Shuowen Wei | Machine Learning Engineer

weisw9@gmail.com | (314) 215-8348 | LinkedIn: www.linkedin.com/in/shuowenwei

Summary

8+ years experienced ML engineer with proven success in building end-to-end ML solutions for different industries. Extensive hands-on experience in deep learning, NLP, scalable ML infra, risk modeling, anomaly detection with a solid quantitative background. Expertise in transforming business resources and requirements into manageable data formats and bringing excellent problem-solving skills to resolve complex issues efficiently and creatively.

- → Deep Learning, Machine Learning, NLP (LSTM, Transformers, BERT, XLNet etc), Graph Neural Networks (DGL).
- → Python, PyTorch, Keras/TensorFlow, Flask/Gunicorn/Nginx , Hadoop, Hive, Spark, Docker, Power BI, R, Matlab, SAS.
- → SQL, C#/.NET, Java/Spring Boot, Angular/TypeScript/JavaScript, HTML, CSS, Shell Script.
- → AWS EC2, S3, RDS, DMS, EMR, ECS/ECR, Lambda, ElastiCache/Redis, SageMaker, Cribl e.t.c.
- → Jenkins, JAMS, Bitbucket, MLOps, DevOps, Databricks, Domino Data Lab, Dataiku, e.t.c.

Experience

Lead Machine Learning Engineer – Finra, Rockville, MD .... 2018.02 - Present

- Strategic planning for leveraging AI/ML to make cyber security more proactive, efficient, and effective. Support Advanced Analytics strategy to ensure that models, data, tools, and infrastructures are secured.
- Led the VPC flow logs anomaly detection project, built scalable ML infra (Jenkins, EC2, S3, Redis, Cribl) for data ETL, feature engineering, and a large scale Autoencoders model to detect abnormal traffic from 34 external facing applications' VPC flow logs (10 million records, 50GB daily).
- Developed an AWS ECS/Docker pipeline for online ML model serving via RESTful APIs (Flask/Gunicorn/Nginx).
- Led the R&D Funding Portal (FP) Risk Monitor System project to help the FP review team to monitor risk of the FPs' webpage contents. Applied transferring learning using BERT to detect potential Finra rules violations from the FP landing pages' contents and analyze sentiment on investors' comments. Used ResNet50 for image embeddings and compared webpage screenshots similarities between different review cycles to detect content changes.
- Led the R&D Corpus of Evidence project to help investigators to find hard evidence among 1.22 million email records. Conducted Network Traffic Analysis (NTA) and Community Detection to reduce the total 635K emails to 198K. Used Transferring Learning (BERT) to encode each email's content to a 768-dimensional semantic vector, and then applied dimension reduction (Truncated SVD) and clustering (K-Means, DBSCAN) to cluster the 198K emails into ~90 clusters. Conducted Topic Modeling on each cluster using Latent Derilicht Analysis (LDA) and Non-negative Matrix factorization (NMF) and ranking each cluster of emails based on the number of security symbols detected using Entity Extraction.
- Built and deployed a documentation classification model (NLP) with an 81% accuracy using 1-D CNN to help the Advertising Regulation team to review securities related advertisements and communications submitted by Member Firms of Finra. It detects suspicious patterns among ~75k annual filings and automatically classifies those web advertisement documents to help analysts prioritize their work and mitigate the biases.
- Built an RNN model (LSTM/GRU) to classify Finra enforcement documents into 17 categories, achieving ~70% accuracy and using Python, PyTorch, NLTK, Gensim on AWS G3 instance with CUDA Toolkit 9.2 and cuDNN 9.2. Tackled highly imbalanced dataset with down-sampling. Tuned the model parameters using grid search and demoed the methodologies in front of the whole Member Regulation Technology Department (~80 audiences).
- Won the Grand Prize and the 2nd place in the group stage in the 2020 FINRA Createathon. Built a Graph Convolutional Neural Network (GCNN) model to help examiners to monitor the risks of Member Firm FinOPs using PyTorch and DGL.
- Won People's Choice Award and the 1st place in the group stage in the 2019 FINRA Createathon.
- Won People's Choice Award and the 2nd place in the group stage in the 2018 FINRA Createathon. Applied LSTM/GRU model to classify regulatory coordinators' (RC) emails into 5 risk categories, conducted grid search on 400+ different combinations of hyperparameters and achieved an accuracy of 74%. Presented the results to the Finra Management Committee including CEO and CTO.

Data Engineer Consultant at Finra – ConsultNet LLC, Rockville, MD ....2017.01 – 2018.02

- Built a deep learning POC model (LSTM/GRU) to classify ~9000 customer complaints into 10 categories with 91% accuracy, using both Keras/TensorFlow and PyTorch. Provided mentorship to a summer intern.
- Won 1st place in the group stage (among 17 teams) in the 2017 FINRA Createathon. Applied Random Forest and XGBoost e.t.c algorithms on broker-dealers' Financial and Operational Combined Uniform Single (FOCUS) Reports to re-evaluate their liquidity risk levels, achieved 60% accuracy overall and 75% accuracy in identifying high risk firms in liquidity. The results are validated by internal regulatory coordinators (RC) and received very positive feedback.

- Applied machine learning algorithms (regression, random forest) to build the High Risk Representatives (HRR) predictive model to detect suspicious behaviors among securities broker-dealers and investment advisors. Exacted and aggregated data from multiple databases for feature engineering using AWS EC2, EMR Cluster and S3.
- Built data ETL pipelines using JAMS and Jenkins for model development and ad-hoc analysis. Deploy and maintain models in the SDLC environment using both internal data management server and AWS services like EC2, S3, RDS, DMS.
- Led the Firm Address Matching project using Python/fuzzywuzzy for investigators to find out which broker-dealer (BD) firms were using virtual office (e.g., P.O Box) addresses to conduct their business. The results received very positive feedback from business users as they helped identify potential fraudulent behaviors by investment firms.

Sr. Predictive Modeler & Full-Stack Software Engineer – Health Integrity, Baltimore, MD 2014.06 – 2017.01

- Full-stack software engineer responsible for developing the company's main product PLATO platform (C#/.NET MVC framework, SQL, HTML, CSS, JavaScript). Handled monthly data ETL and ad-hoc data requests from clients.
- Major contributor of developing the query engine "CLEAR System to System (S2S)" application, for data mining, parsing, and aggregating on federated public records data sources APIs provided by Thomson Reuters Corp. Build complex data models and logic filters to target specific individuals and businesses and generate summary reports.
- Applied both supervised and unsupervised ML algorithms to build the Atypical Antipsychotics Prescriber (ATP) model and Trio Prescriber model to detect patterns of fraud, waste or abuse (FWA) in Medicare Part D data. Worked closely with subject-matter experts (SMEs) to understand medical record data, capture business requirements, build and validate models to enhance and extend the FWA detection processes.
- Automated and optimized model monthly run programs and delivered monthly high risk pharmacy/prescriber reports to the clients. Those reports had led to 71 investigations opened, 35 of which were referred to law enforcement agencies and 11 have been accepted for action, as of May 2016.

Education

M.S. in Computer Science, Wake Forest University, TA/full scholarship, NC, US	2013.08 - 2014.08
M.A. in Mathematics, Wake Forest University, RA/full scholarship, NC, US	2011.08 - 2013.08
B.S. in Applied Mathematics, Wuhan University, Hubei Province, China	2007.09 - 2011.06