API 222: Machine Learning and Big Data Analytics

Milestones

Fall 2024

Homework Calendar

Table 1: Homework Calendar

Posted	**Due**	**Assignment**
September 17	September 25	Problem Set 1
October 7	October 23	Problem Set 2
November 6	November 18	Problem Set 3
November 25	December 6	Problem Set 4

Final homework scores will be calculated as the average of the top 3 homework scores.

Final Project Milestones

Below are the set of due dates for each component of the final project as well as descriptions of what is due on the corresponding date

Group Formation

- **Due Date:** October 23

If you are not a PhD student, we recommend you do the final project in groups of 5. PhD students and other students who prefer to work alone may work individually, though we expect the quality of work done by an indi- vidual to be of the same caliber as the quality of work done by a group of 5.

30 final project groups have been created on Canvas, which each currently have no members. Please join the same group number as the rest of the members of your final project group

Project Abstract

- **Due Date:** November 25

An assignment will be created on Canvas for your project title and abstract. Please submit in the text entry of the assignment using the following format, where you replace XXXXXXXX with your own content:

```
Project Title: XXXXXXXX
Abstract: XXXXXXXX
```

Project Presentation

- **Due Dates:** December 6

Due to time constraints, only 10 groups or individuals will have the opportunity to present their project for 15 minutes to the class on December 6. These groups or individuals will be selected by a vote from the class; however, all groups will be required to submit presentation slides by November 27 at 11:59am. The presentations should contain four slides, and one slide should be dedicated to each of the following topics:

1. The problem motivation, including currently used methods to approach the policy question and how machine slearning can add to the existing approaches.
2. The model you chose and why.
3. Main results.
4. Main open questions or concerns that might hinder adoption in the real world.

Final Project Write-Up

- **Due Date:** December 12

The final project write up is an opportunity for you to demonstrate an innovative and thoughtful application of machine learning to a real-world policy problem. The write up should be 10 pages double spaced in size 12 Times New Roman font

1. The motivation for the project.
2. The data you used.
3. The process you used to clean the data, including any new features you added.
4. The methods you applied, including:

- Methods you tried but did not use and a discussion of why you thought to use those methods and why you did not choose them as your final model.
- A clear explanation of the model you used that could be understood by a smart person with no background in machine learning.
- A discussion of your results, including numeric results and a thoughtful analysis of bias and fairness in your model.
- A discussion of what you would expect for out of sample performance of your model.
- The conclusion you drew from your model.
- A discussion of how your model differs from existing approaches to tackle the problem at hand, which may be both qualitative and quantitative.

5. Recommendations for implementing your model in the setting you address.

For PhD students, you are expected to conduct individual research and write a rigorous final term paper (in the format of a journal paper). It is highly recommended that you meet with Professor Saghafian or Jacob Jameson (TF) to individually discuss this requirement.

Grading Rubric

Refer to the grading rubric for detailed assessment criteria.

Table 2: Grading Rubric for Final Project

Criteria	**Points**
Problem Motivation	1
Description of Data	1
Data Cleaning or Feature Engineering	1
Machine Learning Methods Discussion	1
Justification of Final Model	1
Presentation of Results	1
Explanation of Final Model	1
Discussion of Results	1
Contribution to the Field	1
Recommendation for Implementation	1
Total	**10**

Table 3: Project Milestones and Deadlines

Project Milestones	**Due Date**
Group formation	October 23
Project abstract	November 25
Votes for in-class presentations	November 27
Project presentations	December 6
Final project write-up	December 12

Past Projects and Data Sources

Project Title	**Data Sources**
Predicting Salary and Job Growth from Knowledge and Skills	U.S. Bureau of Labor Statistics' (BLS) Occupational Outlook Handbook (OOH) Data
Rental Insecurity: Predicting Evictions in High Risk Areas	Evictions data by Princeton's Evictions Lab, US Census Data
Predicting Entrepreneurship	Kauffmann Foundation on entrepreneurship
Predicting Household Energy Consumption	2015 Residential Energy Consumption (RECS) survey data
Clustering Chilean Municipalities	Sistema Nacional de Información Municipal Data
Looking the Part?: Using CNN to Analyze Bias in Sentencing of Sex-Offenders in Hawaii	Hawaii Sex Offender Registry
Arrests, Citations, and Warnings: Predicting the Outcomes of Police Stops in NC	Stanford Open Policing Project Data
Sentiment Analysis of Movie Reviews	IMDB Movie Reviews dataset
Unmasking Pretrial Risk Assessments	ProPublica data on COMPAS
Whether Benign or Malignant? Time to Automatically Predict the Cancer!	Breast Cancer Wisconsin (Diagnostic) Data from UCL Machine Learning Repository
Gun Violence Triggers: Examining Predictors of Firearm Incidents in the U.S.	USDA Economic Research Service, State-level reports, Politico election reporting, NRA Political Victory Fund
What Can Be Learned from the Happiness Score?	The happiness score dataset by the World Happiness Report (WHP)
Targeting the Poor in Peru: Analysis of the Current Targeting System and Ideas for Improvement	Peruvian Household Survey (ENAHO)
Varieties of Democracy: Predicting Democratic Backsliding and Women's Political Empowerment	Varieties of Democracy Institute Data from the Department of Political Science at the University of Gothenburg Sweden
Predicting Growth Retardation from Intestinal Parasites in Rural Nepal	Data by the Medical Rescue Association of Turkey (MEDAK) in rural Nepal
Farm Size in a Haystack: Imputing Land Holding Hectarage Through Agricultural Inputs	Data from Lowder et al.'s (2016) article in World Development, Food and Agriculture Organization (FAO) data
Predicting Player Values in FIFA 2019 Video Game	FIFA 2019 dataset
Are President Trump's Tweets Useful for Predicting S&P 500 Movements?	Trump Twitter Archive, S&P 500 trade data
Using Machine Learning Algorithms to Improve Customs Fraud Detection in the Philippines	Philippine Bureau of Customs Data
Identifying Poverty from Household Surveys: An Attempt to Develop Simple, Adaptable Targeting Schemes	Survey microdata from the World Bank's Living Standards Measurement Study
Predictive Puke Prevention: Forecasting Failed Restaurant Health Inspections in Boston	Boston's open data on restaurant code violations, Boston 311 data
Can We Predict the Beliefs of CCP Officials by Using Citizen-Survey Data?	The citizen dataset from China
Understanding Global Startup Ecosystem for Economic Policy-Making: A Data Science Machine Learning Perspective	Crunchbase data
Putting Your Money Where Your Mouth Is: The Impact of Proximity to Good Food on Property Prices in Singapore	National Environment Agency (NEA) and the Housing Development Board (HDB) Data
Mitigating Overages in the NYC Citi Bike System	Citi Bike Data

