Ziying Song

(732) 285-2796 | zs2698@columbia.edu | availability: May - Dec 2025

EDUCATION

Columbia University *New York, NY*

M.S. in Data Science (Expected) *Dec* 2025

Courses: Algorithms, Machine Learning, Deep Learning, EDA & Visualization, Statistical Inference & Modeling, Time Series

University of North Carolina at Chapel Hill *Chapel Hill, NC*

B.S. in Statistics and Analytics Minors: Computer Science, Data Science *May 2024*

Cumulative GPA: 3.9/4.0 Honors: Highest Distinction, Dean's List

Courses: OOP & Data Structures, Database Systems, Machine Learning, Data Analysis, Statistics & Probability, Operation Research

PROFESSIONAL EXPERIENCES

Siemens *Beijing, China*

Research Scientist Intern *May-Jul 2024*

- Developed and optimized a knowledge graph backend in **Java**, improving data retrieval speed by 30% with **JSON-**based filtering.
- Developed **RESTful APIs** with **Spring Boot**, integrated **PostgreSQL**, and deployed to **Docker**, enhancing data storage and enabling real-time retrieval for ML models like GNN **and Transformer-based models** used for entity-linking and knowledge extraction.
- Built a **graph neural network (GNN)** pipeline to extract structured data relationships, boosting entity-linking accuracy by 20%.
- Analyzed datasets using **Python** and **Excel**, applying **feature engineering** and **unsupervised clustering** (K-Means) to detect trends in industrial data. **Led 3 interns** in data preprocessing.
- Improved data reliability by implementing **automated validation tests** and **ML-based anomaly detection (**Isolation Forest, Autoencoder**)**, increasing test coverage by 35%. Presented findings to engineers and product teams for data-informed decisions.

ByteDance *Beijing, China*

Data Scientist Intern *May-Aug 2023*

- Developed and deployed an end-to-end **user intent prediction system**, engineering Python-based data pipelines to automate ETL processes, reducing manual workload by **5+ hours per week**.
- Performed feature engineering and **exploratory data analysis (EDA)** using **Pandas, NumPy, and Scikit-learn**, extracting behavioral signals from user interactions, click-throughs, and sessions. Optimized **SQL queries** for real-time data retrieval and model training.
- **Trained and evaluated machine learning models** (e.g., logistic regression, decision trees, random forests) to enhance prediction accuracy by **20%**. Tuned hyperparameters and assessed performance using **AUC, precision, and recall**.
- Collaborated with product and engineering teams to refine model insights, driving data-informed enhancements in content ranking and recommendation strategies.

PROJECTS

Predicting Stock Price Movements | Columbia University *New York, NY*

- Engineered advanced feature sets from market microstructure data, incorporating volume metrics, price spreads, and temporal patterns to predict stock price movements, optimizing **LightGBM** to achieve the lowest MAE of 4.9874.
- Designed and evaluated models (LightGBM, Random Forest, **GRU, CNN-LSTM**) using **out-of-time validation** to ensure temporal consistency, leveraging feature standardization and rolling window calculations to enhance predictive accuracy.
- Conducted stock-specific analyses across 200 stocks, optimizing trading strategies through model-driven insights.

Breast Cancer Prediction | University of North Carolina at Chapel Hill *Chapel Hill, NC*

- Conducted Exploratory Data Analysis (EDA) on breast cancer malignancy data, utilizing **L1 logistic regression**, **MDS** (optimal with Canberra measure), and **PCA**. Employed a 3:1 dataset split for cross-validation.
- Implemented classification methods with **R**, including Logistic Regression, KNN, QDA, SVM, **Random Forest**, **XGBoost**, and **CNN**.
- Evaluated accuracy, AUC, and sensitivity across different EDA datasets. Achieved 99.14% accuracy and 98% sensitivity with KNN, with Area as the most influential predictor.

SKILLS

Languages: Python, R, SQL, JAVA, Julia, C, JavaScript, HTML, CSS, Matlab

Python Packages: Numpy, Pandas, Matplotlib, PyTorch, Sklearn, Tensorflow, Gensim, BERT, Spark, LangChain

Tools: MySQL, PostgreSQL, Docker, IntelliJ, Tableau, RStudio, MS Excel, MATLAB, AWS, Github, PowerBI, Spark, Scala, Stata