Yuhan (Alison) Yao

yuhan_yao@g.harvard.edu | (917) 495-4827 | LinkedIn | Portfolio | Github | Blog
64 Prentiss St, Cambridge, MA, 02140

EDUCATION

Harvard University, School of Engineering and Applied Sciences | *Boston, MA* *Sep. 2022-May 2024 (expected)*

M.S. in Data Science (Cross-registration at MIT) | **GPA:** 3.96/4.0

Relevant Courses: Machine Learning, Advanced Topics in Data Science, Probability and Statistics, System Development

New York University, New York University Shanghai | *New York, US & Abu Dhabi, UAE & Shanghai, China* *Sep. 2018-May 2022*

B.S. in Data Science (AI track) | Minor in Mathematics | **GPA:** 3.94/4.0

Relevant Courses: Natural Language Processing, Reinforcement Learning, Regression Analysis, Databases, Business Analytics

TECHNICAL SKILLS & CERTIFICATIONS

ML & Statistic Analysis Skills: Deep Learning Models, Regression Models, Decision Tree Models, Clustering Models, Time Series Models, Cross Validation, Bayesian Statistics, A/B Testing, Hypothesis Testing, Data Visualization, Exploratory Data Analysis

Programming Languages: Python (Pandas, Numpy, Matplotlib, Scikit-learn, TensorFlow), SQL, KQL, NoSQL, R, HTML, CSS, JavaScript

Platform & Tools: Jupyter Notebook, Git, MySQL, R Studio, Power BI, Azure Cloud, Docker, LaTeX, Microsoft Office

WORK EXPERIENCE

Data and Applied Scientist Intern, Microsoft Corporation, Seattle, WA [*KQL* | *Python* | *Causal Inference*] *May 2023-Aug. 2023*

- Developed an automated framework to conduct correlation analysis and causal inference between continuous and categorical variables that can be easily generalized to 65k+ pairs of node health signals and customer-impacting events in Azure cloud system.
- Analyzed correlation relationships of 3420 samples on 57 signal-event pairs using Python to establish 7 statistically significant processes and leveraged proprietary auto causal inference engine to validate 9 causal links on 3 processes, enhancing node health anomaly detection pipeline with customer impact analysis.
- Engineered and cleansed 650+ billion rows of big data using Kusto Query Language by using statistical methods to work around database and hardware limitations while ensuring analysis accuracy, pioneering a big data handling method for colleagues.
- Collaborated across 4 teams in different time zones and presented internship outcomes to CVPs and 140+ full-time employees.

Data Science Intern, PayPro Global, Remote [*XGBoost* | *Clustering* | *Power BI*] *Feb. 2021-May 2021*

- Constructed and finetuned a customer lifetime value prediction XGBoost model with 85%+ accuracy, assisting the marketing team to refine customer target strategy.
- Utilized Python to implement K-means and Hierarchical Clustering methods to transform customer recency, frequency, and monetary values, which segmented customers into 3 target groups and engineered 8 features.
- Created and designed a Power BI data visualization report featuring 16 plots on webpage template performance, enabling the frontend team to debug hidden template errors and optimize template functionality.
- Presented actionable business insights to CEO and team leader and wrote requested executive summary detailing model mechanism and suggested marketing strategy for senior leadership.

Computer Vision Intern, Hyron Software Co., LTD, Shanghai, China [*OpenCV* | *TensorFlow* | *CRNN* | *YOLO*] *Jun. 2020-Aug. 2020*

- Preprocessed and extracted structural information from driver's license photos using OpenCV and trained a 95%+ accurate CRNN model with TensorFlow to recognize numbers, dates, and 7000+ Japanese characters, increasing efficiency for DMV.
- Implemented an automated pose detector prototype of abnormal behavior for AirPods factory safety check using YOLOv5 model, preventing theft and larceny.
- Collaborated with 7 team members and successfully delivered 2 fully-deployed AI products to clients in 3 months.

RESEARCH EXPERIENCE

Researcher, New York University Shanghai [*Genetic Algorithm* | *Spatio-temporal network* | *Python*] ***Github***

Shuttle Bus Scheduling Optimization based on Spatio-Temporal Network [publication in progress]

- Proposed a tailored Genetic Algorithm based on Python to solve a black-box optimization problem and devised an improved shuttle bus schedule, which reduced cost by 6.82% while satisfying students' demand.
- Formulated a real-life vehicle scheduling problem into 2 variations of Spatio-temporal networks and constructed a non-closed form objective function with 3 real-life constraints.

SELECTED PROJECTS (Full Portfolio)

NL2SQL: BERT-based Model for SQL Generation, New York University Shanghai [*NLP* | *BERT*] ***Github***

- Designed and built a BERT-based slot-filling classification model that converted questions in human language into SQL statements, which enabled non-programmers to interact with SQL databases effortlessly in Q&A scenarios.

Stable Diffusion: Text to Movie Poster Generation, MIT [*Diffusion Model* | *Prompt Engineering*] ***Github***

- Utilized prompt engineering and hyperparameter tuning to further understand the behavior of stable diffusion model and generated movie posters with manga, Chinese painting, and animation styles.

Bechdel Test: Comparing Female Representation Metrics in Movies, NYU Abu Dhabi [*Data Analysis* | *Visualization*] ***Github*** | ***Blog***

- Employed API to obtain Bechdel scores of 9,300+ movies over 150 years, visualized trend of female representation evolution using Python Seaborn, and showcased quantitatively that more females on set translate into better female representation on screen.

Chinese Traffic Sign Recognition, New York University Shanghai [*CV* | *VGG* | *ResNet*] ***Github*** | ***Presentation***

- Created and trained a self-designed artificial neural network on 6000+ images to accurately classify 58 categories of Chinese traffic signs, which outperformed VGG16 and ResNet50.