 6 deadline (11:59 AM, May 6)

Final project deadline (11:59 AM, May 16)

The course plan is subject to change due to a snow day/delayed start/early closing. If this situation is predictable, the class will be held on zoom. I will announce it accordingly.

Accommodations

Brandeis seeks to create a learning environment that is welcoming and inclusive of all students, and I want to support you in your learning. Live auto transcription is available for all meetings or classes hosted on Zoom and you can turn it on or off to support your learning. Please check for Zoom updates to take advantage of this new feature. To learn more, visit the Zoom Live Transcription webpage. For questions, contact help@brandeis.edu

If you think you may require disability accommodations, you will need to work with Student Accessibility Support (SAS) (781-736-3470, access@brandeis.edu). You can find helpful student FAQs and other resources on the SAS website, including guidance on how to know whether you might be eligible for support from SAS. If you already have an accommodation letter from SAS, please provide me with a copy as soon as you can so that I can ensure effective implementation of accommodations for this class.

Academic Integrity

Every member of the University community is expected to maintain the highest standards of academic integrity. A student shall not submit work that is falsified or is not the result of the student's own effort. Infringement of academic integrity by a student subjects that student to serious penalties, which may include failure on the assignment, failure in the course, suspension from the University or other sanctions. Please consult Brandeis University Rights and Responsibilities for all policies

and procedures related to academic integrity. Students may be required to submit work via TurnItIn.com or similar software to verify originality. A student who is in doubt regarding standards of academic integrity as they apply to a specific course or assignment should consult the faculty member responsible for that course or assignment before submitting the work. Allegations of alleged academic dishonesty will be forwarded to the Department of Student Rights and Community Standards. Citation and research assistance can be found at Brandeis Library Guides - Citing Sources.

Classroom Health and Safety

- Register for the Brandeis Emergency Notification System. Students who receive an emergency notification while attending class should notify their instructor immediately. In the case of a life-threatening emergency, call 911. As a precaution, review this active shooter information sheet.
- Brandeis provides this shuttle service for traveling across campus or to downtown Waltham, Cambridge and Boston.
- On the Brandeis campus, all students, faculty, staff and guests are required to observe the university's policies on physical distancing and mask-wearing to support the health and safety of all classroom participants. Review up to date COVID-related health and safety policies regularly

Student Support

Brandeis University is committed to supporting all our students so they can thrive. If you want to learn more about support resources, the Support at Brandeis webpage offers a comprehensive list that includes these staff colleagues you can consult, along with other support resources:

- The Care Team
- Academic Services (undergraduate)
- Graduate Student Affairs
- Directors of Graduate Studies in each department, School of Arts & Sciences
- Program Administrators for the Heller School and International Business School
- University Ombuds
- Office of Equal Opportunity