Mengyu Zhang

Seattle, Washington, United States
6264384723 | sam.zhang.069@gmail.com
LinkedIn: https://www.linkedin.com/in/sam-zhang-mengyu/

EDUCATION

Johns Hopkins University, Master of Science, Data Science — May. 2023

UCLA Extension, Data Science Certificate Program — Apr. 2021

University of California, Los Angeles, Bachelor of Science, Applied Mathematics Major, Statistics Minor — Dec. 2020

SKILLS

Programming Languages: Python (NumPy, pandas, TensorFlow, PyTorch, scikit-learn, matplotlib), R, SQL, Spark, MATLAB

Tools: MySQL, SQLite, Tableau, Power BI, PostgreSQL, MongoDB, Microsoft Excel, SAS JMP, AWS Athena, AWS S3, Git

Core Competency: Data Analytics, Data Mining, Data Visualization, Database Management, Time Series Analysis, Statistical Modeling, Statistical Inference, Machine Learning, Deep Learning, Big Data Management, Data Structure

PROFESSIONAL EXPERIENCE

Data Science Intern — California, U.S.
Sikka Software Corporation — May. 2022 – Aug. 2022

- Extract and update required data of over **1 billion** patients' records stored on **AWS S3** with **PySpark** and **Boto3**.
- Connected **AWS Athena** with **PyAthena** and **SQLAlchemy** to implement data management and preview on collected data.
- Conducted **exploratory data analysis** with SQL queries and regular QA test for efficient data retrieval over 1 million records to reduce the data requirements' processing time by 45% for the multiple Data Science teams.
- Constructed **BiLSTM** model with **PyTorch** on **AWS EC2** to predict health indicators for patients' dental health conditions.
- Built and maintained 10 interactive **Tableau dashboard** for stakeholders to track and analyze 50+ technical performance.

Data Science Intern — Beijing, China
Inspur Group Company — Jun. 2021 – Jul. 2021

- Built **Selenium WebDriver** to enable automatic web crawler for collecting and updating over 1 million receipt images' data.
- Developed and deployed **YOLOv4** object detection model for receipt classification with **CUDA** and reached 95%+ accuracy.
- Implemented Python solution to automatically generate weekly visualizations and analytics with update in **Microsoft Excel**.
- Built **PowerBI Dashboards** to visualize and present the product evaluation results to the senior management team.
- Partner with third-party, Sales and inventory teams to define project framework and data souring plan.

Data Analyst Intern — Wisconsin, U.S.
BroadStreet Data Co-operative — Sep. 2020 – Dec. 2020

- Developed interactive **Tableau dashboard** with ability to visualize COVID data by location and real-time analysis.
- Conducted statistical analysis (fourth order Runge-Kutta) to investigate infection rate and exposure from Los Angeles' historical data, identified insights on attributes for building dynamic **SEIR** model to simulate separation of COVID-19.
- Engaged in depth-analysis of industry advancements through scientific journals and self-guided academic exploration.
- Provided ad hoc support to VP in assessing the feasibility of proposals to maximize efficiency by enabling strategic planning to increase technical allocation, assigning top-tier technicians to high-priority projects, ensuring clients' satisfaction.

PROJECT EXPERIENCE

Machine Learning in Institution Name Disambiguation from Scholarly Text | IEEE Capstone — Dec. 2021 – Jun. 2022

- Developed 5 **ETL pipeline** with **PostgreSQL** to collect target features from over 1 million data on **AWS Redshift**.
- Performed **Named Entity Recognition** with **spaCy,** created geolocator with **GeoPy** to impute geographical features.
- Utilized various models including **LightGBM** and **XGBoost** in Python for pairwise classification and comparison, examined model performances using Precision-Recall curve and ROC-AUC, successfully achieving a high accuracy of 90% accuracy.
- Applied **Hierarchical Agglomerative Clustering (HAC)** to center and standardize clusters onResearch Organization Registry.