Yiran Xu

929-474-1346 | yx2954@cumc.columbia.edu | https://github.com/Xuuuyeah |100 Haven Avenue, New York, NY

EDUCATION

Columbia University, Mailman School of Public Health *New York, NY*
M.S. Biostatistics | GPA: 4.0 *September 2024 – Present*
- Relevant Courses: Causal Inference, Deep Learning & Neural Network, Model & Trade Derivatives, SQL Programming

The Chinese University of Hong Kong, Shenzhen *Shenzhen, Guangdong, CN*
B.S. Bioinformatics| First Class Honor | mGPA: 3.5/4.00 (Top 20%) *September 2020 – May 2024*
- Relevant Courses: Machine Learning, Molecular Biology, Financial Management, Healthcare Financial Modeling with Excel
- Publication: https://doi.org/10.3389/fnut.2024.1366435
- Honors: 2024 Harmonia College Outstanding Graduate Award; 2022 University Excellent Student Leader Award

SKILLS

- Programming & ML/DL: Python, R, SQL, BigQuery, NumPy, TensorFlow, PyTorch, Model Training, Feature Engineering
- AI Development: AI Agents, LangGraph, RAG, Model Context Protocol (MCP), API, LLM, Prompt Engineering
- Statistics Analysis: Regression, Hypothesis Testing, A/B Testing, Causal Inference, Experiment Design
- Tools & Platforms: Power BI, Streamlit, Azure, Git, Snowflake, Databricks, AWS

PROFESSIONAL EXPERIENCE

Data Science Consultant Intern, Ambit Inc. *New York*, Jun 2025 – Aug 2025
- Led 0-to-1 design and development of multi-agent AI system using **LangGraph** with supervisor-worker architecture, implemented RAG pipelines over 10k+ internal documents and exploring fine-tuning strategies for domain adaptation with **Azure AI API**
- Engineered and optimized prompts to enhance AI reasoning; developed Model Context Protocol (MCP) servers and Python-based backend API services to enable LLM agents to access internal Snowflake data warehouses for automated analytics workflows with SQL**,** enabling end-to-end automation from data extraction, analysis, to Power BI data visualization with 95% completion rate
- Led end-to-end AI lifecycle through **POC** development with **LLMOps** best practice, including git version control, client-facing stakeholder engagement, requirements gathering, and technical demonstrations; delivered AI dashboard with Streamlit UI to CEO and packaged application using **PyInstaller** for company-wide deployment to automate departmental workflows by 25%
- Analyzed 1B+ claims data using SQL in **Spark-based** environment for competitor benchmarking, patient journey mapping, and KOL network identification, synthesizing insights to inform pricing and commercial strategy for pharmaceutical clients

Data Analyst Intern, Charles River Laboratories *Beijing*, Sep 2023 - Feb 2024
- Facilitated strategic transformation analysis from traditional mouse models to gene-edited models by assessing 42 competitors' business lines and core technologies from prospectuses, financial reports, and public sources
- Developed automated data pipeline with **CI/CD** principles using Python to scrape clinical trial data from public registries; evaluated investment trends and clinical trial pipelines, delivering data-driven insights on high-potential disease areas that contribute to actionable insights and drove strategic decisions for establishment of cardiovascular and metabolic R&D units
- Synthesized executive presentations and pitched differentiation strategy to CEO, projecting 10%+ market share capture

Data Analyst Intern, Shenzhen Ruiping Technology Co., Ltd – Startup Experience *Shenzhen*, Jun 2023 - Aug 2023
- Managed full product lifecycle of a new medical device by cross-functionally coordinating with marketing and R&D teams, driving the process from prototyping to market launch generating 100,000+ RMB in annual revenue
- Conducted market trends, competitor offerings evaluations, and quantitative analysis on 2,0000+ pieces of customer feedback using **K-means** clustering, supported new product formula with 14% improvement in efficacy and a 70% decrease in skin irritation
- Evaluated psoriasis-targeting medical devices using internal large language model (**LLM**) for literature review, clinical treatment benchmarking, and ecommerce product analysis, identifying market gaps and cost-effective alternatives

Part-time Assistant, McKinsey & Co *Remote,* Jan 2023 - Mar 2023
- Formulated **market entry** strategy for a top private bank in China's ultra-high-net-worth segment, applied SCP-I model to identify service gaps and craft client personas, supporting differentiated positioning projected to increase new client acquisition by 10-15%
- Supported client segmentation with Censydiam model and competitor analysis for digital strategy for a loan initiative targeting blue-collar workers, enabling the client to tap into an underserved segment and project 5–10% loan portfolio growth

PROJECT EXPERIENCES

AI/ML Lead, Rare Disease Early Diagnosis with AI Algorithm Jun 2025 – Present
- Built **ETL pipeline** on Databricks with Python and SQL for 1B+ **claims data**, performing data cleaning, data validation, and data transformation; conducted statistical analysis to identify trends and patterns, defined primary and secondary evaluation metrics
- Developed Transformer deep learning model with multi-head attention and embeddings using **Python** and **TensorFlow** for rare disease risk prediction in multiple time window from 200K+ time-series patient journeys; performed data preprocessing, feature engineering, hyperparameter tuning, and GPU-accelerated model training; achieved 67% AUPRC in model evaluation on imbalanced data using Focal Loss; implemented Integrated Gradients for model interpretability

Team Leader, Breast Cancer Survival Analysis with R Sep 2024 - Dec 2025
- Conducted EDA, data wrangling, and data cleaning on 2M+ health records, applied statistical analysis, machine learning models (Regression with L1/L2 regularization, SVM, MARS) and hypothesis testing (OLS, LDA, PCA, t-test) in **R** to multiple datasets
- Developed interactive tables, figures, and visualizations with **R Shiny** for reporting; documented code and maintained organized records via **GitHub** for team collaboration and version control

Silver Medal Winner, Kaggle Competition (Parkinson's Freezing of Gait Prediction) Apr 2023 - Jun 2023
- Predicted Freezing of Gait events of Parkinson's patients from 70GB+ **time-series** data by constructing a 1D Convolutional Neural Network with packages like **PyTorch, NumPy, and Pandas**, reaching a final mean average precision of 0.3306 (ranked top 2%)