Shuo Wang

Email: wshuo87@gmail.com https://www.linkedin.com/in/Shuo-Wang-PE Phone: + 1-608-886-2413

PROFESSIONAL SUMMARY

2 years of experience in AI/ML including large language models (LLM), natural language processing (NLP), deep learning, and model development and deployment. 8 years leading teams and managing Communicating and projects.
Currently seeking to transition into a Machine Learning Engineer or Data Science role.

EDUCATION

Master of Information and Data Science, University of California, Berkeley GPA:3.93 August 2024

Coursework: Statistics, Machine Learning at Scale, Natural Language Processing, Machine Learning Operations

M.S. in Civil Engineering, University of Wisconsin-Madison December 2016

M.S. in Civil Engineering, Tsinghua University July 2012

B.S. in Civil Engineering, China Agricultural University July 2010

SKILLS

Programming: Python (Pandas, Numpy), R, SQL, Java, C, C++, MATLAB
Cloud/Platform: AWS, Azure, Google Cloud Platform, Git/GitHub, Databricks, Tableau, PowerBI
Databases/Data Engineering/Distributed Frameworks: NoSQL, Neo4j, Docker, PostgreSQL, Hadoop, Spark

DATA SCIENCE PROJECTS

Very Intelligent Portfolio (Capstone Project) | Amazon SageMaker, S3, Kubernetes, PyTorch, Azure, JupyterLab Summer 2024
(Website Link: Very Intelligent Portfolio, GitHub Link: Source Code)

- Developed Very Intelligent Portfolio (VIP), an open-source Transformer-based tool to optimize portfolio weights and recommend hedging strategies by generating context-aware embedding representations for stocks and replacing the traditional statistical-based stock correlation matrix with cosine similarity matrix used in mean-variance optimization.
- Conducted feature selection, to include historical time series of stock returns and fundamentals (Compustat) as model inputs, and future stock returns and volatility (CRSP) as model target. The scope was 500 S&P stocks spanning from 2018 to 2023.
- Utilized PyTorch with GPU-accelerated Amazon EC2 instance to train the model and extract the stock embeddings. The cumulative portfolio returns improved by ~5% on 2024 test data given user inputted stock combinations and risk tolerance, effectively balancing the risk and returns.

NLP Model for Toxic Comment Detection |Google Colab, TensorFlow, Neural Networks, Keras, Scikit-Learn Spring 2024
(GitHub Link: NLP Model for Toxic Comment Detection)

- Benchmarked against traditional NLP techniques TF-IDF vectorization paired with Naïve Bayes algorithm and evolved to incorporate advanced deep learning models like BART, BERT, DistilBERT, ALBERT, RoBERTa and GPT-3 etc.
- Performed comprehensive error analysis to mitigate model biases and refine false positive/negative predictions.
- Explored hybrid architectures combining BERT with CNN/RNN layers for ensemble approaches, balancing computational efficiency with accuracy.

DistilBERT API Deployment | FastAPI, Docker, Azure, Kubernetes, Redis, Poetry, Istio, Kustomize, Grafana, CI/CD Winter 2023
(GitHub Link: DistilBERT API Deployment)

- Designed and implemented a FastAPI application to serve prediction of an NLP model using DistilBERT from HuggingFace for sentiment analysis. Validated the inputs and outputs with Pydantic models and executed testing with Pytest to reinforce application reliability.
- Configured end-to-end application dependencies with Poetry and containerized the API with multi-stage Docker to reduce image sizes.
- Orchestrated the deployment of the application in Azure Kubernetes Service (AKS) using istio and kustomize, etc., and improved inference time by utilizing Redis cache. Performed load testing with K6 and utilized Grafana for insightful monitoring and system dynamics visualization.

Flight Delay Prediction | Google Cloud Platform, Databricks, Spark, Data Pipeline, Statistical Analysis Summer 2023
(GitHub Link: Flight Delay Prediction)

- Developed a Random Forest model for estimating flight delays using historical airline and weather data from 2015 to 2019 (data size: 31M rows and 200 columns). Utilized PySpark for distributed data processing and reduced running time by 5x compared to Python pandas.
- Conducted rigorous feature and model selection process such as time series based cross-validation, hyperparameter tuning. The model outperformed baseline by 5% in terms of ROC AUC.
- Identified key features influencing flight delays such as taxiing duration and suggesting further improvements to minimize operational costs and enhance passenger satisfaction.

