Danying Xu

EDUCATION

New York University — **New York, United States**
Master of Science in Computer Engineering, Grade: 3.89/4.0 — 2023.09-2025.05

Southeast University — **Nanjing, China**
Bachelor of Engineering in Artificial Intelligence — 2019.09-2023.06

SKILLS

Professional Languages: **Python** ((NumPy/Pandas/Scikit-learn/Pytorch/Tensorflow/SciPy/SQLAlchemy/OpenCV/Matplotlib/ Seaborn/Beautiful Soup/PyGame/Django)**, C++/C, Linux** (Git), **SQL**, **Cypher, Java, Latex**
Tools: Google Colab, Azure CosmosDB, MySQL, PostgreSQL, Github, Gitee, Apache Spark, Apache Hadoop, Protégé, Neo4j
Methodologies: Machine Learning, Deep Learning, NLP, LLM, Database Management (SQL/NoSQL), Big Data, Statistics, Data Analysis, Predictive Modeling, Paper Reading &Writing

PROFESSIONAL EXPERIENCE

Global AI — **New York, United States**
Machine Learning Engineer — 2024.01-2024.04

- Established a **Postgres database** for 30k+ pieces of GDELT news data, boosting time efficiency by 20% on **DBeaver**.
- Implemented and visualized descriptive statistics on 1 million MSCI US Index stocks such as sparsity analysis and correlations.
- Cleaned the data using LOCF, NOCB and periodical average fillings.
- Developed **LSTM** improving stock forecasts by 19.7% with MSE score of 0.4 compared to the linear regression baseline.

Huawei Nanjing Research & Development Center — **Nanjing, China**
Software Development Engineer in Test, Data Communications Department — 2022.08-2022.09

- Conducted **Gray Box Testing** by examining 143 static path graphs with thousands of functions in **C/C++**.
- Conducted **White Box Testing** by **FUZZ test** technology for 872 code files by **Linux**, expected to improve product performance by 30%.

PROJECT EXPERIENCE

AI-Generated Text Detection — 2024.04-2024.05

- Finetuned the **BERT model** on 50k+ human-written and AI-generated data with the accuracy of 75.3%.
- Used the **LlaMa2 model** on **Google Colab** and **Hugging Face** with prediction accuracy of 65.4%.
- Deployed the **ChatGPT3.5 model** through **Azure OpenAI** and **Azure Notebook** with accuracy of 80.1%.

Text Gender Bias Rewriter (Research) — 2022.12-2023.06

- Proposed an NLP framework to reduce data gender bias via pattern transform, neural translation and data aggregation.
- Performed **Seq2Seq model** and **Seq2Seq attention model** (character/word level) on 148k+ Chinese sentences on **Pytorch**.
- Modified Word-Embedding Association Test to Chinese evaluated with **CBOW model**, reducing gender bias by 45.4%.
- Conducted **Coreference Resolution** downstream task on **wwm-RoBERTa Model** with consistent performance around 92% after reducing gender bias.

Recognition of Children's Autism — 2021.11-2023.05

- Extracted degree matrices and adjacent matrices of BOLD signals from 884 brain fMRI data.
- Built **GCN** and **GAT** models for determining autism, achieving the accuracy of 68.9% and 71.3%.
- Used the **Mixup** method for expanding the data size by three times, resulting in 3.4% accuracy improvement for GAT.

Deep Learning-based Explanatory Brain Science — 2020.11-2022.05

- Extracted 1.2 million images from 1297 videos of trained monkeys playing Pac-Man game using **Python (OpenCV)**.
- Developed a **ConvRNN model with AlexNet and LSTM** on **TensorFlow**, predicting players next move with 84.6% accuracy.
- Performed **Class Activation Map (CAM) heatmap** for activation layer visualizations.
- Modified the **Grad-CAM heatmap** for each layer to interpret the brain's decision-making mechanism with visualizations.

