Yue Wu

12112422@mail.sustech.edu.cn · (+86) 153-3360-6607 ·

Education

Southern University of Science and Technology, Shenzhen, Guangdong 09/2021 – 07/2025

Bachelor of Statistics,GPA: 3.80/4.00, Rank: 4/35

- Main Courses:Statistical Learning(98),Time Series Analysis(92),Operations Research and Optimization(90), Probability Theory and Mathematical Statistics(97),Statistical Computation and Software(91)

Research&Project Experience

Application of Deep Learning in Multi-Agent Financial Simulation 07/2023-05/2024

- Read over **40 relevant papers,** including commonly used calibration metrics and optimization methods in the field, knowledge about financial simulators, understanding of Limit Order Book (LOB) data, and summarized relevant **deep learning architecture**.
- I built an **autoencoder** based on the **transformer** and designed a new calibration method to calibrate the multi-agent LOB financial simulator to address lob data representation and spatiotemporal calibration difficulties, which improved by more than **50%**.
- I designed multi-angle experiments in the field to compare with existing algorithms like TransLOB and DepLOB, designed various visualizations to show the results, and constructed an **ablation experiment** to analyze and interpret the model.

Analysis of Bilibili Content Creators 11/2023-12/2023

- Utilized techniques such as **ANOVA, two-way ANOVA, and ANCOVA** to **quantify the significant impact of factors** such as tags, gender, video length, and their interactions on the number of followers.
- Employed **GMM** fitting to **determine the boundary** between long , medium, and short videos on Bilibili with 6.17 and 18 minutes, and explored the relationship between video length and the number of followers.
- Conducted variable selection using methods such as **RandomForest, Lasso, XGBoost, and DCSIS** to **identify the most relevant variables** to the number of followers for further analysis.
- Utilized methods such as **weighted least squares, generalized linear models, and other 5 models** for **regression analysis** and provided **model interpretations**.
- Used **K-medoids** for content creator clustering and **built underestimation mechanism** to identify the most promising and cost-effective categories and video formats. Provided recommendations for content creator positioning.

Exploratory Data Mining Based on New York City Taxi Data 10/2022-11/2022

- Based on **Pyspark**, 6 million taxi data were cleaned and preprocessed, and **data visualization** was performed using **heat maps, Sankey diagrams**, etc.
- Polynomial regression fitting is used to infer the taxi fare method in New York, and reasonable **vehicle dispatch suggestions** are provided to taxi drivers in various regions to maximize profits.
- Combining the **Spearman coefficient, random forest regression analysis, and XGboost regression** analysis the relationship and importance of different factors (such as period, urban area, number of people) and taxi travel prices were obtained.

Work Experience

xDAN AI:AI Data Engineering Intern 03/2024-05/2024

- Based on NLP tools such as TextBlob and NLTK, clean and screen text data for LLM model training.
- Based on papers such as **WizardLM** and **DEITA**, I parallelly completed data selection and diversification of instructions, effectively expanding dozens of famous datasets with cutting-edge data processing tools such as **Distilable** and **DataDreamer**.
- Based on **UltraFeedBack**, quantitatively evaluate LLM data based on complexity and quality.

- Build relevant prompts based on **CoT, Few-shot In-Context**, and other methods to achieve diversified expansion of different style questions and complete high-quality specified tasks.

Peking University Shenzhen Hospital:Alzheimer's disease image processing 09/2023-02/2024

- Learned, installed, and utilized **Freesurfer** to process image data and perform relevant **feature selection** with Python.
- Used models such as **XGBoost** and other 4 models for binary classification on CU-MCI patients, incorporating cross-validation and **Bayesian Optimization** with optuna for **hyperparameter tuning**.Used ensemble strategy which improved by more than **16%**. Trained and tested the models using data from the hospital and **ADNI**.
- Employed **HDBSCAN** techniques and statistical methods to uncover the underlying features of different groups and **provided medical explanations**.

Extra-curricular Activities

Organizing and Participating in a Self-Supervised Learning Seminar 6/2023-8/2023

- Based on A Cookbook of Self-Supervised Learning , we **organized a seminar** to delve into the **self-supervised learning** and **meta-learning**. I studied relevant papers and conducted multiple knowledge-sharing sessions.
- Seminar Link

Organizing and Participating in a Graph Neural Networks Seminar 3/2024-4/2024

- I **organized a seminar** focused on Graph Learning, based on SC224W and Graph Representation Learning; we extensively studied **GNN, Knowledge Graph, Graph Reasoning, Graph Recommendation** and its applications, read several papers, and learned **Pytorch_Geometric**.

Chairperson of the Student Union of Zhiren (Highest-level Position) 9/2023-Present

- Collaborated with multiple student organizations to organize dozens of campus-level events and received the **First Prize** for Outstanding Student Organization (**Highest Rating**) during the tenure.
- Participated in the filming and design of the college documentary.

Captain of the University Orienteering Team and the Orienteering Club 9/2022-Present

- Organized multiple campus-level events to promote orienteering and the club received a Three-Star Club (**Highest Rating**).
- I won **third place** in the 2023 Orienteering and **fourth place** in the 2022 Orienteering at the Guangdong University Games.

Award

- First Class (First Place) of the Merit Student Scholarship 2023
- iGEM International Competition Gold Medal 2023
- First Prize in the China Undergraduate Mathematical Contest in Modelling in Guangdong Province 2023
- Second Price in The Chinese Mathematics Competitions in Guangdong Province 2021
- Student Cadre Model of Excellence(Top 8 at the university level) 2023
- Outstanding College Student Backbone at the University Level 2022

Skills & Languages

Computer Skills: Python,Java,R,SAS,Hadoop,Spark,Pytorch,Linux,Latex
Language: Native in Chinese,Proficient in English(TOFEL 91).