JEFFREY YUAN

Glenview, IL | mailjeffreyyuan@gmail.com | 847-905-6888 | Personal Portfolio | Research Repository | Github

EDUCATION

Northwestern University Evanston, IL

MS in Statistics and Data Science Aug 2022 - June 2026 (expected)

BS in Data Science and Computational and Systems Biology, GPA: 3.98/4.00

Thesis: Deepening Drug Discovery with Causal Inference and Generative AI: Multi-Modal Integration Leveraging Large Language Models and Geometric Deep Learning for Novel Compound Prioritization

- Honors: CME Trading Challenge Top 5%, CME AI/ML Team Challenge Winner, MD+ Datathon Runner Up, NU Summer Undergraduate Research Grant, Molecular Biosciences Summer Grant, NU Conference Travel Grant, and Deans List (8/8)
- Relevant Courses: Deep Generative AI, ML on Graphs, Advanced Machine Learning, Information Management, Data Structures and Algorithms, Statistical Theory and Methods, Data Science with Python, Bioinformatics, Biostatistics
- Teaching Assistant: Advanced ML, Data Science with Python, Data Science Project

EXPERIENCES

CME Group, *Year Round Data Science Intern*, Chicago, IL June 2024 - Present

- Designed automated **Agentic AI** for natural language to GCP Bigquery (**nl2sql**) data retrieval for CME Globex order book, Salesforce, Tag50, Marketing Cloud, and Google Analytics databases, reducing development time by **99%**
- Productionalized a full-lifecycle **Machine Learning Operations (MLOps)** solution on Vertex AI for Market Qualified Leads (MQL) prediction, producing **400 MQLs** with an estimated value of **$254,050**
- Developed **deep learning** models utilizing Commitment of Traders, time & sales, and volatility data to predict weekly returns of 10-year treasury note futures, leading to CME publishing COT data on all benchmark products
- Engineered **BigQuery** code repository for CME API customers to perform **TWAP/VWAP** calculations for benchmark products, enable scalability of Data Services product to 100+ customers and achieve **$360K** in annual revenue

Learngle, *Founding Engineer*, Boston, MA Feb 2023 - Mar 2024

- Built **RAG-LLM** (GPT-4) pipeline for AI/ML technology, clinical case, and clinical implementation question generation
- Designed SuperMemo 2 based **adaptive learning algorithm** for personalized content presentation using MongoDB database
- Created personalized user performance analytics dashboard, advanced analytics, and performance report with Tableau
- Integrated into the Massachusetts General Hospital, Harvard Medical School, Clinical Informatics fellowship program

Significance Lab, Harvard Medical School, *Research Assistant*, Boston, MA Dec 2022 - Present

- Fine tuned 144 **statistical learning** models on the MIMIC IV dataset with Azure ML, optimized each model through 5-fold cross-validation and Bayesian optimization. Used predictions and feature SHAP values for enhanced ER resource allocation
- Performed retrospective cohort study on shock index trajectory, categorized patients into 5 groups via **clustering** algorithms, determined effectiveness with the Bayesian Information Criterion - validated with ANOVA and chi-squared tests
- Leading team of 3 undergrads in a **web-analysis** of Clinical Informatics fellowship pages utilizing the Screaming Frog **web-scraping** API and coordinated with Massachusetts General Hospital physicians to develope data extraction template

PROJECTS

Self-Supervised Sequential Recommendation with Graphs

- Engineered **self-supervised graph neural network** framework for **sequential recommendation** with temporal user-item interaction graphs and multi-level sequential encoders , using structural and temporal signals for future interaction prediction
- Designed a **hard negative sampling** mechanism to select non-interacted items most similar to a user's short-term embedding using **cosine similarity**, integrating **InfoNCE contrastive loss** to optimize short-term user representations

LLM Bias in Clinical Reasoning

- Conducted medical bias analysis of a **LLM** (Llama 3.1) utilizing its publicly available API through **HuggingFace,** evaluating performance against 10,179 STEP-1,2, and 3 medical examination questions
- **Tuned LLM** performance using progressive **prompt engineering**, identifying prompt structures with reduced bias

Airbnb Prediction Problem

- Performed exploratory data analysis, feature engineering, and preprocessing on Chicagoland Airbnb data utilizing Python
- Ensembled boosting and bagging models to **95%** accuracy in host status classification and **$9.60** RMSE in price regression

Deciphering Diabetes

- Analyzed Diabetes data from 130 US Hospitals from 1999-2008 to determine the state of Diabetes in the US, demographic and prior medical care contribution to outcomes, and drug efficiency in treating diabetes, utilizing Python
- Presented drug misuse and demographic bias findings to Northwestern Medicine's medical, admin, and business teams

SKILLS

- Technical: Python (Tensorflow, Keras, Pytorch, Scikit-Learn, Statsmodel, Pandas, Numpy), R, SQL, Java, Git, Tableau, GCP
- Analytical: AI/ML, deep learning, data science, feature engineering, statistics, big data analytics, probability