MIN (MIA) SHI

469-403-7557 ◇ minmiashi@gmail.com ◇ Personal Portfolio ◇ https://www.linkedin.com/in/min-mia-shi/

EDUCATION

The University of Texas at Dallas GPA: 3.95/4.0
Ph.D. Candidate in Political Science – Quantitative Statistical Modeling Focused (*Expected*) Dec. 2024
Master of Science in Business Analytics (STEM) – Data Science & Data Engineer Track May 2024

SKILLS

Programming & Tools: Python, Git, SQL, R, SAS, Stata, Tableau, Power BI, Alteryx
Frameworks & Tools: Django, AWS, Hbase, Redis, Memcached, RabbitMQ, Celery, Amazon S3
Database & Big Data: MySQL, PostgreSQL, Hadoop, Sqoop, Hive, Impala, Pig, Spark
Data Analysis: Machine Learning, Statistical Modeling, Data Visualization, Experimentation
Certificates: Graduate Certificate in Applied Machine Learning, AWS Certified Cloud Practitioner

WORK EXPERIENCES

Data Analytics Intern Jun. 2024 - Present
The Sunwater Institute *North Bethesda, MD / Remote*
Developing scripts to collect data, creating and managing data pipelines, and validate the quality of data.

- Implement web scraping solutions to extract data from various websites and store them in databases.
- Create an ETL process for ingesting the data into the institute's database using AWS S3 and Glue.
- Validate the quality of data by setting data requirements with Pydantic dataclass and BaseModel.

Research Assistant May 2020 - May 2024
The University of Texas at Dallas *Richardson, TX*
Took responsibility for data analysis for 10+ global health/policy projects using advanced statistical models.

- Managed data collection in diverse methods including Qualtrics surveys and web scraping using R and Python.
- Developed 20+ robust statistical models (multi-variable and fixed-effect regression, difference-in-difference, time-series) combined ML models and NLP skills to support correlation and causal inference in research.
- Led a team of five junior assistants, ensuring collaboration and timely project completion and publication.

Data Scientist Student Consultant Aug. 2023 - Dec. 2023
Working for Onyx CenterSource through The University of Texas at Dallas *Dallas, TX*
Led the creation of an AI-driven chatbot, enhancing customer engagement through advanced NLP techniques.

- Employed NLP and MySQL for analyzing and querying an extensive database containing over 10 million entries.
- Achieved 25% improvement in response efficiency and provided 99% accurate predictions using XGBoost model.
- Contributed to a 15% rise in user engagement, increasing customer satisfaction and bolstering company's image.

PROJECTS

Twitter Clone: High-throughput Social Media Backend — *Python, Django* May 2024 - Present
Working on a six-month solo project developing a social media platform's backend using HBase, MySQL, and Redis with Django framework in Python.

- Maximizing query efficiency by storing objects with HBase, MySQL & Amazon S3 based on query complexity.
- Addressing N+1 slow query issues by implementing Redis caching and denormalization.
- Integrating Celery and RabbitMQ to establish asynchronous workers with varying priority levels.
- Implementing a push model for distributing news feeds to followers efficiently.
- Optimizing memory and resource allocation using recursive small batches of asynchronous tasks.

Kaggle Plant Pathology Competition: Leveraging Deep Learning CNNs Nov. 2023 - Dec. 2023
Implemented deep learning models using Python and PyTorch to enhance disease identification accuracy in crops.

- Utilized transfer learning on CNNs with 13042 images in 12 categories, enhancing disease identification accuracy.
- Conducted image transformation, including rotation, flipping, zooming, and noise injections to augment data.
- Fine-tuned ConvNext DL CNN models and achieve 86.8% accuracy, securing a Top 3 ranking in the competition.

Big Data Risk Analysis and Data Visualization for a Trucking Company Aug. 2022 - Dec. 2022
Engineered data visualization dashboards using Tableau, linked to Hadoop, for business risk analysis.

- Processed and analyzed geospatial data with Hadoop, Hive, and Spark, reducing processing time by 40%.
- Developed Tableau visualizations linked to Hadoop and built interactive dashboards for business analysis.
- Conducted linear regression and multivariate analysis, contributing to predictive accuracy by 15%.