CPSC 5310: Machine Learning

The Machine Learning Project

The Machine Learning project is a core component of this course and is designed to integrate theoretical concepts with practical application. Rather than focusing on an open-ended topic search, the project emphasizes deep exploration of machine learning techniques within a single, real-world dataset.

The project has two primary goals:

- Demonstrate your ability to execute an end-to-end machine learning workflow, from precisely defining a problem and understanding the data, to building, evaluating, and refining machine learning models within a reusable pipeline.
- Develop deeper expertise in core machine learning concepts and algorithms by experimenting with multiple modeling approaches, comparing their performance, and understanding their assumptions, strengths, and limitations.

The project is intentionally structured yet exploratory. While the dataset is fixed, you are encouraged to explore different modeling strategies, feature representations, and evaluation techniques. The emphasis is on process, reasoning, and insight, rather than achieving the highest possible performance.

Project Objectives

Throughout the quarter, you will:

- Identify and precisely define a machine learning problem, including why it is important and what success looks like
- Understand the data by identifying data types, candidate features, and relevant target variables
- Prepare the data by acquiring tools and techniques to explore, clean, transform, and analyze it
- Conduct exploratory and analytical work that informs modeling decisions
- Identify and justify prospective ML algorithms suitable for the problem
- Build a reusable ML pipeline, including preprocessing, modeling, and evaluation
- Present and interpret results using appropriate metrics and visualizations
- Suggest strategies to monitor, evaluate, and maintain the model in a real-world setting

Dataset Selection

Your team will choose **one** of the following Kaggle datasets for the project:

- **Adult Census Income**
  https://www.kaggle.com/datasets/uciml/adult-census-incomeLinks to an external site.
- **NYC Yellow Taxi Trip Data**
  https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-dataLinks to an external site.
- **Credit Card Fraud Detection**
  https://www.kaggle.com/datasets/mlg-ulb/creditcardfraudLinks to an external site.
- **Spotify Dataset**
  https://www.kaggle.com/datasets/vatsalmavani/spotify-dataset/dataLinks to an external site.

All datasets are publicly available. A free Kaggle account is required to download the data.

Project Requirements

The following requirements apply:

- You must follow a process closely aligned with Chapter 2 of the textbook (*End-to-End Machine Learning Project*)
- The problem must be non-trivial or challenging in at least one way (e.g., data quality, imbalance, scale, bias, temporal structure)
- You must experiment with multiple ML algorithms and fine-tune at least one model
- Emphasis is placed on process, reasoning, and evaluation, not just final performance

Deliverables

There are four project deliverables, four of which are graded:

- **MLP 1: Initial Problem Statement – Big Picture**
  - Define the problem, motivation, and proposed ML approach
- **MLP 2: Mid-Term Report / Progress report**
  - Baseline model(s), initial feature analysis, early results, and challenges
  - The Progress Report is a mid-quarter checkpoint intended to assess your project's trajectory and ensure that you are on track to deliver a complete and well-reasoned machine learning solution by the end of the quarter.
- **MLP 3:**
  - **MLP 3 b: Final Presentation: slides**
    - Model comparisons, evaluation metrics, insights, and conclusions
  - **MLP 3 a: Final Written Report**
    - Refined models, fine-tuning results, and a reusable ML pipeline
    - **Overall Goal**: Your final deliverables should demonstrate that you can:

1. Frame a real-world problem as an ML task
2. Apply appropriate models and evaluation methods
3. Communicate results clearly to both technical and non-technical audiences
4. Building a reusable pipeline that will be maintained and monitored
5. Reflect critically on your work and learning process

- **MLP 4: Post-Project Reflection**
    - Reflection on experimentation, evaluation choices, limitations, and lessons learned

- Deliverables (1) through (4) are graded work.
- Due dates are posted on Canvas.

Presentation

- Presentation Quality Expectations
    - Slides should be:
        - Concise
        - Visually engaging
        - Focused on figures, diagrams, and charts
    - Avoid dense text blocks.
    - Every slide should clearly support your narrative.
- Timing and Audience
    - Plan for approximately 10 minutes.
    - Assume a non-technical audience:
        - Emphasize the problem motivation, real-world relevance, and impact.
        - Explain your ML approach intuitively (what the model is doing and why it helps).
        - Technical details and metrics are encouraged, but they should support the story, not overwhelm it.

Use of External Work

Researching existing work is expected and encouraged, as most real-world ML problems have been studied before. However, the following rules apply:

- You must cite all external sources that influenced your project.
- You may draw inspiration and high-level ideas, but:
    - You may not copy code, notebooks, or solutions
    - Your implementation must be your own
- A good rule of thumb:
    - Read external material to understand the problem space
    - Put it away
    - Design and implement your own solution

- Cite the inspiration

Failure to follow these guidelines may be treated as an academic integrity violation.

Note:

The most important is for you to enjoy the experience! This is not supposed to be a high-pressure class or a high-pressure project, so just try to do a beautiful job, and don't worry. If you find yourselves worrying about or not understanding the project, let me know! We can spend class- time discussing your topic, and I'm always happy to talk to you in person about it.