Project Title	Data Sources
Accurately Predicting Response Time for Servicing 311 Requests in the City of Boston	A dataset on 311 engagements made public by the city of Boston
Predicting Bike-Share Use in Buenos Aires	The City of Buenos Aires' open data
Predicting Fuel Poverty in the US	2009 Residential Energy Consumption Survey (RECS) data provided by the Energy Information Administration (EIA)
A Comparative Analysis of Machine Learning Methods for Predicting Progression from Mild Cognitive Impairment to Alzheimer's Disease	The National Alzheimer's Coordinating Center (NACC) Data
Public Health Knowledge and Machine Learning	Demographic and Health Survey (DHS) dataset for the Democratic Republic of the Congo (DRC)
Breaking Barriers: Laws that Increase Women Business Ownership	World Bank data on WBL Index
Identifying Hidden Electoral Geographic Clusters to Support Tailored Campaign Strategies for the 2020 Election	U.S. General Elections 2018 - Analysis Dataset from the MIT Election Data and Science Lab
Predicting Intergenerational Economic Mobility from Google DataCommons	Opportunity Atlas by Raj Chetty, Google DataCommons
Oil Price Forecast Based on Twitter Text of "Market Experts"	WTI 3-month calendar spreads (2014 - Present) data, Twitter text archive
Predicting Portuguese High School Students' Aspirations to Higher Education	The Cortez and Silva survey data
Is There an Association Between Weather and Crime in Boston?	Boston Crime dataset
Yemen Microloans	Data from a microfinance lender in Yemen
Predicting Police Misconduct in Chicago	The Invisible Institute Data about complaints against Chicago police officers
Predicting TNC Tips in Chicago	Chicago TNC Data
Machine Learning as an Affordable Alternative to Estimating Education Quality	Trends in International Mathematics and Science Study (TIMSS) Data
Improving the Efficiency of K-12 Schools in Brazil	Data collected by own on primary and secondary schools of Brazil's Federal District
US Wage and Employment Drivers in Local Economies	Bureau of Labor Statistics (BLS) Occupational Employment Statistics (OES) data
Gun Violence in Virginia: A Descriptive Analysis of County-level Factors	Virginia Department of Health (VDH), Injury and Violence database, American Community Survey 5 Year Estimates, U.S. Department of Agriculture, Economic Research Service, Robert Wood Johnson Foundation County Health Rankings, US Department of Justice Bureau of Alcohol, Tobacco, Firearms and Explosives' database of licenses to sell firearms
A Sense of Hopelessness: Mental Health Analysis Using the Medical Expenditure Panel Survey	Medical Expenditure Panel Survey (MEPS) Data for Social, Economic and Health Research
Predicting Global Technology Startup Success: What Organizational, Management, and Location Decisions Matter?	Crunchbase data
Predicting County-Level Life Expectancy Using Population-Level Socioeconomic Characteristics in the U.S.	Atlas of Rural and Small-Town America and the U.S. Census, U.S. Small Area Life Expectancy Estimation Project (USALEEP)
Relative Importance of the Restaurant Environment with Obesity Prevalence Rates in 500 Largest US Cities Using Traditional vs Machine Learning Models	MenuStat data on U.S. chain restaurants, AggData on locations of restaurant chains, and ESRI US census track boundaries

Project Title	**Data Sources**
[Un]Biased Machines: How Racial Discrimination Permeates Innovation in the Criminal Justice System	ProPublica data on criminal defendants in Broward County Florida
A Study of Workplace Mental Health Issues in the Tech Space	OSMI 2014 survey dataset on attitudes towards mental health and frequency of mental health disorders in the tech workplace
Using Machine Learning to Help Develop and Deliver Appropriate Financial Services for Users at the Base of the Pyramid in Central America	Data from a social-impact-oriented cable and internet provider in Nicaragua for the study
Market Definition in E-Commerce: A Machine Learning Approach Using Hierarchical Clustering to Identify Competitors	Transaction data from the Comscore Web Behavior Database
Can Machine Learning be Used to Better Predict PM2.5 Levels of Air Pollution?	The UC Irvine Machine Learning Repository dataset on atmospheric observations
Predicting Dropout of Public High School Students in the South	Data from the High School Longitudinal Study of 2009 (HSLS:09) from the National Center for Education Statistics
One Step Ahead: Using Machine Learning to Inform Containment Strategies During Ebola Epidemics	The World Health Organization's (WHO) weekly Ebola figures, the World Bank's 2011 Integrated Household Survey, an open-source dataset on NGOs in Sierra Leone
Understanding the Expansion of Social Protection, a First Approximation	The ILO's Social Protection Department dataset