MIN (MIA) SHI

Teaching Assistant, School of Economic, Political & Policy Sciences at UTD
Cecil H. Green Hall | 800 West Campbell Road, GR 3.314 | Richardson, TX 75080-3021
Min.Shi@utdallas.edu ◇ minshimia@gmail.com
UTD Profile ◇ Personal Portfolio

Education

The University of Texas at Dallas — **Aug. 2019 – Dec. 2024**
Ph.D. Candidate in Political Science – Quantitative Statistical Modeling Focused — *GPA: 3.95/4.0*

The University of Texas at Dallas — **Aug. 2021 – Aug. 2024**
M.S. in Social Data Analytics and Research — *GPA: 3.95/4.0*

The University of Texas at Dallas — **Aug. 2022 – May 2024**
M.S. in Business Analytics (Data Science & Data Engineer Track) — *GPA: 4.0/4.0*

The University of Texas at Dallas — **Aug. 2022 – May 2023**
Graduate Certificate in Applied Machine Learning — *GPA: 4.0/4.0*

The University of Texas at Dallas — **Aug. 2019 – May 2022**
M.A. in Political Science — *GPA: 3.95/4.0*

Shandong University — **Sept. 2016 – Jun. 2019**
M.L. in International Politics — *GPA: 88.78/100*

Daito Bunka University — **Sept. 2017 – Aug. 2018**
Exchange Student in Political Science

Shandong University — **Sept. 2012 – Jun. 2016**
B.A. in Japanese — *GPA: 87.37/100*

Scholarships

Keith Lankford Tayer Fellowship — **2024**

John Forrest Kain Scholarship — **2023**

Government and Political Science Scholarship — **2022**

Work Experiences

The Sunwater Institute *Data Analytics Intern* — **May 2020 –May 2024**
↪ North Bethesda, MD / Remote

- Developed scripts to collect data, created and managed data pipelines, and ensured data quality.
- Implemented web scraping solutions to extract data from websites, storing over 1 million records in databases.
- Created ETL process for ingesting data using AWS S3 and Glue, boosting data processing efficiency by 40%.
- Automated speech-to-text and speaker identification using AWS Transcribe, achieving over 99% accuracy.

The University of Texas at Dallas *Graduate Research Assistant* — **May 2020 –May 2024**
↪ Richardson, TX / Part-time

- Took responsibility for data manipulation and model building for 10+ global health and policy analytics projects.
- Directed the data gathering processes, utilizing diverse methods like sampling, surveys, and web scraping.
- Developed robust statistical models, including multi-variable regression, fixed-effect regression, difference-in-difference, and time-series models, to facilitate correlation and causal inference studies.
- Oversaw a team of over five junior research assistants, ensuring smooth collaboration and timely completion.

Onyx CenterSource *Data Science Student Consultant* — **Aug. 2023 – Dec. 2023**
↪ Dallas, TX / Consultant

- Lead a team to develop an AI-powered online chatbot using SQL, NLP, predictive models, and radio UI interface.
- Utilized NLP and MySQL to analyze and query a vast database of 10 million+ rows, with high ETL efficiency.

- Upgraded payment info query service, reduced response times by 99%, provided prediction for payments.
- Contributed to a 15% increase in user engagement, improving customer satisfaction and company's reputation.

Lucion Technology Corp., Ltd. *Marketing Data Analyst* **July 2017 – Aug. 2017**
↪ Jinan, CN / Intern

- Served as a Data Analyst Intern responsible for data management, data visualization, and business analysis.
- Improved the efficiency of data extraction by 40% through data optimization in MySQL.
- Employed Microsoft Visio to visualize intricate network structures and aided in product comprehension.
- Produced Business Intelligence (BI) reports, offering insights based on user structures and competitor analysis.

Data Science & ML & Backend Projects

Twitter Clone: High-throughput Social Media Backend **May 2024 - Present**

- Overall: Optimizing a social media backend using technologies like HBase, MySQL, and Redis.
- Maximizing query efficiency by storing objects with HBase, MySQL, and 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.

US Top 4 Airlines Financial Performance Analytics **Jan. 2024 - May 2024**

- Analyzed financial data from a 20-year dataset of over 10,000 rows, covering net income, revenue, and expenses across the US airline industry. This deep dive provided insights into long-term financial trends and shifts.
- Conducted financial performance analytics for the top 4 airlines, identifying key turning points related to major events, alliances, and partnerships over the period.
- Assessed operational trends and competitive positioning of each airline, deriving specific business model recommendations based on a two-decade comparison with competitors.

Kaggle Plant Pathology Competition: Leveraging Deep Learning CNNs **Nov. 2023 - Dec. 2023**

- Overall: 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.

Forecasting Stock Prices Through NLP Examination of Newspaper Articles **May 2023 - Dec. 2023**

- Developed automated web scraping for 7,000+ WSJ articles, increasing data acquisition efficiency by 30%.
- Employed various vectorizers for WSJ article analysis, such as Tfidf Vectorizer, n-grams Count Vectorizer, etc.
- Utilized Naïve Bayes and Random Forest models, enhancing S&P 500 prediction accuracy by 12%.

