Kangrui Liu

+1 (202)868-9664

krliu67@umd.edu Linkedin/kangrui-liu Github/krliu67

Education

University of Maryland **College Park, Maryland, US**

Master of Science in Survey and Data Science at Joint Program of Survey Methodology *Aug 2023 - May 2025[expected]*

Related courses: Machine Learning, Statistical Modeling, Data Collection, Applied Sampling, Multiple Imputation, Web Scraping, Model-based and design-based Inference, Total Survey Error, Modern workflow for Data Science

Ningbo University of Technology **Ningbo, Zhejiang, CN**

Bachelors of Science in Information and Computing Science at School of Statistics and Data Science *Sep 2019 - May 2023*

Related courses: Linear Algebra, Calculus, Probability Theory, Data Structure and Algorithm, Database Management, Object-Oriented Programming, Web System, Data Visualization Techniques, Mathematical Modelling, Big Data Applications

Experience

Jianxin Technology **Guangzhou, Guangdong, CN**

Intern *Feb 2023 - May 2023*

- Corporate Resource Library Optimization
  - Designed and deployed a MinIO-based file storage and transfer system, leveraging Fast Transfer and Instant Upload features; reduced average upload latency by 30%, enhancing document accessibility and team productivity.
  - Developed a dynamic file prioritization algorithm based on metadata (e.g., file size, classification level), which improved system throughput and ensured high-priority files were processed within 5 seconds under peak load.
- Power Plant Gate Detection Project
  - Labeled over 1,200 images using LabelMe to identify gate status (open/closed) for downstream ML model training.
  - Implemented preprocessing pipelines in Python (e.g., grayscale conversion, bounding box filtering), improving feature extraction accuracy by 18% for initial model iteration.

Projects

Boosted Pseudo-Weighting for Nonprobability Samples to Improve Population Inference

Research Assistant, Supervised by Prof. Yan Li and Dr. Lingxiao Wang *Dec 2023 - Present*

- Developed and implemented the "PS-GBM" pseudo-weight construction methods, integrating gradient boosting method with traditional propensity-score adjustments to enhance the representativeness of nonprobability samples.
- Conducted Monte Carlo simulations to evaluate the performance of these methods, demonstrating improvements in bias reduction compared to traditional logistic regression-based approaches.
- Using the National Health And Nutrition Examination Survey III as a nonprobability sample and the 1997 National Health Interview Survey as a reference, the 15-year incidence of diabetes was estimated with a 20% improvement in relative bias compared with traditional methods.

Reddit Data Analysis – Public Response to Affirmative Action Case

Data Display and Computing *Dec 2023*

- Collected and processed 15,000+ Reddit comments via Reddit API discussing the *SFFA v. Harvard* affirmative action case. Cleaned and tokenized unstructured text data and built document-term matrices for downstream NLP modeling.
- Applied Latent Dirichlet Allocation to uncover 6 dominant discussion topics; performed sentiment analysis using `TextBlob` to categorize emotional valence over time.
- Conducted time series analysis to reveal a 40% increase in negative sentiment following the Supreme Court decision, highlighting polarization trends in public discourse.

Strength Evaluation on Listed Companies of Zhejiang Province

Data Analysis *May 2022*

- Crawled and compiled 2019 annual report data of 418 publicly listed companies in Zhejiang Province using Python and `requests` + `BeautifulSoup`.
- Selected top 3 industries (Manufacturing, Finance, Real Estate) and extracted financial KPIs (e.g., ROE, P/E ratio, debt ratio).
- Conducted Principal Component Analysis (PCA) to reduce dimensionality and derive composite strength scores across industries; visualized cross-industry differences using biplots.