Knowledge-Based Question Answering (KB-QA) System — 2022.03-2022.04

- Crawled 1k+ La Liga game data from Wikipedia and FBref webs for fact extraction.
- Aligned the knowledge by regularization, Google translation API and Python (difflib).
- Built the ontology map on **Protégé** and knowledge database on **Neo4j**.
- Developed a **BiLSTM-CRF model** to segment natural language to generate queries in **Cypher**.

Implemented a dynamic web page with the locations of La Liga football clubs for visualization using Python (pyecharts).

Design of Prediction Model for NBA Games Analysis 2022.02-2022.03

2022 Winter GEARS in North Carolina State University

- Adopted the **Local Outlier Factor (LOF), Isolated Forest (IForest)** and **PCA** for data processing on 14,532 NBA data.
- Filtered the outliers using **K-Nearest Neighbor (KNN)** for 2,240 players from 1946 to 2004.
- Utilized **Bayesian Classifier**, **Logistic Regression**, and **LSTM**, to predict outstanding players with best of 96% F1 score.
- Applied **Grey Prediction**, **XGBoost**, and **MLP**, to predict game results with best of 0.25 MSE score.

Optimal Management Model of Forest Carbon Sequestration 2022.02

Finalist in 2022 American Mathematical Contest in Modeling, MCM/ICM

- Established an optimization model based on **Canadian Carbon Budget Model (CBM-CFS3).**
- Applied the model on 10k+ data of the White Mountain National Forest (WMNF) through **Lingo,** demonstrating a 25.3% improvement in deforestation rate.

Named Entity Recognition (NER) for Financial Data Extraction 2021.11

- Used 48k Groningen Meaning Bank (GMB) sentences to implement NER tasks in the field of NLP.
- Finetuned **the ELMo model** pretrained on 1b data with best 0.81 F1 score on 17 labels.

Time Series Data Analysis based on the Phytium-Kylin Systems 2021.07-2021.09

Grand Prize in the 17th "Challenge Cup" National Competition for Extracurricular Academic Science and Technology Works

- Designed a **time series anomaly detection model** combining **Spectual Residual and CNN**.
- Evaluated the model on 590k+ KPI data on **Azure**, achieving 0.73 F1 score and 81.1% accuracy.
- Developed the platform on **PK System** for Southeast University data center, forecasting a 24.7% potential increase in total benefits.

Data Anomaly Detection and Early Warning 2021.08

- Used algorithm A for robust statistical test to determine sensor impact risk on a total of 5k+ time series data.
- Applied Anderson-Darling test (AD), and Kolmogorov–Smirnov test (KS) to determine sensors with normal fluctuations.
- Performed the autocorrelation coefficient to determine sensors with strong risk fluctuations.
- Built a risk anomaly early warning model using linear regression, RBF kernel regression and polynomial regression for time series to achieve the overall optimal MSE score of 0.25.

Major Selection Recommendation System 2021.06-2021.07

- Crawled 11k+ enrollment data from 100+ university official websites.
- Build a **MySQL database** using Python (SQLAlchemy).
- Developed on **Django** framework and **Gitee** platform in implementing front-end and back-end interaction.

Predictions for Potential 5G Customers 2021.05

- Used the **K-Means** on 140k+ China Mobile users to classify 5G user behaviors, resulting in the silhouette score of 0.29.
- Improved the silhouette score to 0.37 using **Gaussian Mixture Model (GMM)** by **Python (Scikit-learn)** on the same task**.**
- Leveraged multiple regression tree **CART** for **Gradient Boosting Decision Tree (GBDT)** to build a 5G consumer behavior prediction model, achieving the best accuracy of 91.41% with grid search.

Examination Scoring System 2020.08

- Generated graphical user interfaces using **C++ (Qt)**, allowing users to log in and redirect to different interfaces.
- Implemented functions of a main system that supports sequential scoring and subsystems that supports direct modification and viewing for scores and questions, as well as exporting scores.