Spotify Song Genre Prediction | Kaggle, Python, Scikit-Learn, Keras December 2022

- Led a team of 5 members to develop automated genre classification models for 42,305 songs on Spotify. Enhanced library customization and user experience via achieving 71% accuracy, a 16% improvement from the baseline model.
- Conducted exploratory data analysis including data balancing and feature scaling (MinMaxScaler, StandardScaler) on the raw dataset, and selected 22 audio attributes after data processing and feature engineering.
- Explored and optimized various models including Random Forest, XGBoost, Neural Networks, and Logistic Regression. The Random Forest model was selected as the champion model.

Fashion MNIST Multi-Class Classification | Python, TensorFlow, Neural Networks October 2022

- Processed 60K data points from Fashion MNIST dataset that includes 784 features (a 28*28 greyscale image) and a label from 10 classes. Visualized summary statistics of the features using Python Seaborn.
- Developed a Neural Network model with three hidden layers and tuned the model with various activation functions (e.g., Tanh, ReLu) and optimizers (e.g., Adam, SGD) using Python TensorFlow, achieving an impressive 99% testing accuracy.

Data Analysis and Engineering for a Hypothetical Restaurant | SQL, Neo4j, Relational Databases August 2022

(GitHub Link: Data Analysis and Engineering for a Hypothetical Restaurant)

- Parsed sales nested JSON files to produce CSV files with proper parent/child table linkage. Designed and created staging tables that can accept raw data from CSV files. Validated data in the staging tables using SQL.
- Developed a graph database in Neo4j for BART subway system from queries on stations, lines, and travel times data tables. Identified the shortest paths to improve the efficiency of the restaurant's delivery service using public transportation.
- Designed SQL queries to extract data from sales, customers, and orders data tables to identify zip code areas with the largest population and orders. Utilized Google Maps API to create heatmaps and presented recommendations for delivery service.

CO2 Emissions Analysis | R, Python August 2022

- Conducted an extensive analysis of car transmission impacts on CO2 emissions using Agency (VCA) data (2000-2013).
- Analyzed over 7,000 entries after data cleaning to establish correlations between transmission types, fuel types, and engine capacities with CO2 emissions. Consolidated data by car models to remove bias due to repetition.
- Employed linear regression to demonstrate that diesel-fueled and manual transmission vehicles emit significantly less CO2.
- Identified potential model limitations including collinearity and omitted variables such as the drag coefficient and improvements in transmission technology.
- Explored the impact of regulatory and technological changes over time, highlighting potential shifts in the relationship between transmission type and emissions in recent years.

Voting Difficulty Analysis Project | R, Statistical Analysis, Wilcoxon Test June 2022

- Led a statistical investigation into voter difficulties between Democratic and Republican voters during the 2020 U.S. election, utilizing the American National Election Studies data consisting of over 8,000 survey responses.
- Defined a novel metric for voter difficulty that included both voters and those who intended but failed to vote, enhancing the breadth of the analysis.
- Applied a non-parametric Wilcoxon rank-sum test to rigorously compare voting difficulties, substantiating the hypothesis that Democratic voters faced more challenges.
- Identified a meaningful difference in difficulty levels that may influence election outcomes, advocating for targeted political strategies to alleviate voting barriers.

WORK EXPERIENCE

Managing Technical Consultant, ERM - Environmental Resources Management, Washington D.C. January 2023 – Current

Team Management

- Led the team to successfully complete 30+ projects annually, delivering on schedule and within budget.
- Increased team working efficiency by 10% via coordination between clients, project managers, and teammates.
- Received the Global Recognition Award (top 1 %) by the CEO for outstanding leadership and project execution.

Solar Energy, Electric/Gas Distribution/Transmission Line projects, Maryland, Colorado, California, and Virginia.

- Presented technical concepts to clients, prepared proposals, maintained network, and bid for projects.
- Directed the exploratory data analysis (EDA) and developed prediction models for stormwater management analysis.

Senior Civil Engineer, JCL Consulting LLC, Chantilly, VA October 2021 – July 2022

Data Center Development Projects, Loudoun County and Prince William County, VA. ($1.0B)

- Conducted EDA and developed a scenario planning model for Land Development Design, influencing key client decisions.
- Reviewed and validated forecast models for water and storm design, saving millions of dollars for the client.

Water Resources Engineer, iDesign Engineering, Inc., Calverton, MD August 2018 – October 2021

I-95 Highway Transportation Projects, Maryland Transportation Authority (MDTA), Baltimore and Harford Counties, MD. ($1.1B)

- Corrected Pond Design model enacted by MDTA, saved millions of dollars for client, and received official "thank you" letter.
- Performed data analysis and validation, created construction schedules, and increased contractor efficiency by 50%.

Water Resources Engineer, Endesco, Inc., Rockville, MD December 2016 – July 2018

Purple Line, Prince Georges and Montgomery Counties, MD. ($2.4B)

- Improved forecast model and saved 10% cost for Stormwater Management & Erosion and Sediment Control design.