SHIHAO WANG

GitHub | (352) 871-0803 | shihaow@usc.edu | LinkedIn

EDUCATION

University of Southern California **Los Angeles, CA**
Master of Applied Data Science Expected Graduation: May 2025
- **GPA**: 3.5/4.00

Relevant Coursework: *Database Management, Machine Learning, Data Mining, Data Visualization, Data Science Practicum*

University of Florida **Gainesville, FL**
Bachelor of Arts in Economics, Statistics (Minor) May 2023
- **GPA**: 3.91/4.00 **Major GPA**: 4.00/4.00

Relevant Coursework: *Linear Algebra, Regression Analysis, Probability Theory, Econometrics, Time Series and Forecasting, Business Intelligence and AI, Business Finance, Operational Research, Multivariate Statistics, Public Speaking*

TECHNICAL SKILLS

- **Programming Languages**: Python (*Numpy, Pandas, Pyspark, ScikitLearn, Tensorflow, Pytorch*), Linux Command, R, Microsoft Excel (*Pivot Table, VLOOKUP, etc.),* Stata, Knime , Git
- **Database**: SQL, MongoDB, Databricks, Snowflake, Kubernetes, Docker, Hadoop, Google Cloud, Amazon S3, Sass, Twitter API, Google API
- **Data Visualization**: Tableau, Power BI(*Dax Query*), Microsoft Office, Excel

EXPERIENCE

Navy Federal Credit Union **Virginia, VA**
Data Scientist Intern May 2024 – August 2024
- **Queried and analyzed** over 2MM loan and employee records using **MySQL**, driving **20% more efficiency** in data processing.
- Conducted exploratory data analysis, identified variable distributions and correlations, and executed essential **data cleaning and feature engineering** in **Python** to enhance model accuracy by **15%**
- Developed **production-ready** machine learning models (regression) in **Python** to forecast **member satisfaction scores with 90% accuracy**, identifying key operational metrics like "days to initial contact" and "loan officer call acceptance rate." which reduced operational delays by **20%**
- **Designed and deployed an interactive Power BI dashboard** using **DAX**, enabling dynamic predictions of member satisfaction scores with through real-time parameter adjustments. This provided the workforce optimization and business teams with actionable insights, contributing to a **10% improvement** in achieving **Q3 OKR goals**

National Laboratory of Pattern Recognition **Beijing, China**
Research Data Analyst May 2023 – August 2023
- Automated ETL pipelines (Python, SQL) to streamline data ingestion; conducted exploratory data analysis to identify variable distributions and correlation; performed essential data cleaning and feature engineering processes in **Python**
- Trained machine learning models (**Regression** and **Tree-Based** Model) to predict the sales in **Python**, achieving a **88%** prediction accuracy; provided data-driven recommendations about better market campaign strategy and optimized supply chain operations
- Developed interactive dashboards using **Tableau** for analyzing real-time data about customer characteristics and measuring several business metrics (churn rate,sales distribution) for stakeholder and cross-functional team cooperation

University of Florida **Gainesville, FL**
Data Analyst Research Assistant, Economics Department January 2023– May 2023
- Leveraged 3MM+ data and employed advanced data analytics skills to visualize intricate spatial patterns of wetlands and adjudication decisions across four distinct regions in the US using *Geospatial* packages in **R Studio**
- Developed regression models and quantified tangible impact of proximity to wetlands on real estate property prices, explaining housing price variations (Adjusted-R^2) to the degree of ~67-69%
- Presented research findings to an audience of 50+ professors and students; identified a 15% higher value in housing in relation to wetland; presented final visualizations to a diverse set of stakeholders

PROJECTS

IOT Real-Time Crowd Density Tracker Aug 2024 – Dec 2024
- Led a team of 3 in conducting **market research** across a $125 billion security sensor industry and performed over **20 campus interviews** to assess product adoption potential and pricing strategies, identifying key opportunities and aligning the product with user needs
- Designed an IoT solution handling real-time data streams from **Android** sensor. Using **Termux, thingsboard.io, AWS EC2** and **SageMaker** to implement an **XGBoost** machine learning model with *python* achieving a **95%** accuracy rate in crowd density predictions.
- Developed a full-stack web application using **React.js** for dynamic and responsive user interfaces, **Node.js** and **MySQL** for backend development and data management, and **WebSocket APIs** for real-time data updates. enabling seamless interaction with processed data

Weed Dispensary Distribution on Public Sentiment Towards Weed Aug 2023 – Dec 2023
- Leveraged Twitter API and *Tweepy* library to gather **10K+ tweets** (Twitter API) on cannabis related topics; utilized reverse geocoding technique for spatial dataset and performed text preprocessing on the cleaned dataset in *Python*
- Generated density maps and visualized sentiment trends in **Tableau**; uncovered a positive correlation between the presence of weed dispensaries and people's sentiment towards weed; revealing a **25% higher dispensary approval rate** in positive sentiment regions.
- Employed *PyTorch* and *TensorFlow* frameworks to implement **transfer learning** methodologies, using pre-trained deep learning model *BERT* and *RoBERta*, achieving a **85%** accuracy in tweets sentiment prediction

Spotify Recommender System June 2023 – July 2023
- Transformed raw data to analysis ready tables and built Batch Processing **ETL pipeline** by assembling a dataset of 5,000+ songs curated for personalized song recommendations generator in **Python** and Spotify API
- Explored music streaming data using **Python** to discern distinct user preferences; deployed clustering algorithms for user segmentation, and enhanced recommendation accuracy by analyzing the top songs for individualized insights
- Developed high-performing **classification algorithms** (XGboost, Decision Tree, RandomForest) with grid search hyperparameter tunning, yielding an F1 score of **0.93** which served as the base for delivering accurate and personalized song recommendation