Zhihan Li

408-427-1658 | zhihanli@andrew.cmu.edu | linkedin.com/in/zhihan-li-126429192 | https://github.com/leoliiiiii

Education

Carnegie Mellon University Pittsburgh, PA

Master of Information System Management – Business Intelligence and Data Analytics Expected 12/2023

Coursework: Cloud Computing, Computer Vision, Unstructured Data Analytics, Applied Econometrics, Business analytics.

University of California, Berkeley Berkeley, CA

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

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

Skills

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

Tools & Technologies: Spark, AWS, Azure, GCP, Kubernetes, Helm, Docker, MySQL, Hbase, Neo4j, Kafka, Samza, Stata, Terraform.

Professional Experience

Machine Learning Engineer Intern | ScriptChain Health, Boston, MA 05/2023 - present

- **ML Modeling:** Developed a deep learning model using pytorch, consisted of Graph Attention Network, Graph Pooling (Differentiable Pooling), and Transformer, to predict cardiovascular disease readmission risk.
- **Data Processing:** Built a pipeline pre-processing EHR data and creating co-occurrence matrices for model training.
- **Model Tuning:** Performed pre-training on balanced dataset and fine-tuning on the entire dataset, leading to 0.72 testing accuracy.
- **Deployment:** Deployed the model onto AWS, introducing CVD readmission prediction service to the company's product portfolio.

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

- **ML Modeling:** Developed a deep learning 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:** Replaced the model's softmax layer with negative sampling algorithm and derived the corresponding loss function, leading to ~200% increase in training speed to save training cost.
- **Data Migration:** Developed a server using flask that converted PostgreSQL tables into Hive tables to automate and accelerate the data migration process for saving data storage cost.

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

- **Project Management:** Assisted in team management, including completing a three-year project plan, preparing monthly reports for Accenture's management board, etc., achieving cumulative profitability increase by ~30%.
- **Product Management:** Participated in the entire process 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.

Academic Projects and Independent Research

Cloud Computing: 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.
- **Web Tier:** Implemented the web tier using Vert.x after researching the performance of different web framework, such as Jooby.
- **Data Engineering:** Performed ETL on ~ 1TB raw twitter data using Spark based on optimized schema design for efficient queries.
- **Storage Tier:** Implemented MySQL backend with AWS RDS, loaded the data, and configured connection with the web tier.
- **Deployment & Automation:** Containerized the microservices with Docker, deployed the microservices in a Kubernetes cluster with an elastic load balancer, and automated the process with helm chart and terraform transcripts.

M&A Target Prediction (Independent Research advised by Professor Anastassia Fedyk @Haas School of Business) 09/2021 - 12/2021

- **Experiment Design**: Selected 690 target companies and control companies matching industry and firm size from S&P Capital IQ.
- Performed **hypothesis testing** to examine the correlations between a potential acquisition and various financial variables.
- Trained **topic models** on a total of 690 financial disclosures to obtain topic features using Latent Dirichlet Allocation.
- **ML:** Tuned logistic regression classifiers to predict acquisition targets, reaching state-of-the-art performance (~0.1 R-squared) with much fewer data and demonstrating the power of topic features in increasing model predictability over ~150%.