CS/DS-549 Spark! Machine Learning X-Lab Practicum

Time/Location: Tue/Thu 2:00pm–3:15pm, Fall 2023
Seminars/Meetings: Wed 6:30 – 8:30pm
Location: CDS 164
Course Number: CS/DS 549
Course Credits: 4

Instructors:

Oindrilla Chatterjee (oc@bu.edu)

Office hours: Schedule on https://calendly.com/ochatter/15min

Thomas Gardos (tgardos@bu.edu)

Office hours: Schedule on https://calendly.com/tgardos_bu/15min

Project Managers and TE:

PM: TBA

PM: TBA

TE: TBA

Piazza: https://piazza.com/bu/fall2023/dscs549 (access code to be shared via email)
Please use Piazza, not email, for all questions, including grading or missing class, etc. (use the private message to instructors for such requests)

GradeScope: https://www.gradescope.com/courses/576078
If you don't have access, contact the instructors via Piazza or email.

Course Description

Traditional machine learning (ML) courses underemphasize—or in some cases—eschew machine learning deployment practices and software engineering principles, to ensure beginners focus their attention on developing a solid understanding of ML. While justifiable, this practice perpetuates an ever-widening gap between industry expectations and student skills.

The X-Lab Machine Learning Practicum affords students opportunities to work on real-world, semester-long projects while highlighting architectural, infrastructural, and foundational considerations involved in building and deploying a machine learning pipeline. Ultimately, we hope to bridge the aforementioned gap between computer science and ML engineering, through project-based learning.

We will divide our discussion into 4 parts:

1. Teamwork and project management

2. Problem definition, data preprocessing, and exploratory research

3. Designing and developing ML pipelines

4. Delivery and maintenance

We begin by exploring the various phases of software development methodologies. In this part of the course, we will study software development models and learn to use scrum, a flexible, iteration-driven approach to project management. We will also learn the delicate art of managing client expectations—a task that often involves understanding how software engineers and clients view time estimates differently.

In the second part of the course, we will develop an understanding of our project's needs and constraints and explore possible ML-based solutions. Importantly, we will also investigate and determine the need for machine learning, since not every problem requires (or even benefits from) ML. Having definitively established a need for machine learning, we will learn ways to preprocess and clean data, to understand it better. Here, we will look at techniques to augment sparse data, unearth hidden correlations, and contend with vast datasets.

In the third part of the course, we will focus on the design, architecture, and development of ML pipelines. We will use popular open source tools to build and train machine learning models. We will also focus on applying machine learning techniques to specific domains (e.g. computer vision, natural language processing, etc.).

Finally, we will wrap up our discussion by devising a delivery and maintenance plan. Here we will study continuous delivery and deployment, two distinct but compatible approaches to releasing reliable software in short cycles.

Prerequisites

To ensure that students get the most out of this class, we require students to have taken one of CS 505 (Intro to Natural Language Processing), CS 542 (Intro to Machine Learning), CS 585 (Image and Video Computing) or have equivalent experience. You must have a strong programming background especially with proficiency in Python. Familiarity with web and/or mobile application development is helpful, though not required. Please consult with course staff during office hours if you have questions about the prerequisites. Assignment 1 (the diagnostic test) will help you assess your readiness to take this class.

Hub Learning Outcomes

Hub Unit #1: Teamwork/Collaboration

Learning Outcome #1: *As a result of explicit training in teamwork and sustained experiences of collaborating with others, students will be able to identify the characteristics of a well-functioning team.*

The X-Lab Machine Learning Practicum affords students opportunities to work on real-world, semester-long projects while highlighting architectural, infrastructural, and foundational considerations involved in building and shipping an enterprise machine learning pipeline. Students in CS/DS 549 will explore and apply the various aspects of collaboration necessary for developing an enterprise machine learning pipeline, including developing a team agreement, assigning roles and responsibilities, pitching project ideas, making use of scrum, operating in sprints, pair programming, and presenting a well-designed, functional final project to both their peers and to their clients from industry. Teams will use a team agreement and mid-term review as a framework to establish expectations and provide feedback against those agreements.

Learning Outcome #2: *Students will demonstrate an ability to use the tools and strategies of working successfully with a diverse group, such as assigning roles and responsibilities, giving and receiving feedback, and engaging in meaningful group reflection that inspires collective ownership of results.*

As a team, students will work together to develop an understanding of their project's needs and constraints and explore possible ML-based solutions. Considerable theoretical and practical knowledge is explored and practiced, not only to ensure students are grasping course concepts but also to ensure students are adequately making progress toward the final project. Throughout the course of project development, students will be expected to

provide each other with continuous feedback and implement said feedback to improve the functionality of their project. Students will learn the value of scrum, which ensures that all team members are aware of how each individual's work is progressing. Establishing a consistent meeting cadence as a team and with the project's PM will provide students the opportunity to reflect on their progress, both individually and as a team.

