Andy (Xiangyu) Cui

xiangyucui@outlook.com | (402)-853-3000 | Linkedin | Portfolio

EDUCATION

Northeastern University — Boston, MA
M.S. in Artificial Intelligence of Khoury College — Dec 2023
University of Nebraska-Lincoln — Lincoln, NE
B.S. in Computer Science of Arts Science College — May 2020

PROFESSIONAL EXPERIENCE

King 7 Club Corp — Jan. 2025-Present
Senior Software Engineer — *Los Angeles, CA*

- Developed and deployed a responsive website using **Node.js** for the frontend and **FastAPI** for the backend, with component-based architecture for modular scalability. Implemented dynamic UI with custom **JavaScript** logic and **CSS** animations to ensure a seamless user experience across devices and languages.
- Hosted static assets via **GitHub**, containerized the full-stack environment with **Docker**, and deployed the application on **AWS EC2**, ensuring consistent and reproducible builds across development and production. Utilized **PostgreSQL** as the backend database to securely manage and store user data.
- To optimize access for both international and Chinese users, implemented intelligent **DNS-based** traffic routing: international traffic is served through **AWS Global Accelerator**, while Chinese users are redirected to a mirror deployment on **Alibaba Cloud**. This global traffic segregation strategy reduced cross-region latency by up to **90%**, significantly improving page load times and user experience across all target markets.
- Integrated **Google Analytics Reporting API** to track traffic and user behavior, enabling the development of interactive dashboards to monitor key metrics such as page views and user engagement across platforms like TikTok, Red Book (Xiaohongshu), and YouTube. This real-time data integration streamlined reporting workflows and automated insight generation, improving backend operational efficiency by **80%**.
- Enhanced backend performance and frontend delivery using **CDN** and caching strategies, and structured API responses in **JSON** format for clean integration.

CAC Auto Group LLC — Feb. 2024-Dec. 2024
Data Engineer — *Southborough, MA*

- Developed and maintained a predictive pricing system for vehicles on CarGurus using **AWS serverless architecture**, enhancing market compatibility and streamlining operations. Leveraged key AWS services including **S3**, **Lambda**, **DynamoDB**, **SNS**, **CloudWatch**, and **Kinesis**, and used **Python** with **AWS CloudFormation** for scalable infrastructure deployment.
- Designed and implemented a **fully serverless** data pipeline to continuously monitor target data sources using **Kinesis streams** and **Lambda triggers**, eliminating the need for traditional polling. This approach reduced infrastructure and processing costs by **80%**, while maintaining high scalability and responsiveness.
- Integrated real-time monitoring to track market data fluctuations, enabling automated detection and adjustment of vehicle prices in response to deviations. This solution boosted daily operational efficiency by **80%** and improved pricing accuracy by over **50%** compared to industry standards.

PROJECTS

Job Recommendation System Design — Jan 2025

- Developed a user interface for job searching using **Axure RP 10**; Applied content-based filtering using **TF-IDF** and cosine similarity, achieving 82% precision in matching user skills to job descriptions; Conducted **collaborative filtering** in **Python** with implicit user feedback, improving recommendation diversity by 18% via matrix factorization.
- Leveraged **deepseek API** to dynamically adjust recommendations based on real-time user feedback; Reduced cold-start bias by 30% through RL-driven exploration of niche roles.

Stock Price Prediction with Deep Learning — Oct 2024

- Collected the historical stock price and other financial assets data on the company of interest; Conducted data preprocessing by applying min-max scaling in **Sklearn** to normalize stock price values, ensuring consistency across the dataset.
- Implemented **LSTM**, **GRU**, and **Transformer models** in **PyTorch**, optimizing hyperparameters (e.g., number of layers, optimization methods) through grid search, increasing model accuracy by 20%; Visualized opening and closing price trends to assess model performance in Python.

Auto QA Chat Agent for Customer via NLP — Dec 2023

- Collected and cleaned large-scale Amazon customer Q&A datasets, ensuring high-quality data for effective AI model training; Developed a conversational AI agent in **PyTorch** leveraging state-of-the-art NLP models, including BERT and GPT2, fine-tuned for question answering (QA) tasks.
- Optimized hyperparameters for **BERT** and **GPT-2** to improve contextual understanding and response generation; Conducted 10-fold cross-validation for BERT and 5-fold for GPT-2, achieving average BLEU scores of 0.9 for BERT and 0.8 for GPT-2.
- Integrated continuous learning capabilities to adapt to evolving customer queries, improving long-term operational efficiency; Successfully automated 70% of routine customer inquiries, significantly reducing operational costs.