Yiming Cao

Data Scientist/Machine Learning Engineer

800 Park Ave, Fort Lee, NJ 07024, US | +1 530-953-3833 | nickcao1124@gmail.com

Education

Columbia University, New York, NY, U.S. Sep 2025— Dec 2026

Master of Science in Environmental Health Data Science

University of California,Davis, Davis, CA, U.S. Sep 2021—June 2025

Bachelor of Science in Statistics

Core Course: Regression Analysis, Time Seiries Analysis, Multivariate Data Analysis, Data Science, Biostatistical Methods, Machine Learning for Public Health, Causal Inference, Deep Learning, Neural Networks & Transformer Models, Large Language Models & Fine-Tuning, Public Health GIS

Skills

Programming: Python (NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, PyTorch), R , SQL (MySQL, Hive), Apache Spark, Git,Tableau

Machine Learning & Data Science Techniques: Regression Analysis, Tree-based Models (Random Forest, XGBoost), Clustering (K-means, DBSCAN), Topic Modeling(LDA) ,A/B Testing, Dimensionality Reduction (PCA), Natural Language Processing , Deep Learning (Neural Networks, LLM fine-tuning, Agent development, RAG), Causal Inference (PSM,DiD), Model Evaluation

Work Experiences

Tencent, Beijing, China June 2025—Aug 2025

Tencent SSV, Tech Ecosystem Department,Data&AI Center: Data Secience Intern

- Built an **BI dashboard** for the "Digital Care platform"on team's Dataops warehouse, monitoring funds, beneficiaries, NGOs, and merchants.
  - Designed a **multi-dimensional indicator system** and iterated based on evolving business needs.
  - Identified and resolved critical business logic gaps, ensuring data integrity and actionable insights.
- Developed Gongfu AI marketing assistant (**LLM-powered Agent**) to empower rural communities with automated content generation.
  - Architected a **prompt-engineered workflow** that transforms keywords and short text into persuasive, localized marketing copy.
  - Successfully deployed on the rural support platform, reaching thousands of users and scaling event promotion capabilities.
- Automated **SFT fine-tuning** dataset preparation for the Gongfu AI agent.
  - Engineered a multi-threaded Python pipeline to generate large-scale Q&A pairs, reducing manual annotation costs by ~10x.
  - Delivered high-quality aligned training data, directly enhancing the model's robustness and generalization in marketing scenarios.
- Designed automated **LLM-based data cleaning pipeline** leveraging in-house DataOps framework.
  - Cleaned and validated 40k+ CRM education records(in natural language) with zero errors.
  - Increased efficiency from manual row-by-row cleaning to automated batch processing, achieving ~20x acceleration.
- Integrated MinerU **OCR** into the **RAG** ingestion flow for blurred, vertical historical docs, improving retrieval coverage and precision.

Projects

Natural Language Processing and Topic Modeling on User Review Dataset

- Grouped customer reviews into clusters and uncovered latent semantic structures using Python.
- Preprocessed text data through tokenization, stemming, stop-word removal, and feature extraction with Term Frequency–Inverse Document Frequency (TF-IDF).
- Applied unsupervised learning models including K-Means clustering and LDA.
- Identified latent topics and representative keywords for each review cluster to enhance interpretability.
- Visualized training results via dimensionality reduction using Principal Component Analysis (PCA).

Customer Churn Prediction in Telecommunications Industry

- Developed predictive algorithms for telecom service providers to estimate customer churn probability using Python and Apache Spark.
- Performed data preprocessing including cleaning, categorical feature transformation, and standardization to ensure model readiness.
- Trained supervised machine learning models such as Logistic Regression, Random Forest, and K-Nearest Neighbors, with regularization and hyperparameter tuning to mitigate overfitting.
- Evaluated model performance (Accuracy/F1-score) through k-fold cross-validation and conducted feature importance analysis to identify key drivers of customer churn.