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, Databases, Bayesian Statistics

University of California, San Diego (GPA 3.7/4.0) September 2019 – June 2023

B.S. in Applied Mathematics / Computer Science

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

SKILLS

Software: Python, Java, PostgreSQL, C++, R Programming, MATLAB, PyTorch, Stan, HTML

RESEARCH EXPERIENCE

Columbia University, Statistics Department March 2024 – Present

Advisor: **Prof. Parijat Dube** New York

- Developed language model for improved topic analysis in police narratives.
- Designed algorithms that capture subtopics within document clusters and calculate document similarities, resulting in a selection of documents that maximize their within diversity with respect to generic topics.
- Built topic a modelling pipeline to process police narrative data and create topics for decision analysis.
- Finetune NLP model with cuML library using dimension reduction methods such as UMAP, and clustering strategies such as k-means and HDBSCAN.

Columbia University, Statistics Department January 2024 – Present

Advisor: **Prof. Daniel Rabinowitz** New York

- Investigated forensic algorithmic integrity, particularly Metropolis-Hastings algorithm convergence, and diagnostic effectiveness across diverse data sets.
- Performed comparative evaluations of differing probabilistic genotyping system (PGS) to determine consistency in likelihood ratio outcomes.
- Reviewed independent scientific committee reports to assess potential biases within DNA sample analysis impacting client defense strategies.
- Conducted literature reviews in forensic science models to identify false assumptions that affects predictions.

University of California-San Diego, Finance Department March 2022 – February 2023

Advisor: **Prof. William Mullins** San Diego

- Built crypto promotion databases using Twitter and TikTok APIs, leveraged the collected promotion data to construct regression models to analyze the effects of financial guidance from social media influencers.
- Applied NLP with RoBERTa model to perform sentiment analysis on Twitter posts, effectively categorizing them into distinct attitude-based segments. This resulted in an 60% enhancement in group working efficiency.
- Constructed web automation to fill 90 forms per minute and to implement web scraping, establishing an SSN database.

PROJECTS

Recipes Website Database Application October – December 2023

Group Leader New York

- 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.

Algorithmic Fairness in Dropout Prediction December 2023

Group Leader New York

- Performed data wrangling and exploratory data analysis on a de-identified students record from a Portugal university.

- Implemented Logistic Regression and Boosting Algorithm and tuned parameters for dropout prediction.
- Conduct statistical testing to analyze the differences in model performance based on metrics such as accuracy and recall.

Telecom Customer Churn Prediction May – June 2023
Group Leader San Diego

- 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.

Portfolio Optimization Project March – May 2023
Group Leader San Diego

- Performed data wrangling to process closing price data from ~1000 companies, deriving normalized return.
- Constructed optimized portfolios using SVD and Gradient Descents, rating portfolios with Sharpe ratio.

Graph Generator Application December 2022
Group Leader San Diego

- Developed a Graph generator in C++ employing tuple embedded unordered map. This tool efficiently 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 with a provided graph.
- Created a Huffman Coding Tree to compress and uncompress input files.

Auto-grader Ticket System March 2022
Group Member San Diego

- 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.