Kaiyan (Kaylee) Li

likaiyan02@gmail.com | LinkedIn | GitHub

SUMMARY

Data scientist with 4+ years of experience in machine learning, statistical modeling, and data-driven decision-making. Skilled in predictive analytics and delivering actionable insights to optimize business outcomes.

EDUCATION

University of Chicago – MS in Applied Data Science — Sep 2024 – Expected Mar 2026

- GPA: 4.00/4.00

University of North Carolina at Chapel Hill – BS in Mathematics & Statistics — Aug 2021 – May 2023

- Graduate with Highest Distinction and Honors; Phi Beta Kappa
- GPA: 3.89/4.00

SKILLS

- **Programming/Tools:** Python (pandas, NumPy, scikit-learn, statsmodels, XGBoost), R, SQL, MongoDB, Tableau
- **Machine Learning/AI:** Predictive modeling, statistical analysis, linear regression, logistic regression, decision trees, random forests, gradient boosting (XGBoost, LightGBM), support vector machines, k-means clustering, Gaussian mixture models, deep learning (CNNs, RNNs, transformers)

WORKING EXPERIENCE

Cmind Inc – Boston, MA — Jun 2023 – May 2024

Data Scientist Intern

- Trained and optimized boosting models (XGBoost, LightGBM) for earnings surprise prediction, applying hyperparameter tuning and trend-based feature engineering, improving accuracy from 65% to 70%
- Automated ETL pipelines using **Python** (pandas, SQLAlchemy) and **MySQL**, migrating 500K+ records from MongoDB to Oracle DB to enhance data consistency, improve query efficiency, and reduce data update time by 30%
- Independently performed feature importance analysis using the **Gain-Based method**, enhancing model interpretability and guiding financial analysts in decision-making
- Developed interactive dashboards (heatmaps, trend analysis) using **Python (Plotly, Matplotlib)**, increasing LinkedIn post engagement by 25% and providing stakeholders with improved data-driven insights for investment strategies

PROJECT EXPERIENCE

Robustness of PageRank Centrality on the Undirected Networks – Chapel Hill, NC — Aug 2022 – May 2023

Supervisor: Mariana Olvera-Cravioto

- Applied **random graph models** (Chung-Lu, stochastic block models) to real-world network data from SNAP, analyzing structural properties and optimizing parameters for simulations
- Evaluated PageRank centrality robustness using statistical metrics and **simulations** on synthetic networks, replicating real-world characteristics based on extracted parameters
- Identified key factors affecting rankings, demonstrating how variations in data distribution impact network performance and decision-making in connected systems

Fitness Center Customer Churn Analysis – Chicago, IL — Sep 2024 – Nov 2024

- Engineered a data pipeline to clean and process 25K+ customer records, ensuring data consistency for prediction
- Built and compared classification models (logistic regression, decision tree, random forest, CatBoost), with CatBoost achieving 92.6% accuracy, utilizing key predictors like payment type and facility usage frequency
- Presented insights to cross-functional teams, enabling data-driven customer retention strategies that reduced churn and stabilized revenue

Toxic Comment Classification – Chapel Hill, NC — Aug 2022 – Dec 2022

- Preprocessed Kaggle text data using normalization techniques to ensure high-quality inputs for model training
- Developed and fine-tuned deep learning models (Bi-LSTM, TextCNN, DistilBERT) using PyTorch, Keras, and TensorFlow, achieving 95% accuracy in toxic comment detection
- Communicated findings in reports and presentations, demonstrating DistilBERT's superior performance and its impact on improving content moderation and online user experience