Boya Zeng

Chicago, IL ● mayazeng@outlook.com ● (608) 622-3620 ● LinkedIn ● Portfolio

EDUCATION

University of Chicago Dec 2024
M.S. in Applied Data Science | GPA: 4.0/4.0
- **Courses:** Machine Learning, Data Analysis, Big Data & Cloud Computing, Natural Language Processing, Data Mining, Deep Learning

University of Wisconsin-Madison May 2022
B.S. in Computer Science & B.S. in Data Science

TECHNICAL SKILLS

- **Programming Language**: Python (Pandas, NumPy, Scikit-Learn, Pytorch, TensorFlow), SQL, R, C/C++, Java, HTML, CSS, JavaScript
- **Machine Learning Algorithms & AI:** Time Series Forecasting (Prophet, ARIMA), LLM (RAG, Vector Database), NLP (Topic Modeling, BERT, NER), Classification & Regression, Statistical modelling, Bayesian Inference, A/B Testing, Deep Learning
- **Data Engineering & Data Visualization:** Tableau, Power BI, ETL, Database (Vertica, DynamoDB, SQL-based DBs), Excel, VBA
- **Cloud Computing & Big Data:** Google Cloud Platform (GCP Dataproc,GCS, Vertex AI), AWS (S3, Athena, EC2, IAM), Azure, Spark
- **MLOps & Deployment:** Git, CI/CD, MLOps, Databricks, MLflow, agile methodologies, Streamlit, Flask

PROFESSIONAL EXPERIENCES (SELECTED)

Royal Cyber Inc. Chicago, IL
Data Scientist Intern (Capstone Project) Mar 2024 – Dec 2024
- Developed a contextual zero-shot product recommendation system using fine-tuned **Large Language Models (LLMs)** and **Retrieval-Augmented Generation (RAG)**, addressing cold-start issues and enabling real-time personalized fashion recommendations in e-commerce platforms via a **Streamlit chatbot** integrated with a **Qdrant vector database** for semantic retrieval.
- Optimized model performance by applying **LoRA fine-tuning** and **4-bit quantization,** reducing training loss by 47.8% and validation loss by 28.7%, while improving recommendation accuracy measured by Hit Ratio, NDCG, Perplexity, BLEU Score, and BERTScore.

Realix AI Chicago, IL
Data Scientist Intern June 2024 – Sep 2024
- Created automated **ETL pipeline** for data stored in **AWS DynamoDB** using **Python and Boto3** to extract, clean and preprocess transcript data generated by user and AI leadership coach for improving our Large Language Model used for leadership coach.

The Trade Desk Shanghai, CN
Business Intelligence Analyst Intern Sept 2022 - Aug 2023
- Deployed an **E2E data pipeline** on **AWS** (Athena, S3) and Tableau dashboards to track audience trends, market share shifts, and competitor overlap beyond Google/Facebook. Enabled real-time insights on Samsung owners engaging with competitors, securing an **$88 million+ deal with Samsung** in the US.
- Lead cross-functional initiatives with **Agile methodologies (JIRA)** to design and implement a web-based automation tools for file format validation and metadata management using **Python**, **JavaScript**, **HTML**, CSS, and **Git**, reducing manual processing time by **90%**.
- Built **ETL** process infrastructure using **SQL** to improve workflow efficiency. Developed and presented visualizations of key internal metrics derived from complex data to non-technical stakeholders and senior leaders using **Power BI.**
- Created a custom ad-hoc reporting platform using **Python Streamlit** to generate tailored performance reports for internal stakeholders.

Tencent Shanghai, CN
Data Scientist Intern Jun 2021 – Sep 2021
- Identify AI market trends across industries with advanced **NLP** techniques to analyze 200K+ unstructured news article, performing topic detection with **LDA** and **BERTopic**, **sentiment analysis** with a custom logistic model, and SpaCy **NER.**
- Delivered sector-based AI trend reports used by internal product leads for competitive tracking and investment targeting.

Alibaba Group Hangzhou, CN
Data Analyst Intern *Aug 2020 - Oct 2020*
- Built a **churn prediction** model (**Random Forest/XGBoos**t) with >95% accuracy, validated through **A/B tests** with **5pp lift** in retention.
- Utilized **PCA**, **K-means clustering** and **Apriori** algorithms to analyze **over 3 million** records of customer transaction data, demographic data, and product SKUs. Deployed a 5-tiers customer segmentation system (classification) based on shopping patterns and provided personalized product recommendations via a searching dashbaord.
- Collaborated with **cross-functional** and **marketing** teams to launch a series of **targeted** campaigns through **A/B testing**, leveraging previously defined **power user** classification. This initiative led to **10%** increase in repeated purchases (**stat sig**).

PROJECT EXPERIENCES

- **Divvy Bike Demand Forecasting**: Predictive modeling using **SARIMA, Prophet, Orbit, LSTM** on four years of Divvy bike usage data (2019-2023). [GitHub Link]
- **Credit Card Fraud Detection:** Deep learning models (Multilayer Neural Network & **CNN)** on an imbalanced dataset of 300k+ transactions (0.17% fraudulent), achieving 97.50% accuracy and 100% precision. [GitHub Link]
- **Amazon Review Big Data Analysis**: Applied **PySpark, Spark NLP, ALS, and K-Means** on 620K+ reviews for sentiment analysis, product recommendation, and customer segmentation. [GitHub Link]
- **Nasdaq Stock Prediction (MLOps):** Built a **Databricks AutoML** pipeline with MLflow & EvidentlyAI for stock trend forecasting and model monitoring. [GitHub link]