MIA SHI

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

WORK EXPERIENCES

Data Scientist May 2025 - Present
Reframe Data Services *North Bethesda, MD*

- Designed and deployed production-ready ML/AI pipelines on AWS (S3, Glue, Lambda, EC2) processing 100+ GB daily, improving efficiency by 40%.
- Built and integrated predictive and prescriptive ML models (logistic regression, gradient boosting, neural networks) into business workflows, achieving 80–90% accuracy across multimodal datasets (audio, video, text).
- Developed and optimized LLM-based NLP pipelines (summarization, Q&A, entity extraction), improving data quality and insight generation for stakeholders.
- Delivered interactive dashboards (Dash/Streamlit) to translate model outputs into actionable business insights for non-technical partners.

Data Scientist Jun. 2024 - May 2025
The Sunwater Institute *North Bethesda, MD / Remote*

- Developed predictive NLP pipelines (speech-to-text + entity extraction) achieving >90% transcript accuracy for congressional hearing data.
- Designed end-to-end ML/ETL workflows in Python, SQL, and PySpark to automate ingestion, transformation, and modeling for large-scale education and policy datasets.
- Implemented automated data validation rules (schema checks, missing values, distribution drift) and anomaly detection on multiple datasets, reducing pipeline failures and transcription errors by 75%.
- Collaborated with cross-functional teams to deliver production-ready predictive analytics and dashboards supporting policy decision-making.

Data Analyst May 2020 - May 2024
The University of Texas at Dallas *Richardson, TX*

- Built and evaluated 20+ predictive/statistical models (logistic regression, GLMs, time-series, NLP) for international political economy, global health and education projects.
- Applied NLP techniques to free-text survey data, contributing to peer-reviewed publications and evidence-based policy recommendations.
- Managed 10+ concurrent projects, leading a team of five research assistants through data collection, cleaning, modeling, and presentation.

SELECTED PROJECTS

AI-Powered Chatbot for Customer Engagement (2024)

- Developed GenAI chatbot using Python + MySQL with NLP and XGBoost, achieving 99% classification accuracy and improving query efficiency by 25%.

Kaggle Plant Pathology Competition (Top 3, 2023)

- Applied CNN transfer learning (ConvNet) on 13k+ images, with augmentation techniques boosting classification accuracy to 86.8%.

Big Data Risk Analysis with Hadoop & Tableau (2022)

- Processed large-scale geospatial datasets using Hadoop, Hive, and Spark and delivered interactive Tableau dashboards to support business risk analysis.

EDUCATION

Ph.D. in Political Science (Quantitative Methods & Data Science) — UTDallas — GPA: 3.95/4.0
M.S. in Business Analytics (Data Science & Engineering Track) — UTDallas — GPA: 3.95/4.0

SKILLS

Programming & Tools:Python, SQL, PySpark, R, Git, Docker, AWS (S3, Glue, Lambda, EC2), Streamlit, Dash
ML & AI:: Predictive & Prescriptive Modeling, Logistic Regression, GLMs, Time Series, Decision Trees, Gradient Boosting, Random Forests, Neural Nets, NLP, LLMs
MLOps & Systems: FastAPI, MLflow, Airflow, CI/CD, Model Serving & Monitoring, Cloud ML (AWS, Google)
Database & Big Data: SQL Server, PostgreSQL, MySQL, Spark, Hadoop, Hive, OpenSearch
Specialized:Feature Engineering, LLM Prompt Design/Evaluation, LangChain/LangSmith (familiar)
Certificates: AWS Certified Cloud Practitioner, Certificate in Applied Machine Learning