Zhihan Li

408-427-1658 | 6nardoli@gmail.com | linkedin.com/in/zhihan-li-126429192 | https://github.com/leoliiiiii

Education

Carnegie Mellon University Pittsburgh, PA

Master of Information System Management GPA: 3.66 Expected 12/2023

Coursework: Cloud Computing, Distributed System, Large Language Models, Computer Vision, NoSQL Database Management.

University of California, Berkeley Berkeley, CA

B.A. in Applied Mathematics, Data Science GPA: 3.71 08/2017 – 12/2021

Coursework: Machine Learning, Natural language Processing, Data Structures, Data Mining, Database, Probability, Linear Algebra.

Skills

Programming Languages: Java, Python (PyTorch, Keras, Scikit-learn, Pandas, NumPy, SciPy, Flask, NLTK, Gensim), Scala, R.

Tools: AWS, Azure, GCP, Spark, Kubernetes, Docker, MySQL, Kafka, Samza, Redis, MongoDB, Cassandra, Git, Terraform, HTML.

Professional Experience

Machine Learning Engineer Intern | ScriptChain Health, Boston, MA 05/2023 – 08/2023

- **Data Processing:** Performed **ETL** using AWS GLUE on ~300GB Electronic Health Records.
- **ML Modeling:** Developed a deep learning model with **PyTorch**, consisted of **Graph Attention Network**, **Graph Pooling**, **Transformer**, and **Multi-instance Multi-label Classification**, to predict CVD readmission risk, achieving ~0.72 testing PR AUC.
- **Backend & Deployment:** Converted the trained model to **ONNX** format for better performance and greater flexibility, integrated with a **Django** web server, containerized using **docker**, and deployed on **AWS SageMaker**.

Data Scientist Intern | Pingan Technology, Shenzhen, China 05/2022 – 07/2022

- **ML Modeling:** Developed a NLP model (**Skipgram**) that generated vector representation for car damages to detect fraudulent car insurance claims, achieving an accuracy of 58% outperforming the rule-based detection mechanism in use (52%).
- **Model Optimization:** Optimized the model with **negative sampling** and new loss function, increasing training speed by ~200%.
- **Data Migration:** Developed a **flask server**, converting **PostgreSQL** tables into **Hive** tables, to automate the data migration process.

Business Analyst Intern | Avanade (Accenture BMW Project Team), Beijing, China 10/2020 – 01/2021

- Assisted in **project management** and reported monthly to the Accenture board, achieving project profitability increase by ~30%.
- **Product Management:** Participated in the entire lifecycle of developing two modules for BMW's IT platform, including surveying clients' demands, forming implementation plans, and performing system integration testing and user acceptance testing.

Projects

Twitter Analytics Web Service 03/2023 – 05/2023

- Developed a **web service** consisted of three **microservices** (QR Code Processor, blockchain validator, and Twitter Recommender System), achieving stable throughput of around 140k, 90k, and 16k RPS respectively with an AWS budget of $1.28 per hour.
- **Data Engineering:** Performed **ETL** on ~1TB raw twitter data using **Spark** based on optimized **schema design** for efficient queries.
- **Web & Storage Tier:** Implemented web servers using **Vert.x** and a **MySQL backend** with **AWS RDS** and configured connections.
- **Deployment & Automation:** Containerized the microservices with **Docker**, deployed in a **Kubernetes** cluster with an elastic load balancer, and automated the process with helm chart and terraform scripts.

Stream Processing for Ride-hailing Service 03/2023 – 04/2023

- **Streaming Layer:** Engineered and partitioned streaming pipelines using **Apache Kafka** on AWS EMR.
- **Processing Layer:** Leveraged **Apache Samza** for **real-time analysis**, pairing clients with drivers by preferences and geolocations.
- **Personalized Ad-targeting System:** Integrated static user data with live client streams to make effective ads recommendations.

NYC Fare Prediction Application 03/2023 – 04/2023

- **Model Development:** Developed a **XGBoost** model to make fare predictions deployed to a **Google Vertex AI endpoint**.
- **Web Server:** Created a Flask app deployed to **Google App Engine**, integrating the model with multiple Google Cloud APIs for named entity recognition, speech-text conversion, etc. to enable **real-time voice queries and prediction responses**.
- **Computer Vision:** Integrated a hybrid solution querying both a custom **Vertex AI AutoML model** and the Cloud Vision API to identify NYC restaurants and make corresponding fare predictions.