Research & Information Literacy

Learning Outcome #1*: Students will be able to search for, select, and use a range of publicly available and discipline-specific information sources ethically and strategically to address research questions.*

When developing their projects, students will use both publicly and privately-available data sets and employ popular open source tools to build and train machine learning models. Students will be encouraged to read papers and evaluate open source models to identify the machine learning approach most suited to the problem they are seeking to address. They will conduct an ethics assessment to understand the potential risks and areas of bias involved with their chosen models.

Learning Outcome #2*: Students will demonstrate understanding of the overall research process and its component parts, and be able to formulate good research questions or hypotheses, gather and analyze information, and critique, interpret, and communicate findings.*

The semester-long project demands that students formulate and iterate on their topic, which includes: problem definition, data preprocessing, and exploratory research; designing and developing ML pipelines; and delivery and maintenance. In order to adequately analyze data used for each project, students will learn ways to preprocess and clean data, using techniques to augment sparse data, unearth hidden correlations, and contend with vast datasets.The course will wrap up with final presentations to industry partners, delivering the final work product which includes thorough documentation of the code and data before the end-of-semester.

Ethical Reasoning

Learning Outcome #1*: Students will be able to identify, grapple with, and make a judgment about the ethical questions at stake in at least one major contemporary public debate, and engage in a civil discussion about it with those who hold views different from their own.*

Issues of bias, transparency, and fairness in the field of machine learning are gaining widespread exposure in the public sphere. Students will be provided an overview of ethics and responsibility issues in ML along with a framework to assess their work based on a set of ethics and responsibility principles. They will apply an assessment that covers issues of explainability and traceability, i.e. the ability to explain the model's behavior at a high level and for a specific input and the ability to trace how the model was trained including underlying assumptions, acceptance criteria, and performance of the model itself. They will gain experience examining issues of bias in the collection of the underlying data as well as issues of equity and justice in the context of applying the model itself.

Learning Outcome #2*: Students will demonstrate the skills and vocabulary needed to reflect on the ethical responsibilities that face individuals (or organizations, or societies or governments) as they grapple with issues affecting both the communities to which they belong and those identified as "other." They should consider their responsibilities to future generations of humankind, and to stewardship of the Earth.*

Computing and data science students are finding themselves entering workplaces that are underregulated by the public sector leaving a vacuum to be filled by self-regulation and public pressure. Students will gain an understanding of the risks inherent to machine learning models and applications. They will acquire both the language used to describe these risks as well as the processes needed to evaluate them. They will also grapple with determining their own values and personal responsibilities, so they are better equipped to operate responsibility in workplaces and in a world where societal expectations are surpassing government regulation. They will practice this skill through dialogue around their assessments with partners/ clients as well as students and instructors.

Other Outcomes (e.g., School, Department, and/or Program Outcomes)

As a result of completing this course, students will be able to:

1. Plan, execute, and manage complex machine learning projects
2. Create reproducible and deployable ML pipelines
3. Improve teamwork and communication

Instructional Format, Course Pedagogy, and Approach to Learning

In addition to lectures, we will also have team collaboration time. During these times, we will work on assimilating material covered in lecture into our projects. These are meant to be hands-on work sessions.

Books and Other Course Materials

There is no required textbook for this course. Pertinent readings and lecture notes will be posted in the Course Schedule and Piazza.

Courseware

We will be using the typical suite of software tools for this course:

- Blackboard: Used to support current grade status. However, the grade you see in Blackboard will not be completely accurate until the end of the semester as it doesn't take into account participation and peer evaluation.
- Gradescope: Assignment grades and feedback about the assignments
- Piazza: Class discussion and assignment details. Piazza should be where you go first and has links to all information/software used in the course.

Assignments

Assignments serve 2 purposes:

- Cement material learned in class
- Track team and project progress

Assignments due dates will be posted in the Lecture Schedule and in most cases may be submitted up to 24 hours late with a 5% late penalty. No late submissions will be accepted after 24 hours. Assignments must be submitted on Gradescope. To account for emergencies, we will drop the assignment with the lowest score from your final grade calculations. Gradescope due dates will be the final arbiter.

Qualitative Assignments

While several of the assignments are testing grasp of programming content, there are several assignments that are relevant to collaborative work, client engagement, and importantly ethical reasoning. These assignments are explained in further detail below:

Team Agreement
We'll have a discussion on how to facilitate effective functioning of teams in a project-based learning structure. We'll share material on the research grounding for high performance teams and outline the GRPI model of teaming: Goals, Roles, Processes, and Interactions. Students will construct a team agreement following this format and be graded based on

the completeness of the contract in addressing each component. The team will review and revise the team agreement at the mid-point in the semester and provide a final peer evaluation for the individual component of the team grade.

