PEI YI (CINDY) LI

+1 (929) 737-5221 | pl2923@columbia.edu | www.linkedin.com/in/pei-yii | New York

EDUCATION

Columbia University New York, NY
M.S. in Industrial Engineering Aug 2024 – Expected Dec 2025

- Courses: Data Analytics, Optimization Models, Data Mining, Cloud Analytics, Quantitative Pricing

University of British Columbia Vancouver, BC
B.S. in Computer Science and Biology **GPA: 3.8/4.0.** Sept 2018 – Dec 2022

- Courses: Advanced Relational Databases, Introduction to Artificial Intelligence, Machine Learning, Genetics.

SKILLS

Languages: Python (Numpy, Pandas, Matplotlib, Sympy, scikit-learn), SQL, Java, R, VBA, Matlab, Excel
Others: MySQL, MongoDB, PostgreSQL, PowerBI, Tableau, AWS, API usage, PyTorch, Epic (EHR system), GCP, Apache Spark

WORK EXPERIENCE

Memorial Sloan Kettering Cancer Center New York, NY
Data Science Intern Mar 2025 - Present

- Built AI-powered 3D medical image segmentation models using deep learning architectures like nnU-Net, enabling automated identification of clinically relevant anatomical structures from imaging data.
- Designed scalable preprocessing and inference pipelines to support robust performance across 5+ imaging formats and 1,000+ annotated scans, ensuring consistent model accuracy across diverse clinical datasets and real-world variability.
- Integrated backend ML models with a custom software extension, designed and developed the UI, enabling real-time, physician-in-the-loop segmentation via REST API and supporting future deployment of the tool.

Anote *(AI Team)* New York, NY
Machine Learning Engineer (Part-time) May 2025 - Aug 2025

- Led a team of 3 in the development of an agentic LLM-powered chatbot to streamline prior authorization workflows in hospital and clinic infusion centers, reducing manual effort in billing by retrieving insurance coverage and drug details.
- Deployed an end-to-end pipeline integrating real-time web scraping using Selenium, structured formulary databases, and Pinecone-based vector search to retrieve eligibility based on chat input.
- Built and refined a conversational interface that enables clinical staff to interact with backend models seamlessly, delivering real-time coverage insights through a secure and scalable AI assistant.

Regeneron Pharmaceuticals New York, NY
Data Science Student Intern (Industry Collaboration) Jan 2025 - May 2025

- Cleaned and standardized complex clinical trial data from the AACT database, including filtering 45,000+ studies, handling inconsistent entries, mapping medical conditions, and engineering features across study design, geography, and therapeutic area.
- Developed two predictive ML models to estimate clinical trial duration and predict early trial termination, and engineered feature pipelines and identified key duration drivers, achieving an F1-score of 81% for terminated trial classification.
- Deployed an interactive Gradio dashboard enabling users to input trial design parameters and receive insights from model prediction on expected duration and termination risk to enable feasibility assessment and scenario testing.

YF Technology Vancouver, BC
Data Analyst Aug 2022 - Aug 2023

- Delivered actionable reports using Tableau dashboards by monitoring key performance indicators in TikTok campaign to maintain client relationships and provide actionable insights to redesign influencer campaign to align with target audience.
- Created targeted advertising copy based on Click Through Rate (CTR) models using Python, collaborated closely with marketing team to deliver and execute model, resulted in a notable 15% increase in average CTR for campaigns.
- Implemented a predictive model using machine learning to analyze customer engagement on TikTok ads, partnering with the marketing team to refine audience segmentation and targeting strategies, increasing ad targeting accuracy by 22%.

PROJECT EXPERIENCE

Healthcare Insurance Plan Recommender (Python, Next.js, Groq AI) - Devfest 2025 Award Winning

- Developed a full-stack ML-powered insurance plan recommendation system, mapping demographic inputs to optimized health plans using CatBoost, achieved best model performance: $R^2 > 0.85$.
- Engineered a chatbot interface leveraging Groq and Retrieval-Augmented Generation (RAG) to answer real-time insurance queries with structured and unstructured data retrieval.
- Designed and deployed the backend architecture for fast plan scoring, clustering, and output formatting; collaborated closely with frontend team (Next.js) to ensure seamless UX.