BUS-212A-1: Advanced Data Analytics

M/W, 5:40 – 7:00 pm, Lemberg TBD

Instructor:

- Ahmad Namini, PhD, anamini@brandeis.edu
  Office: Sachar 017

OVERVIEW

This is a four-credit course that teaches theory and best practices of modern data science/machine learning with attention to business intelligence, as well as predictive and prescriptive modeling. This course covers data preparation and visualization and provides hands-on experience with major models currently used within industry. The course will rely heavily on packages, tools and software paradigms utilizing the Python programming language. Students will become experienced in methods of reproducible research, creation of models, and professional communication of results.

Students will be exposed to major techniques currently applied to business data, always in the context of real problems and the search for business value. This course will regularly move between theory, practice and coding, working through the challenges of technology and managing data. In addition, students will develop habits of efficient and reproducible data science workflow.

LEARNING GOALS

Upon successful completion of this course, students will:

- Understand the challenges of using data analytics to provide business value.
- Appreciate the different types of machine learning projects.
- Recognize the uses and differences in training, validation, and test data subsets to carry out machine learning.
- Detail methods for data exploration and data wrangling to permit valid machine learning model building.
- Understand the differences between supervised and unsupervised machine learning models.
- Use machine learning models to appreciate a model's applicability, theory, and usage in a real-world setting.
- Select, fit, and evaluate machine learning models to understand performance.
- Apply models to understand natural language processing.
- Apply models to understand deep learning (neural networks).
- Demonstrate habits of reproducible collaborative project development.
- Prepare clear, informative, and professional reports utilizing visualization and narrative techniques.

Success in this course is based on the expectation that students will spend a minimum of nine hours of study time per week in preparation for class.

COURSE REQUIREMENTS

Book(s): There are suggested books for this course. These books help students to understand the various coding examples used in the course, however, many other books exist that explain theory and practice.

- Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python, Galit Shmueli, Peter C. Bruce, Peter Gedeck, and Nitin R. Patel, John Wiley & Sons, 2020.
- Thoughtful Machine Learning with Python, Mathew Kirk, O'Reilly Media, Inc., 2017.
- Grokking Deep Learning, Version 10, Andrew W. Trask, Manning Publications, 2017.
- Introduction to Machine Learning with Python, Andreas C. Muller and Sarah Guido, O'Reilly Media, Inc. 2017.
- Feature Engineering for Machine Learning, Alice Zheng and Amanda Casari, O'Reilly Media, Inc., 2018.

Please note that the Brandeis University Library permits full access for Safari Books online free to all students. The portal can be accessed at here.

Prerequisite: A functional Python and probability/statistics background is required. All students should have access to a computer and obviously know how to use it.

Class Participation: Class participation is expected of **everyone** in this course, and **class attendance is required**.

Communications: We'll make regular use of LATTE. All lecture notes, handouts, assignments, and supporting materials will be available via LATTE, and any late-breaking news will be communicated via email. Please check your Brandeis email and the LATTE site regularly to keep apprised of important course-related announcements.

Approach: This course is designed to provide students with an understanding of techniques for managing information abundance and for attending to the costs and benefits of information processing in decision contexts. As such, always consider the typical decision-making situations in business settings and work towards making data-driven decisions. Readings focus on the theory of decision-making, data structure, and analytic models. In addition, articles and cases illustrate typical decision problems and the application of the techniques that are discussed. Python code will be available for all models and techniques utilized.

Technologies: Throughout the course, students have the choice of using the public computers at IBS and/or personal laptops. Ideally, students should bring their laptop to class, but use it only for course-related software. Tablets and other mobile devices will not be reliable tools in the course, and nearly all of the software provided should run on either Windows or Macs.

Special Accommodation. If a student has a documented disability on record at Brandeis and wishes to have a reasonable accommodation made, please see the Instructor immediately. Please keep in mind that reasonable accommodations are not provided retroactively.

Grading

- **Assignments - 100**%

Assignments. For each assignment, students will do analysis and document their results in a reproducible, well-documented report. The correctness (quality) of the results is important, of course, but so is completeness. Completeness will include visualizations and interpretation of results along the way. Finally, a report that explains the entire project should be submitted. The report should have an overall conclusion of what has been learned from the project.

For a successful project, a student should consider the following when creating the report:

- Business context and problem definition, actual or hypothetical.
- Insights pertaining to the data.
- Business benefits, limitations, or tradeoffs of the tool and analytics techniques: R*equirements*, e.g., Clarity, Accuracy, Complexity, Interpretability, and Explainability. These *Requirements* may pertain to the Problem, the System (Data and/or Analytics), the Users, the Organization, the Management, or some combination thereof.
- User reflections about the analytics tool and techniques, in terms of tool functionality, usability, flexibility, persuasiveness, and overall interpretability vs. accuracy of models;
- What additional data would be good to gather and integrate with the data for a more extensive analysis?

Reflections should consider including the following:

- *Domain-specific reasoning*, e.g., basketball players who score a lot of points are good, but we need to figure out their ball passing patterns to really understand their contribution to the team.
- *Modeling process-specific reasoning*, e.g., the linear regression fitted line showed this result quickly and easily, but there are a lot of outliers. Maybe a nonlinear model would fit the data better, after a lot of iterative modeling, without overfitting the data.
- *Language/tool-specific reasoning*, e.g., Python machine learning tools show a simple model with a small RMSE, but a different tool shows an ensemble model with smaller RMSE.
- Good communication of the project statement, methods and conclusions should be coherent, complete, and persuasive. Remember to check spelling, grammar, and readability. Good writing and visual presentation counts.

Academic Honesty. All students are expected to be honest in all academic work. Please consult Brandeis University ***Rights and Responsibilities*** for all policies and procedures related to academic integrity. Students may be required to submit work to TurnItIn.com software to verify originality. Allegations of alleged academic dishonesty will be forwarded to the Director of

Academic Integrity. Sanctions for academic dishonesty can include failing grades and/or suspension from the university. Citation and research assistance can be found at LTS - Library guides.

COURSE OUTLINE

Course Contents

1. **Types of Machine Learning Projects**
2. **Probability and Statistics**
3. **Machine Learning Modules with Python**
4. **Overview of the Machine Learning Process**
5. **Data Exploration and Data Wrangling**
6. **Data Visualization**
7. **Feature Engineering and Dimensionality Reduction**
8. **Supervised Learning Models - Classification and Regression**
    a. **k-Nearest Neighbors**
    b. **Naïve Bayes Classifier**
    c. **Discriminant Analysis**
    d. **Linear, Multiple and Logistic Regression**
    e. **Support Vector Machines**
    f. **Decision Trees**
    g. **Combining Models**
9. **Model Performance Evaluation**
10. **Unsupervised Learning Models**
    a. **Cluster Analysis**
11. **Advanced Models**
    a. **Natural Language Processing (NLP)**
    b. **Deep Learning – Neural Networks**
12. **Real-World Examples**
13. **Time Series – if time permits**