Wei-Han (Zoe) Hsu

zoe70416@gmail.com | (332) 249-9798 | Website | LinkedIn | GitHub

EDUCATION

New York University   **New York, NY**
M.S in Data Science (GPA: 3.7/4.0)   Sep 2022 – May 2024

Taipei Medical University   **Taiwan**
B.S in Medical Laboratory Science and Biotechnology (GPA: 3.9/4.0)   Sep 2016 – Jun 2020

Core Coursework: Data Science, **Probability and Statistics for DS,** Linear Algebra, **Natural Language Processing, Machine Learning, Big Data, Deep learning, Causal Inference, Computer Vision**

SKILLS

Libraries: NumPy, Pandas, Matplotlib, ML framework (PyTorch, TensorFlow/Keras, sci-kit learn), MLflow
Programming Languages: Python, SQL, R, C++
Tools: AWS, Spark, Hadoop, Git, Snowflake, Docker, Kubernetes, Tableau, PowerBI, Statistical software (SPSS)

WORK EXPERIENCE

New York University IT - Anomaly Detection System for Network Security   **New York**
Machine Learning Engineer- Student Researcher   Sep 2023 - Present
- Developed scripts for the **ETL process**, engineered **unstructured** network data through stream processing.
- Designed/Implemented **cost-effective**, **novel intrusion detection system (AE model)** for full data center monitoring, surpassing traditional methods limited to internet traffic.

Data Scientist - Student Researcher   Feb 2023 - Aug 2023
- Performed exploratory data analysis, preprocessing and feature engineering on a 100GB dataset.
- Built/Tested **SOTA anomaly detection** algorithms (Tensorflow Autoencoder-based methods, unsupervised) achieving 73% F1 score in identifying malicious behavior.
- Implemented **MLOps** best practices using **MLflow, Docker and Kubernetes**.

Medidata, a Dassault Systèmes company - MedidataAI Department   **New York**
Commercial Analytics Intern   May 2023 - Aug 2023
- Leveraged **analytical expertise** to evaluate sales performance for pharmaceutical companies (customers).
- Employed **SQL/Python** for efficient exploratory data analysis, data balancing and data transformation.
- Developed and implemented a **Random Forest model** for revenue prediction, conducting thorough **feature importance analysis** to identify the most effective promotion channels (phone, in-person, or email).
- Anticipated a **3% improvement in targeted customer engagement** through data-driven optimization.
- Created impactful **data visualizations** via **Tableau** to simplify complex information for stakeholders.

aetherAI- Medical device startup   **Taiwan**
Medical Technician/ Medical Data Management   Jul 2018 – Jun 2022
- Conducted **SQL analysis** on 1M labeled data, elevating prediction model precision.
- Pioneered aetherAI's HEMA project and oversaw clinical trial data management and familiar with **ICD-10**.
- **Collaborated cross-functionally** to facilitate clinical validation processes for FDA approval.

Sleep Center of Shuang-Ho Hospital   **Taiwan**
Research Assistant   Aug 2020 – Mar 2021
- Undertook the **preprocessing** and **statistical analysis** of **medical data** (**Python** and **SPSS**)
- Conducted research on multiple predictive modelings (Random Forest, Naive Bayes, LR, SVM, KNN)
- Achieved over **80% accuracy,** predicting the apnea subtype (**multiclassification task)** using **sklearn**.
- Improved internal clinical decision-making using the model and **co-authored** an article in **Diagnostics, 2021**.

PROJECTS

Life Expectancy Prediction | Python, AWS, S3, Sagemaker
- Uploaded and updated the dataset on **AWS S3** for efficient storage.
- Utilized **XGBoost regression** using **Sagemaker** to train a predictive model, obtaining a remarkable 90% r-squared score for life expectancy forecasting.

Recommendation system based on the ListenBrainz Dataset | Python, Spark, Pyspark
- Analyzed large-scale data using **Spark transformations (Pyspark).**
- Implemented **Spark's ALS algorithm**, outperforming the popularity-based baseline by 0.5MAP.