GE (GLORIA) GAO

Tel:+1 (929) 316-3474 | gg2797@columbia.edu | Github

Education

Columbia University — New York, NY
M.Sc. in Business Analytics (STEM) — *Sep 2021 - Dec 2022*

Nankai University — Tianjin, China
B.Ec. in Finance — *Sep 2016 - Jun 2020*

Technical Strengths

Language Proficiency: Native Chinese speaker, Proficient in English
Programming Languages: MySQL, Python, R, Julia, Tableau, Big Data Techniques (PySpark), Bash
Software & Techniques: Git, LATEX, TensorFlow, PyTorch, Spark, Docker

Publication

- Deep Learning-Based Convert Brain Infarct Detection from Multiple MRI Sequences (Submitted)
- DeepMed: A Python Package for semi-parametric causal mediation analysis (In progress)
- Efficient Principal Component Transformer for Long Sequence Prediction (In progress)

Internship Experience

Tencent America — Palo Alto, CA
Data Scientist Intern — *Jul 2022 - Sep 2022*

- Collaborated with cross-functional teams to run A/B testing experiments for international games, including experiment design and analysis; communicated experiment results to the leadership team to assist business decisions.
- Constructed an Uplift modeling pipeline, including feature engineering, relevant sanity checks, modeling (meta-algorithms/casual-forest/ TARNet/CFRNet/DragonNet), and evaluation metrics.
- Enriched measures to target customers in the marketing campaigns.

Alinea Invest — New York, NY
Data Scientist Intern — *Jan 2022 - May 2022*

- Created Enhanced Graph Embedding with Side Information using 9k+ user data and four types of side information; recommended stocks to users and enabled users to maintain optimal portfolios.
- Clustered users into five categories using the K-Means clustering algorithm to validate customer personas.

Oppo Research Institute (AI Lab) — Shenzhen, China
Data Scientist Intern — *Aug 2020 - Oct 2021*

- Constructed a decision-making algorithm in Python; Enhanced post-processing rules for predicting the following app users will launch; increased top-10 accuracy by 7%.
- Developed, implemented, and tuned a Principal Component Transformer; reduced the space-time complexity to O(Lk) and enhanced the context awareness; increased top-5 accuracy by 6% in a real-world dataset.
- Applied Transformer-XL to user interaction logs. Increased learning sequence length by 450% and reduced evaluation time by 300% compared to vanilla Transformer.
- Evaluated the effectiveness of app embedding by setting up a dashboard in Embedding Projector to visualize user embedding with different dimensionality reduction methods (e.g., T-SNE, UMAP, PCA).

Lufax (AI Lab) — Shanghai, China
Data Scientist Intern — *May 2019 - Oct 2019*

- Developed 20+ marketing sales tables using MySQL. Optimized query procedure daily; reduced the query time by 30%.
- Developed an AutoML platform in Python with cross-functional teams (i.e., Business Intelligence team and Data PM team); provided end-to-end marketing solutions; boosted the revenue of 15% for the marketing team.
- Reduced processing time by 50% by creating an automated feature engineering and selection pipeline in SQL and Python to generate informative features (500+ features).

Research Experience

Computational Image Analysis Laboratory, Columbia University

Supervisor: Prof.Binsheng Zhao ;Dr.Sicheng Zhao *Apr 2022 - Present*

- Proposed a novel framework based on deep learning to detect Convert Brain Infarct from multiple MRI sequences,i.e., T1 and Flair; achieved a sensitivity of 80% and less than five false positives per case.
- Conducted cross-sequence Liver registration between CT sequences through body part regression, alignment, and discrete dense displacement sampling.

Semi-parametric Causal Mediation Analysis

Supervisor: Prof.Zhonghua Liu *Aug 2022 - Present*

- Initialized a Python package for semi-parametric causal mediation analysis; estimated the natural (in)direct effects of a binary treatment on an outcome of interest.
- Implemented deep neural networks and other competing methods(Lasso/RandomForest/GBM) to estimate the nuisance parameters involved in the influence functions of the causal parameters.

Review of Big Data Sampling Methods

Supervisor: Prof.HaiYing Wang *Sep 2022 - Present*

- Discussed existing functions for sub-sampling in different software packages (Ptyhon/R/Julia).
- Illustrated underlying algorithms and performed reproducible numerical comparisons.