Yuhan Pu

530-746-1426 | puyuhan@bu.edu |

Education

Boston University Boston, MA, United States
M.S. in Statistical Practice GPA: 3.53 *Aug. 2023 – Jul. 2024*

University of California, Davis Davis, CA, United States
B.S. in Statistics Major GPA: 3.01 *Sep. 2018 – Jul. 2023*

Publication

Ma, C. and **Pu, Y.** (2021) 'Research Progress of Fine-grained Visual Classification: Basic Framework, Challenges, and Future Development', 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Science and Technology Innovation (IAECST), 2021 3rd International Academic Exchange Conference on, pp. 413–419. doi:10.1109/IAECST54258.2021.9695701.

Experience

Consulting Assistant Intern Oct. 2022 – Dec. 2022
Accenture (China) Co., LTD; remote *Beijing*
I helped with follow-up work of a bank, including but not limited to the arrangement of funds, as well as the storage arrangement and protection of data in the cloud.

Trainee in Accountants Assistant Apr. 2021 – May 2021
Baker Tilly International Limited (Chongqing Branch) *Chongqing*
I counted and reviewed the property losses and supervised the legality of tax administration of China Mobile. I asked for their annual property loss audit information list. I created workbooks to compare and determine whether their actions on tax were trustworthy and legal. I found some flaw on their materials and helped them correct their unnoticed mistakes.

Research and Projects

Comparison of LLM for Information Retrieval Tasks | *Python, LLM, G-eval* Sep. 2023 – Dec. 2023
Collaborated Research with Fidelity Investment

- Investigated the efficiency and effectiveness of state-of-the-art LLMs in information retrieval.
- Focused on optimizing performance through advanced prompting techniques and evaluated models across key metrics such as accuracy, response time, and computational cost.
- Organized team efforts to test models using APIs from Huggingface, analyze results statistically, and visualize performance disparities using G-eval.
- Selected a model balancing high accuracy and resource efficiency, contributing actionable insights for real-world implementation.
- This project underscored expertise in LLM evaluation, prompting optimization, and data-driven decision-making in AI application.

Virtual Obstacle | *R, EDA, Causal Mediation Analysis, NN, Random Forest* Oct. 2023 – Dec. 2023
Collaborated Research with Boston University Medical Campus

- Analyzed human movement strategies during obstacle avoidance using comprehensive datasets that included physical characteristics (BMI, age, and leg length) and total foot travel distance (TD).
- Verified the experimental methodology through causal mediation analysis (CMA) by replicating results using manual implementations.
- Applied k-fold cross-validation to ensure robustness in exploratory data analysis (EDA) and statistical modeling.
- Investigated distinctions between obese and non-obese groups and evaluated the impact of individual characteristics on motor responses.
- Found that individual traits, including BMI and movement strategies (adduction, abduction, or direct impact), had predictive relevance for TD.

Text While Walking | *R, EDA, Linear Regression, Decision Tree, Random Forest* Jan. 2024- Apr. 2024
Collaborated Research with Boston University Medical Campus

- Examined the interplay between cognitive tasks and physical movement by analyzing the effects of texting while walking under varying obstacle conditions.

- Conducted a detailed exploratory data analysis (EDA) on gait variables such as stride length, stride width, and step width, alongside typing metrics like speed and accuracy.
- Applied statistical modeling, adjusting for confounding variables, to uncover relationships between physical characteristics (e.g., BMI, leg length) and multitasking performance.
- Utilized interaction analysis and hypothesis testing to validate findings, showing that task complexity and obstacle height significantly affected gait stability and typing efficiency.
- This project bridged cognitive and biomechanical research, contributing insights for safer mobile device usage and improved human-computer interaction design.

Investigating Neonatal Hearing Loss | *R, EDA, CMA, Logistic Regression* Jan. 2024 – Apr. 2024
Collaborated Research with Boston University Medical Campus

- Analyzed a seven-year dataset of 18,676 newborns to investigate associations between neonatal abstinence syndrome (NAS) and hearing loss identified during newborn screening.
- Conducted exploratory data analysis (EDA) and applied causal mediation analysis (CMA) to assess both direct and indirect effects of NAS on hearing outcomes.
- Expanded the study to examine maternal opioid use and its impact on gestational age as a potential pathway to hearing loss.
- Addressed missing data using alternative inference methods and identified significant indirect effects, highlighting the interplay between prenatal substance exposure and neonatal health.
- This project demonstrated expertise in applying statistical methodologies to uncover complex relationships in medical datasets and provided actionable insights for preventive neonatal care strategies.

LLM Applications to Contracts | *Python, LLM, XML parsing, G-eval* Jan. 2024 – May. 2024
Collaborated Research with Fidelity Investment

- Investigated the use of LLMs to automate the comparison of accounting standards and contracts.
- Applied advanced natural language processing techniques, including tokenization, similarity scoring (e.g., cosine similarity for text embeddings), and document segmentation for handling large regulatory texts.
- Developed workflows for preprocessing documents using PDF-to-HTML conversion, text chunking, and contextual embedding via transformer-based LLMs.
- Evaluated model outputs through XML parsing, G-eval metrics, and custom "Xxx Distance" similarity measures to ensure precision in identifying textual changes.
- Presented findings demonstrating LLMs' capability to enhance compliance processes by increasing speed and accuracy while minimizing manual effort.
- This work showcased expertise in document preprocessing, LLM fine-tuning, and evaluation metrics for large-scale text analysis.

Readmission of Patients | *Python, Feature Selection, XGBoost, SVM , NN* Apr. 2024 – May.2024

- Developed and evaluated machine learning models to predict patient readmission within 30 days using the Nationwide Readmission Database (NRD)
- Conducted extensive data cleaning, exploratory data analysis, and feature selection to identify the top 12 predictors.
- Used a balanced dataset after resampling techniques and trained models such as Decision Tree, XGBoost, SVM, and MLP Neural Network with K-fold cross-validation.
- Achieved significant improvements in model performance metrics, with Decision Tree and XGBoost showing strong results in the training dataset, and MLP excelling in generalizability to a separate test dataset.
- This project demonstrated expertise in handling imbalanced datasets, feature engineering, and model evaluation across multiple metrics.

Technical Skills

Languages: R, Python
Developer Tools: Git, Google Cloud Platform, VS Code, Visual Studio, Microsoft Office