PLA6113 Exploring Urban Data with Machine Learning

EUDML Project deliverables

This course asks you to conduct **ONE data analysis project** for the semester, grounded in available **urban data** and applying some of the advanced **machine learning** techniques covered throughout the course - 1) finding out the relationship between variables and forecasting, 2) classification, 3) pattern recognition, or 4) forecasting and prediction. Working as a group is strongly encouraged to consist of a **maximum of 4** people.

Firstly, the important dates of the project are as follows:

- Project proposal submission: **03/09/2022**, 11:59pm (ET)
  **Please arrange a zoom meeting (or a Wednesday in-person meeting) with me to discuss a topic and data before 03/06/2022*
- Midterm presentation: **03/23/2022** in class
- Midterm packet (paper and presentation slides): **03/25/2022**, 11:59pm (ET)
- Final presentation: **04/27/2022** in class
- Final packet (paper, presentation slides, and technical documentation): **05/04/2022** 11:59pm (ET)

Project guideline

The project constitutes the quantitative research framework that addresses a specific city problem or urban questions of your choosing. The range of potential problems include issues discussed in class or in readings, challenges defined in NYC strategy documents (such as PlanNYC, Vision Zero, etc.), or problems you come across as a planner (e.g. COVID-19 pandemic, social justice, anti-racism, climate change). Projects focusing on the New York area are highly recommended due to data constraints, but other cities/countries are welcome.

Defining a suitable problem, given data and time constraints, is critical to the success of your analysis. Problems and analysis that focuses on or addresses issues of sustainability or social equity and equality or other planning related perspectives should be prioritized. The analysis through machine learning approach is an important piece, but it is more important to demonstrate a logical approach to try to solve the problem identified and communicate effectively how your results can be implemented and lead to operational or policy change.**You should draw data from Open Data platform or other publicly available datasets and clearly state a hypothesis to be tested and the connection between the data you select, the methodology you use, and the problem chosen.**

For the midterm, you will be asked to do the following:

- Project proposal (03/09/2022)

The proposal is a **1-2 page text document** outlining the project scope, research question, the data that are going to be used and the machine learning approaches you're going to be using. Also the (anticipated) individual contributions of each team member should be specified.

- Exploratory analysis (03/23/2022)
  The initial step in your project is to conduct exploratory analysis and develop a concept paper for your proposed topic. Exploratory analysis could include both quantitative methods and qualitative methods, fully using your skills. The midterm concept paper and the presentation should include the following:
  - Problem statement and research questions
  - Literature review of at least three (3) publications
  - Data sources and description
  - Methodological framework development (list of machine learning methods to be used/considered - if you don't know what to use, we can discuss together)
  - Exploratory analysis (e.g. findings from data preprocessing, descriptive statistics, data visualization)
  - Expected impact of findings

  A paper should be **no more than five (5) pages** including figures, tables, maps, and references. Also, you will be asked to give a **7 minutes presentation** during the class on March 23th.

For the final, you will be asked to do the following:

- Main analysis - applications of machine learning
  After the midterm, you should start to work on a primary analysis. **The analysis must implement at least one machine learning algorithm discussed in our class.** You may also use techniques we don't explicitly cover, but do not stray too far from the course topics.

  While you are building a machine learning solution, you should answer, or at least keep in mind, the following questions:
  - What question(s) am I trying to answer? Do I think the data collected can answer that question?
  - What is the best way to phrase my question(s) as a machine learning problem?
  - Have I collected enough data to represent that problem I want to solve?
  - What features of the data did I extract, and will these enable the right predictions or pattern recognitions?
  - How will I measure success in my application? Model evaluation?
  - How will the machine learning solution interact with other parts of my research?

  The final full paper and the presentation should include the following:
  - Introduction: introduction to the urban topic/problem at-hand you are aiming to solve/answer
  - Literature review and theoretical framework
  - Methodologies: clear description of the specific question(s) being asked/hypotheses being tested, data sources, variables, and methodology

  - Results and implications: results (visualized in some manner) and insights/explanations from the analyses. You should also address the implications of your findings for urban operations, policy, and/or planning.
  - Conclusions and Next Steps: conclusion synthesizing your analysis/insights, laying out the limitations and commenting on how you could improve the analysis, etc. with additional data and opportunities for future analysis
  - References: bibliography of works cited in APA format

  A final paper is expected to be **a total of 10 pages**, including figures, tables, maps, and references. Also, you will be asked to give a **10 minutes presentation**.

- Technical documentation (**data and code**)
  Reproducibility is an important concept of data analytics. As a proper practice, you should try to organize your data analytic environment. Please assume that you will collaborate with someone and share structured technical documentation including data and Jupyter notebooks with appropriate markdown or notes. Please submit your folder (compressed) used for the final project analysis. This folder should include at least a data folder and script folder, and Jupyther notebooks should be run without any error.

Grading

Based on the syllabus, full marks of the midterm and the final grades will be a total of 60 (including a project proposal). The projects will be graded on the following criteria:

- Creativity of topic and question-driven aspects (midterm 5)
- Data collection and preprocessing (midterm 5)
- Exploratory analysis (midterm 10)
- Methodological framework and application of machine learning techniques (final 15)
- Is paper clearly written and intelligently formulated? (midterm 5, final 5)
- Communication through the presentation (midterm 5, final 5)
- Is the technical documentation suitable for reproducibility? (final 5)