Jacob Zhu

New York, NY | +1 (929)626-0940 | yz9029@nyu.edu

EDUCATION

New York University **New York, NY**
Bachelor of Arts in Data Science and Mathematics Sept. 2022 – May 2026

- **GPA:** 3.6/4.0, Dean's List
- **Major Coursework:** Discrete Math, Numerical Analysis, Principles of Data Science, Causal Inference

PROFESSIONAL EXPERIENCE

China Construction Bank Jiangsu Branch **Nanjing, China**
Data Management Intern Dec. 2024 – Jan. 2025

- Contributed to big data risk control solutions, focusing on anomaly detection and fraud prevention algorithms (Isolation Forest, One-Class SVM) and real-time data processing (Apache Kafka, Flink)
- Participated in the 4th CCB AI Competition; used models like CatBoost, LightGBM, and AdaBoost Classifier to predict customer default risk based on 24,000 customer profiles and 6 months of credit card repayment data, achieving an AUC of 0.99 and an F1 score of 0.95 with the AdaBoost Classifier
- Conducted feature extraction and built a multi-class classification model to analyze fund investment preferences of approximately 500,000 potential clients. Optimized model performance with LightGBM using down-sampled data, achieving an F1 score of 0.64 and accuracy of 0.76

PROJECTS

2024 US Presidential Election Prediction with Machine Learning Aug. 2024 – Oct. 2024

- Analyzed historical election and polling data to predict the 2024 US Presidential Election outcome
- Explored various machine learning models such as Linear Regression, Random Forest, XGBoost, Gradient Boosting, and employed a Stacking Regressor for refined predictions
- Validated predictions against historical election results to evaluate accuracy

Spam Classification Using Machine Learning and Deep Learning Jul. 2024 – Aug. 2024

- Processed a dataset of spam and ham messages using tokenization, lemmatization, and stopword removal, and visualized word distributions through word clouds and bar charts
- Developed machine learning models, including Naive Bayes, Decision Tree, Logistic Regression, and Gradient Boosting, alongside a Bidirectional LSTM deep learning model for classification
- Assessed model performance using confusion matrices, classification reports, and accuracy scores

Diabetes Dataset Analysis and Regression Modeling May. 2024

- Conducted an in-depth exploration of the diabetes dataset, featuring EDA with correlation heatmaps, pair plots, and missing value visualizations
- Developed and evaluated regression models, including Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting, using RMSE, R^2, and cross-validation scores
- Examined the distribution of the target variable with the Shapiro-Wilk test to assess normality

Spotify Song Feature Analysis and Classification Mar. 2024 – Apr. 2024

- Analyzed the distribution and relationships between song features and popularity using histograms, scatter plots, and correlation analysis
- Developed regression, classification, and PCA-based models to predict song popularity
- Examined statistical differences in song popularity with tests like Mann-Whitney U and t-tests

SKILLS

Programming Languages: Java, Python
Libraries and Tools: NumPy, Matplotlib,SK learn, LaTeX
Languages: English (Fluent), Chinese (Native)