Analysis of the Effect of COVID-19 on US Trade and US Firms **May 2023 - Jul. 2023**

- Synthesized data and created fixed-effect regression models to identify correlations and causal mechanisms.
- Developed and Implemented machine learning and deep learning models to conduct counterfactual analysis.
- Presented research at the 2023 Applied Data Science International Conference to 200 professionals, receiving recognition for clarity and actionable insights.

Analytical Insights and Marketing Strategy Guidance for a Food Company **Feb. 2023 - May 2023**

- Handled data cleaning over 1.3 million raw data records using Python, ensuring data quality and accuracy.
- Developed interactive dashboards in Tableau, enhancing data accessibility and supporting business analytics.
- Employed SAS to construct regression and time series models, leading to a 15% increase in forecasting accuracy.

Big Data Risk Analysis and Data Visualization for a Trucking Company **Aug. 2022 - Nov. 2022**

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

Payroll Management System Database Design via MySQL **Jun. 2022 - Aug. 2022**

- Led a group of five in conducting business requirements analysis and designing a payroll management database with MySQL consisting of 13 tables.
- Created stored functions, procedures, and triggers to calculate employees' payroll per two weeks, fill in new employee information, and send PTO reminders automatically.
- Performed extract-transform-load, data cleaning, and query optimization.

Modeling U.S.-China Trade War's Effect on US Firms using ML and Time Series **Jan. 2022 - May 2022**

- A project aimed at exploring how the US-China trade war affects Multinational Corporations (MNCs) through an ML content analysis of policy changes and a time series GARCH modeling approach using stock data.
- Utilized Pandas, NumPy, Matplotlib & Seaborn in data cleaning, visualization, and transformation.
- Leveraged sentiment analysis to explore how the US frame 2018 US-China trade war
- Applied regression analysis in exploring the causal mechanism between trade war and S&P 500 revenues.
- Built machine learning (ML) models in predicting the profound influence of the trade war on US firms.
- Used time-series GRACH models to evaluate MNCs' revenue & volatility quantified via stock data in Stata.
- Presented at 2022 International Society for Data Science and Analytics Conference.

Content Analysis of News Coverage about US-China Trade War **Aug. 2022 - May 2022**

- Led an analysis on how news organizations frame the 2018 US-China trade war during the 2018-2022 period.
- Leveraged machine learning skills such as top modeling and sentiment analysis to explore a collection of over 500 news articles.
- Implemented time-series analysis and chi-squared test in modeling sentiments change tendencies among news coverage.

COVID-19 Worldwide Cases Synchronous Dashboard using Tableau **Dec. 2021 - Jan. 2022**

- Designed a synchronous Tableau dashboard with advanced interactive functions to explore the COVID-19 severity.
- Built a Tableau story to dig into the factors affecting the severity of COVID-19 by country and found out the deep connection between multiple aspects of factors with COVID-19 severity.

Selected Course Work

Data Science	Data Management	Data Modeling
Deep Learning	Big Data	Predictive Analytics for Data Science
Natural Language Processing	Cloud Computing Fundamentals	Modeling for Business Analytics
Causal Analytics and A/B Testing	Database Fundations for BA	Regression and Multivariate Analysis
Programming for Data Science	Information Management	Applied Data Analytics with Python
ML for Socio-Eco and Geo Data	Data Collection	Applied Regression
Content Analysis using ML	Data Visualization	Introduction to Quantitative Methods
OOP in Python	Digital Consulting Project	Social Science Research Methodology
Data Structure & Algorithm	Practical Practicum Project	Prescriptive Analytics

Technical Skills

Programming	Python, R, SQL, Java, Stata, SAS
Tools	Alteryx, Tableau, Jupyter Notebook, Excel Charts, R Shiny, LaTeX & TeX
Database & Big Data	MySQL, PostgreSQL, Mango DB, Amazon RDS, Hadoop, Sqoop, Hive, Impala, Pig, Spark
Automation	Alteryx, Appian, Accelq, Uipath
Certificates	Graduate Certificate in Applied Machine Learning at UTD, Google Data Analytics, AWS Certified Cloud Practitioner, Alteryx Designer Core Certificate, Appian Certified Associate Developer, ACCELQ Automation Engineer
Languages	English, Chinese, Japanese

Personal Statement

I am advancing my expertise in Data Science and Data Engineering through my education at UTD and my diverse intern and project experiences.

In my recent role as a Data Analytics Intern at Sunwater Institute, I have been honing my skills by creating an ETL process for ingesting data into the Legis1 database and ensuring the quality of IP4 data by setting data requirements with Pydantic and Dataclasses.

As a Data Scientist Student Consultant, I spearheaded the development of an AI-driven chatbot using advanced NLP techniques. This project significantly enhanced customer interaction and increased engagement by 15% through improved response efficiency. My work with NLP and machine learning, particularly using the XGBoost model, advanced the company's customer service capabilities.

Eager to apply my analytical acumen and innovative approach, I am excited about the opportunity to contribute to your team and push the boundaries of what data can achieve in the tech industry.