Toby Liu

Data Scientist | San Francisco, CA | tobyliu2004@gmail.com | Github | LinkedIn | Portfolio

EDUCATION

UC Santa Barbara, B.S. in Statistics & Data Science, B.A in Economics, Letters & Sciences College Honors Program, Expected June 2026

Relevant Coursework: Statistical Machine Learning(Graduate Level), Data Science Principles, Probability & Statistics, Linear Algebra, Design of Experiments, Econometrics, Vector Calculus, Regression Analysis

TECHNICAL SKILLS

Programming Languages: Python, SQL, R
Libraries & Frameworks: Pandas, NumPy, Scikit-learn, XGBoost, TensorFlow, PyTorch, Matplotlib, tidyverse
Tools & Platforms: Git, VSCode, Google Cloud Platform (GCP), AWS, Airflow, Jupyter Notebook, Tableau
Databases & Querying: PostgreSQL, MySQL, BigQuery
Machine Learning & Statistical Modeling: Clustering, Time Series, A/B Testing, Causal Inference, NLP

EXPERIENCE

Houston Methodist (Medical Artificial Intelligence & Innovation Lab) Houston, TX, **Machine Learning Engineer Intern**, May 2025 – August 2025

- Built Random Forest and XGBoost models to predict prostate cancer treatment responses from gene expressions data (TCGA-PRAD), achieving **70% AUC** on test sets, enabling faster patient stratification
- **Reduced input feature space by over 99%** using statistical testing (Fisher's exact test) and feature selection, improving model interpretability and training time, while eliminating data leakage
- Deployed and evaluated a deep learning classifier using TensorFlow using kidney cancer data (TCGA-KIRC), refining architecture and adding early stopping to improve generalization
- Created a modular ML pipeline to apply both workflows across other TCGA cancer types, **cutting modeling time by 50%** and enabling rapid experimentation

Boston Children's (Computational Biology Department), Boston, MA, **Data Science Intern**, June 2024 – August 2024

- Analyzed **50,000+** genomic sequences using DANPOS (Python) to evaluate environmental impacts on nucleosome dynamics and protein-DNA occupancy in collaboration with the department head
- Built exploratory data analysis pipelines and statistical workflows that **improved pipeline efficiency by 8%**
- Created interactive R visualizations that highlighted statistical deviations and biological trends in sequencing data for research team

PROJECTS

ML Model Audit Copilot |Python, SQL, SHAP| March 2025 - April 2025

- Built a modular ML audit framework to catch drift, data leakage, fairness issues, and schema mismatches across model pipelines
- **Automated 7+ audit workflows** with CLI + SQL support, enabling side-by-side model comparisons and batch reporting for version tracking
- Integrated SHAP explainability and demographic bias breakdowns (e.g., group-wise MAE), **cutting audit time by 60%** and streamlining model validation

Hospital Cost Prediction Pipeline |Python, XGBoost, SHAP, Streamlit | February 2025 - March 2025

- Developed an XGBoost model to predict inpatient procedure costs from SPARCS data, **achieving 91% R^2** with stratified regression and Optuna tuning to estimate impatient costs across New York Hospitals
- Added fairness audits across race and gender, drift simulations and unsupervised anomaly detection (Isolation Forest), **cutting manual QA time by 50%**