Mengyao Zheng

Boston, MA | mengyaozheng@hsph.harvard.edu (primary) | mengyaoz@mit.edu | (617)866-3815

EDUCATION BACKGROUND

Harvard University, *MSc in Biostatistics (with a concentration in Data Science),* **GPA: 4.00/4.00** **09/2021 – 05/2023 (Expected)**

- **Relevant Courses:** Statistical Inference, Systems Development for Computational Science, Data Science, Machine learning for Healthcare (MIT)

University of Liverpool, *BSc in Financial Mathematics,* **GPA: 3.99/4.00 (Top 1)** **09/2016 – 07/2021**

- **Relevant Courses:** Stochastic Process, Econometrics of Time Series, Financial Engineering, Numerical Analysis
- **Selected Awards:** 2020 National Scholarship, 2021 Provincial Excellent Graduates, Best Overall Academic Performance Award

TECHNICAL SKILLS

Programming & Tools: Python (SciPy, NumPy, Pandas, Scikit-learn, Seaborn, PySpark, Tensorflow, PyTorch), Java, R, SQL, AWS, Linux (Bash), Git, Tableau, Hive

Knowledge: Machine Learning, NLP, Distributed Computing, Statistics (Hypothesis testing, A/B testing, etc.), Casual Inference, ETL Pipeline

PROFESSIONAL & RESEARCH EXPERIENCES

Novo Nordisk Inc – *Advanced Analytics Intern & Co-op*

06/2022 – Present (full-time on-site in summer, part-time online during fall semester) **Plainsboro, NJ**

- Developed an end-to-end solution with machine learning (ML) to identify target patients and related health care providers (HCP) for market potential assessment and marketing outreach analysis, which helped the company save ~$200K in outsourcing consulting costs. Communicated model results and complex analyses to leaders with presentations tailored to both technical and non-technical audiences.
- Constructed 80+ patient-level features out of raw medical claims data with Snowflake **SQL**, performed several data quality checks, conducted EDA and feature selection, established the data pipeline.
- Conducted literature review on *ML using noisy labels*, redesigned and implemented the *Anchor-and-Learn* algorithm, trained and fine-tuned various models (LASSO regression, random forest, XGboost) in **Python**, with the top model achieving AUC of 0.97. Applied the redesigned algorithm to successfully identify potential 700+ target patients out of 32K+ based on 4M+ historical medical claims made by the patients.
- Interviewed clinical experts to understand the disease development, conducted detailed model interpretation & diagnosis, innovatively integrated ML with domain knowledge by applying rule-based prediction validations, which led to acceptance of the final ML model by clinical experts.
- Communicated with sales teams to define and prioritize their needs, refined performance metrics, identified marketing weaknesses, delivered in-depth HCP profiling and segmentation analysis to assist sales outreach optimization. Proactively built a one-stop interactive HCP profiling dashboard in **Tableau** integrating maps, tooltips, filters, sorting buttons, navigation page, etc.

Harvard University, School of Public Health – *Research Assistant*

01/2022 – 06/2022 **Boston, MA**

- Examined changes in US restaurant advertising from 2012 to 2016 and investigated the relationship between those changes and community-level income, race/ethnicity, and population density using longitudinal analysis.
- Proposed and calculated the measure of per capita restaurant advertising (PCRA) integrating data from Census, Kantar Media, and Nielsen Ad Intel. Performed data cleaning, aggregation, and transformation in Python including geocoding all the addresses of restaurants with parallel computing. Visualized the change in restaurant advertising across the US on map, and constructed various mixed effects model in **R**.

Nanyang Technological University, Computer Networks and Communication Lab – *Machine Learning Research Intern*

03/2019 – 10/2019 (full-time onsite) **Singapore**

- Developed *ObfNet* **[GitHub],** a privacy-preserving inference approach that obfuscates the inference data samples while preserving the inference accuracy of a pre-trained deep neural network within the Internet of Things (IoT) edge computing framework. Implemented the *ObfNet* scheme in **Python TensorFlow**.
- Fine-tuned the CNN and MLP models and used data augmentation, batch normalization, dropout, and L2 normalization to mitigate over-fitting. Tested various combinations of CNN and MLP in the *ObfNet* scheme based on FSD, MNIST and ASL datasets in the scenarios of computer vision and voice recognition; all achieved high performance on accuracy maintenance (within 1% drop) and privacy preservation (completely unrecognizable by human).
- Made first-author publications: **journal paper published in IEEE IoT-J** (Impact Factor = 9.9), **conference paper in ACM AIChallengeIoT'19**

Renmin University of China, Big Data Finance Lab – *Research Assistant*

10/2019 – 01/2020 **Suzhou, China**

- Coordinated a team of seven people to design, prototype, test, publish, and promote a free, high-quality online Chinese stock market Fama-French (FF) factor web application. Conducted user research, competitor analysis, and then prepared the Product Requirement Document (PRD).
- Automated the computation of daily and monthly FF factors (FF-3 and FF-5) using **MySQL and Python**, developed the data pipeline, achieved code management and version control using Git, and visualized the time series data using **Python Matplotlib**. Backtested the FF models in Chinese stock market. Led a team to explore the problem of "shell-value distortion" and improved FF models accordingly.

PROJECTS

Context-Aware Legal Case Citation Prediction Using Deep Learning [Presentation and GitHub] **03/2022 – 06/2022**

- Scraped 100K+ legal texts from Harvard Law School's database using **Python Scrapy** and applied data cleaning and tokenization.
- Constructed ML models for citation prediction using both supervised and unsupervised learning methods: 1) built LSTM and CNN models with embedding layers using **TensorFlow**; 2) Leveraged the pre-trained Legal-BERT model to obtain the embeddings and then used the FAISS index to predict the most similar legal case for the in-text citation. The top model (the LSTM model) achieved a 200x accuracy boost as compared to the baseline random model

Causal Inference Analysis on The Effect of Quitting Smoking on Weight Gain Based on the NHEFS dataset **04/2022 – 05/2022**

- Visualized marginal distribution differences of confounders across treatment and control groups using Python Seaborn, built a regularized logistic regression model to estimate the propensity score using **Python Sklearn**, performed hyperparameter selection via cross-validation, deployed the inverse propensity score re-weighting (IPW) and propensity score matching (PSM) to estimate the average treatment effect (ATE).

Dashboard and Visualization for Comparing Job Offers [Website and GitHub] **09/2021 – 12/2021**

- Web-scrapped all the salary and cost of living data in 50 major US cities, cleaned the data using **R tidyverse and dplyr**, created a dynamic **R Shiny** dashboard with interactive visualization tools, and constructed word clouds to visualize the high-paying occupations in various cities. Used GAM, decision tree, and gradient boosting trees to predict the apartment price.

EXTRACURRICULAR EXPERIENCES

Harvard Venture Club – *Chairman at the Web 3 & Metaverse Forum and Community* **2022**

Plug and Play (China) AI Entrepreneurship Hackathon – *Team Leader at the Top 1 team* **2018**

"Psychology of Love" Course – Best *Student counselor (provided love mentoring to 20+ students)* **2017**