Ethics and Responsibility Assessments

Students will be provided with a foundational framework for assessing potential issues of ethics and responsibility which they will apply in a series of case study assignments and a class discussion. Three additional assignments will provide them with an opportunity to apply this ethics and responsibility assessment to their own project as well as the project of another student team in the class. Students will be asked to engage both their clients/ partners around issues or questions that arise as well as in class discussions with the instructor and other students as well as a guest lecturer specialized in the topic. Students will be graded individually on the completeness of their assessment and the quality of reflection for the two assessments. Students will be asked to meet as a team to develop a mitigation plan that outlines steps they will take and recommend to the partner to address potential issues of ethics and responsibility. We expect students will be able to identify ethical considerations in the following areas:

- Representation and blind spots
- Product/project intent
- Potential unintended harms (e.g. from inaccurate assessments)
- Technical vulnerabilities, limitations, and risks
- Data collection, privacy, storage, and security
- Auditability of algorithms: automation of human processes, testing, and monitoring
- Auditability of models: purpose of models, input data and training risks and bias, ability for human termination
- Disclosures
- Accessibility
- Use of work by other creators

Project Description

While students will be presented with an initial project description from Spark!, they will be responsible for revising and reframing the project description based on a preliminary meeting with the client and any issues that arise with the ethics and responsibility assessment. A key component of the grade will be the students demonstration of their

grasp of the project and ability to present this understanding back to the client along with any concerns or considerations regarding ethics or responsibility issues.

Grading

Grade Weightings

Your final grade will be a weighted sum of grades received in the following categories:

% of Grade	Category	Notes
15%	Assignments	Grading rubrics are available on individual assignment pages. You must pass the diagnostic assessment to receive a passing grade in the class.
15%	Attendance and Participation	Students are graded on their in person attendance, class participation and individual contributions towards the project.
50%	Project	Students are graded on their individual contributions and overall quality of the project (see below)
5%	Peer evaluations	Each student will receive midterm and final peer evaluations from members of their project
5%	Midterm presentation	Student teams will present their projects to the rest of the class. All students on a team will receive the same grade.
10%	Final presentations	Student teams will present their projects to the rest of the class. All students on a team will receive the same grade, barring special circumstances.

Given that this is a practicum, the project is central to this course and is worth 50% of your final grade. We will be partnering with [BU Spark!](#) to work on a semester-long, machine learning project. Projects are sourced from external partners and are complex enough to provide students with real-world ML experience.

Project Grading

Projects will be graded on:

% of Project Grade	Category	Description/Notes
30%	Objectives Achieved	Ultimately the goal of the project is to deliver towards the client's expectations. Did the project accomplish a sufficient number of (possibly revised) objectives?
10%	Client Relationship & Management	• Did the team develop a positive, constructive rapport with the clients? • Did the team clarify client expectations and problem definition sufficiently? • Were project deliverables defined and updated with client input and approval?
25%	Algorithmic Soundness	• Is the training process well documented and reproducible? • Is the evaluation process rigorous and reproducible? • Is the dataset available and any preprocessing or augmentation well documented?
25%	Code quality, organization and reproducibility	• Is the source code for the application well-structured with meaningful folder structure and intuitive filenames? • Is the environment documented and easily recreated? • Is there a top-level README that orients the user to the repository, installation and operating instructions and summarizes results? • Is the project replicable? Has a clean installation been attempted?
10%	Team Process	• Did the team meet periodically and operate in sprints? Was project progress incremental?

Community of Learning: Class and University Policies

Course members' responsibility for ensuring a positive learning environment (e.g., participation/ discussion guidelines):

Integrity & Conduct

We take the [Student Responsibilities](#) guide very seriously and in particular: "civility and respect for others within the University." In this class we should all strive to be the model for what we want our University and industry to be.

Attendance & Absences

Due to the sequential nature of the product creation experience and the goal of completing a product demo by the end of the semester, attendance is required. Missing more than 3 classes may affect your final grade. If you must miss class for any reason, please email ahead of time. Absence from project meetings should be considered equivalent to absence from lecture.

Academic Conduct Statement

Computing is an inherently collaborative endeavor. In most cases, you will find open source projects or code snippets on the internet that you might want to use in your own projects. While this is permitted, you *must* cite your sources appropriately. You are also responsible for ensuring that you have the original author's permission to use their work. The Open Source Initiative maintains an excellent page on [the different types of software licenses](#) and what you can and cannot do with them.

Using code you have borrowed from the internet without permission and/or attribution is an instance of plagiarism, which is a violation of the [Academic Code of Conduct](#). If you are in doubt about whether something might be construed as plagiarism, please check with course staff and in general—err on the side of caution. Remember, source code with no mentioned license is, by default, not available for reuse.

Collaboration on Assignments and Projects:

Unless explicitly stated, collaboration on assignments and projects among teammates is both allowed and encouraged.

Use of Generative AI for course work

Generative AI tools are permitted for coursework, with a strong recommendation to maintain transparency by appropriately citing their usage.

Disability Accommodations:

If you are a student with a disability or believe you might have a disability that requires accommodations, please contact the Office forDisability Services (ODS) at 617-353-3658 to coordinate any reasonable accommodation requests. For more information, please see http://www.bu.edu/disability.

Course Feedback

There will be a formal course evaluation at the end of the term.

We also appreciate feedback at any time. You are welcome to do that via office hours, email, Piazza or you can submit feedback anonymously HERE.

What works for you? What doesn't? Do you have an idea how to improve something?