Zijun Yi

Applied Scientist | +1 515-357-3516 | Seattle, WA
jimzijun@gmail.com | linkedin.com/in/zijunyi/ | github.com/jimzijun

PROFESSIONAL SUMMARY

Data engineer with 3+ years of experience, leveraging a data engineering background to build and optimize complex models. Expertise in **Hierarchical Bayesian discrete choice models**, **market share simulation**, **recommendation systems**, and **algorithmic optimization**. Proven ability to apply **Deep Learning (PyTorch)**, **NLP (RAG, Bias Analysis)**, and **Computer Vision** in end-to-end projects, from research to production.

WORK EXPERIENCE

Data Engineer | Bases Data Science, NielsenIQ | *Remote, US* 03/2022 – *Present*

- Enhanced the calibration process for a **Hierarchical Bayesian discrete choice model**, analyzing and integrating promotion ratio as a new target to **improve market share simulation accuracy** against real-world data.
- Optimized the core market share simulation engine by **vectorizing NumPy** operations and eliminating inefficient memory handling. This resolved critical memory overflow bottlenecks, **doubling the model's item capacity** and enabling more complex simulations.
- Improved the data quality for an **LLM-based RAG** system by building a **computer vision** pre-processing pipeline. The module uses Paddle OCR and k-means clustering to automatically assess image readability and flag low-contrast text that would degrade the LLM's review generation.
- Owned and re-architected the end-to-end share simulation and calibration platform (Django/Kubernetes), implementing robust data validation, job monitoring, and error handling that **reduced production failures by 80%** and **cut processing time by 30%**.

Research Assistant | Science of Science and Computational Discovery Lab | *Syracuse, NY* 02/2021 – 12/2021

- Investigated systemic gender bias in **GloVe word embeddings** by adapting a **2-alternative forced choice (2AFC)** psychological framework. **Extended this methodology** to also quantify bias in **Amazon's Polly** speech-to-text service, analyzing occupation/attribute associations.
- Developed a **topic modeling** pipeline using **Latent Semantic Analysis** and **TF-IDF** to power a content-based scientific paper recommendation engine. Scaled the vectorization and analysis by processing a corpus of **over 30 million documents** with **Spark** and loading embeddings into Elasticsearch.

INDEPENDENT PROJECTS

Time-Series Forecasting for Bakery Demand Optimization

- Designed and deployed an end-to-end forecasting system to predict daily bakery sales and optimize inventory, managing the full project lifecycle. Rigorously **compared multiple time-series models**, including **Prophet**, **ARIMA**, and **XGBoost**, to identify the most accurate solution.
- Engineered an automated daily data orchestration pipeline with Prefect to pull, clean, and process sales data, retrain models, and forecast the upcoming week, enabling the owner to reduce waste.

SCIENTIFIC PUBLICATIONS

- Acuna, D.E., **Yi, Z.**, Liang, L., Zhuang, H., Predicting the usage of scientific datasets based on the article, author, institution, and journal bibliometrics - [Mar. 2022]

EDUCATION

M.S. in Applied Data Science | Syracuse University Sept. 2020 - Dec. 2021
B.S. in Information Management & Technology | Syracuse University Sept. 2016 - May. 2020

CORE SKILLS

Applied Science & Machine Learning: Machine Learning, Deep Learning, NLP, Transfer Learning, Computer Vision, HB Modeling, Market Share Simulation, Recommendation System

Frameworks and Libraries: PyTorch, Scikit-learn, Pandas, NumPy, Git, Spark
Infrastructure & MLOps: Git, Docker, Kubernetes, Elasticsearch, Flask, Django Airflow, Prefect