Yilin Zhu

yz3323@columbia.edu | (858) 539-3692 | https://ericzyl.github.io/

EDUCATION

Columbia University, New York City September 2023 – December 2024
M.A. in Statistics

- Relevant Coursework: Computational Statistics, Machine Learning, NLP, Inference, Regression Models, Databases, Bayesian Statistics, Modern Analysis, Data Analysis, Applied Data Science

University of California, San Diego (GPA 3.7/4.0) September 2019 – June 2023
B.S. in Applied Mathematics, Minor in Computer Science

- Relevant Coursework: Mathematical Statistics, Optimization, Advanced Data Structures, Algorithms, Stochastic Process, Computation Theory, Graph Theory, Time Series, Combinatorics, Computer Organization

SKILLS

Software: Python, Java, C/ C++, SQL, R Programming, MATLAB, PyTorch, Stan, HTML/CSS

RESEARCH EXPERIENCE

Columbia Engineering | DitecT Lab June 2024 – Present
Advisor: **Prof. Sharon Di & Zhaobin Mo** New York

Build Robust & Explainable GNN for Spatial-temporal Predictions

- Implemented multimodal mobility nowcasting using EAST-Net and ST-Net models to predict mobility patterns, utilizing datasets representing urban mobility trends in major cities and regional trends during the pandemic.
- Managed data input, normalization, and splitting into training, validation, and test sets for large-scale mobility datasets, configured and executed training scripts for both models using PyTorch, and evaluated model performance with RMSE, MAE, and MAPE metrics.
- Enhanced models with temporal covariates, integrated Heterogeneous Mobility Information Network (HMIN), and applied Memory-Augmented Dynamic Filter Generator (MDFG) for dynamic parameter adjustment.
- Analyzed SHAP values for feature importance and conducted evaluations via t-tests.

Columbia University | Statistics Department & IBM Research March 2024 – Present
Advisor: **Prof. Parijat Dube** New York

Enhance Document Similarity Measures

- Implemented NLP algorithms to enable customized concept importance in queries, improving document similarity measures to help users find specific cases in legal documents quickly and accurately.
- Developed document similarity metrics by tuning SBERT and generating concept overlapping matrix, conducting transfer learning by solving an optimal transport problem with Word Mover Distance approach.
- Designed document similarity metrics with concept distance, formulating into an optimal transport problem.
- Implemented an unsupervised sentence ranking method in long articles using KSVD and PageRank, combine intra-article and corpus-wide relevance to enhance document summarization tasks.

Develop Language Model for Improved Topic Analysis in Police Narratives

- Developed Python pipelines and tuned BERT with UMAP for better topic representation in police narratives, aiding decision analysis and optimizing report accuracy for law enforcement.
- Visualized criminal patterns over time with interactive temporal charts, highlighting shifting themes for trend analysis, and presented results at an industry event to discuss with professionals.

Legal Aid Society | DNA Unit January 2024 – Present
Advisor: **Prof. Daniel Rabinowitz** New York

Access Algorithmic Fairness in Forensic Probabilistic Models

- Analyzed statistical assumptions in forensic models and designed hypothesis tests, equipping attorneys with data-driven arguments, strengthening case strategies and outcomes in juvenile trials.
- Investigated forensic algorithmic integrity, particularly Metropolis-Hastings algorithm convergence, and diagnostic effectiveness across diverse data sets.

University of California-San Diego | Finance Department March 2022 – February 2023
Advisor: **Prof. William Mullins** San Diego

Analyze Celebrities' Influence on Crypto Markets

- Built crypto promotion databases using Twitter and TikTok APIs, identifying positive signals via RoBERTa.
- Conducted regression analysis to access the effects of social media influencers on financial markets.
- Implemented web automation to create SSN database and identified immigrants for research studies.

PROJECTS

Wine Quality Analysis May 2024

- Implemented regressions, ANOVA, and MCMC to assess the impact of physicochemical properties on wine quality, addressing data contamination for accurate statistical analysis.
- Optimized Gibbs Sampler by selectively updating covariance, maintaining data imputation accuracy, and reducing runtime by 60%, accelerating project experiments and timely analysis completion.

Recipes Website Database Application October – December 2023

- Designed E/R diagrams and constructed a PostgreSQL database for a recipe website.
- Engineered a recipe-sharing web platform leveraging Python Flask and SQLAlchemy, integrating functions like user authentication, recipe look-up, uploads, saves, and reviews.
- Created a dynamic user interface with HTML and JavaScript, deploying the application via Google Cloud Platform.
- Implemented a collaborative filtering recommendation system to provide personalized recipes suggestions to users.

Telecom Customer Churn Prediction May – June 2023

- Applied forward feature selection with AIC, designed exploratory data analysis using group bar chart.
- Utilized machine learning methods such as XGBoost, SVM, Regressions, and Random Forest, contributing to accurate customer churn forecasts, and empowered proactive decision-making.
- Evaluated ML models via cross validation to ensure a robust model, achieving AUC score of 0.92.

HIV Transmission Graph Application March 2023

- Developed a C++ Graph generator in C++ to analyze HIV data. This tool read input edge list from CSV files and facilitates essential graph operations including neighbor and edge weight retrieval.
- Implemented Dijkstra's Algorithm and Up-Trees data structure to find weighted shortest paths, connected components, and smallest connecting threshold, identifying transmission clusters and infection pathways.
- Created Huffman Coding Tree to compress files, optimizing the storage of large HIV datasets

Auto-grader Ticket System December 2022

- Employed Minheap structure in Java to create priority queue as a foundational component in the ticket system.
- Developed comprehensive test cases (approximately 500 lines) to thoroughly assess system functionality.

PRESENTATION

- Yilin Zhu. "Developing Language Model for Improved Topic Analysis in Police Narratives" Present on Data Science Day at Columbia University 2024.
- Yilin Zhu. "Statistics in Social Sciences" Present at STEM Graduate Lunch Talk at Columbia University 2024.
- Yilin Zhu. "Alternating Direction Method of Multipliers with Applications" Presented at Statistics PhD Seminar in Columbia University 2024.

TEACHING EXPERIENCE

Applied Linear Algebra (Spring 2023) March – June 2023
Instructor: **Prof. Christian Klevdal** San Diego

HONORS

- Provost Honors 2019 – 2023

VOLUNTEER

- Notes taker in Applied Linear Algebra and Combinatorics for students with disabilities.
- Help organize club tennis try-out.