# Matthew (Matt) Ohanian

matt.ohan@gmail.com · linkedin.com/in/matthewohanian · github.com/mattohan567 · mattohanian.com · +1 (818) 522-4089

Data Scientist and Machine Learning Engineer with a track record of shipping production ML systems and Generative AI solutions. UC Berkeley AI Hackathon winner (Anthropic Track).

## Experience

**Full-Stack AI Engineer**
*California University of Science and Medicine* — *January 2026 - September 2026*
- Built source-grounded healthcare RAG backend and privacy-first cross-platform mobile app for MS caregiver support in collaboration with physicians and medical students at California University of Science and Medicine
- Designed 5-stage RAG pipeline (guard rails, HyDE, hybrid vector+BM25 retrieval, FlashRank reranking, grounded generation) passing a 90-example eval with 1.000 Recall@5, 0.851 MRR, 0.896 faithfulness, and 1.000 no-answer success
- Engineered crisis detection, prompt injection defense, and scope detection for healthcare-sensitive queries with self-auditing citation enforcement

**Machine Learning Engineer & Consultant** — **Remote**
*EcoRecycle* — *January 2026 - July 2026*
- Trained and evaluated regression-based vehicle pricing models for an end-of-life vehicle auction platform used by automotive recyclers
- Recommended MLOps improvements including recurring retraining, validation-set comparison, drift monitoring, model promotion gates, rollback-aware deployment, and structured feedback capture

**Data Scientist** — **San Francisco, CA**
*Qventus* — *February 2025 - June 2025*
- Developed surgical case prediction models using ensemble methods (XGBoost, LightGBM) with cross-validation on healthcare snapshot data, optimizing OR scheduling for hospital partners
- Built real-time inference MLOps pipelines on AWS SageMaker with model deployment, latency optimization (sub-50ms), and automated monitoring, processing 100K+ daily predictions while tracking data drift and model performance
- Engineered 50+ predictive features from healthcare data using Snowflake SQL with automated feature validation and data quality checks
- Applied Transformers (BERT via Hugging Face) to surgical procedure text to compute semantic similarity and cluster cases, helping clinicians identify related procedures and streamline case review

**Data Scientist** — **San Francisco, CA**
*Raydiant* — *August 2024 - February 2025*
- Architected ETL pipelines using AWS Glue, S3, Lambda, and Athena to process 3M+ customer records, reducing pipeline errors by 45%
- Built churn prediction model achieving 85% AUC-ROC with SHAP analysis, enabling targeted retention campaigns
- Created real-time Salesforce dashboards for leadership team, driving data-driven decision making

**Data Scientist** — **Cape Town, South Africa (Remote)**
*Axaitech* — *May 2022 - August 2022*
- Built 11-class cancer classification model using deep learning, neural networks, and ensemble methods, achieving 95% accuracy while reducing false positives by 40% vs. baseline
- Applied Random Forest feature selection to reduce gene expressions from 58,000 to 16 critical biomarkers without accuracy loss
- Presented findings to clinical stakeholders, informing diagnostic workflow improvements

## Projects

**InDaCapitol - Congressional Stock Trade Predictor** *(Python, LightGBM, Supabase, Next.js, React, FastAPI)* — Production App
- Built production full-stack ML system with daily ETL across 6+ sources, parsing 14K+ disclosures into 16.7K feature vectors and 16.6K predictions
- Trained June 2026 segmented LightGBM buy/sell models: classifiers at 0.695/0.697 AUC and regressors at 0.0813/0.0774 MAE
- Shipped Next.js/Vercel dashboard with prediction badges, 99.9%/99.6% feature/prediction coverage, 73.3% realized accuracy, and stable drift

**JobLens - Privacy-First Job Email Classifier** *(Transformers.js, ONNX, WebAssembly)* — Chrome Extension
- Published Chrome Web Store extension with on-device inference using Transformers.js + ONNX/WebAssembly, keeping email classification fully in-browser
- INT8 quantized for browser-edge deployment, reducing model size from 128MB to 34MB (73% reduction) while preserving accuracy
- Fine-tuned BAAI/bge-small-en-v1.5 with custom classification head on 28,500 emails (20.5k real + 8k synthetic via DeepSeek/Ollama) with curriculum learning across data formats

## Education

**University of San Francisco** — **San Francisco, CA**
*Master of Science in Data Science, GPA: 3.9/4.0* — *July 2025*

**University of California, Los Angeles** — **Los Angeles, CA**
*Bachelor of Science in Data Theory, GPA: 3.8/4.0* — *June 2023*

## Technical Skills

**Languages & Frameworks**: Python, SQL, R, PyTorch, TensorFlow, scikit-learn, Transformers, Hugging Face, FastAPI
**Machine Learning**: XGBoost, LightGBM, Random Forest, Deep Learning, Neural Networks, NLP, Computer Vision, Feature Engineering
**Cloud & MLOps**: AWS (SageMaker, Glue, Lambda, S3, Athena), Snowflake, Docker, Kubernetes, MLflow, W&B, CI/CD
**Model Deployment**: ONNX, Model Quantization, Real-time Inference, Model Serving, REST API, Model Monitoring, A/B Testing
**Generative AI**: LLM APIs (Claude, GPT-5, Gemini, Mistral, DeepSeek, Grok, Ollama), LangChain, CrewAI, Multi-Agent Systems, Embeddings, Synthetic Data
**Data Engineering**: PostgreSQL, Supabase, PySpark, Redis, Celery, ETL Pipelines, Salesforce, Pydantic, Plotly, Tableau, Git