

SUNIL ALETI

thesunilaleti@gmail.com | (832) 292-4956 | sunilaleti.dev | linkedin.com/in/sunilaleti/ | Blog

SUMMARY

Machine Learning Engineer / Data Scientist with 5+ years shipping ML and Generative AI systems across financial services, retail, and technology. Specialized in MLOps, RAG, agentic AI, and LLM fine-tuning (LoRA/QLoRA, PEFT, SFT), plus classical ML and statistical modeling. Turns business needs into production models that cut risk, automate workflows, and drive measurable impact.

SKILLS

Machine Learning & Analytics: Predictive Modeling, Statistical Analysis, Data Analysis, Classification, Regression, Feature Engineering, Hyperparameter Tuning, A/B Testing, Clustering, PCA, Random Forest, XGBoost, CatBoost, SVM, Model Evaluation (AUC, KS, Gini, RMSE)

Generative AI & LLMs: LLM Fine-Tuning (LoRA, QLoRA, PEFT, SFT, Instruction Tuning), RAG, Agentic AI, MCP, Claude SDK, LangChain, Hugging Face Transformers, Embeddings, Vector Databases

Deep Learning: PyTorch, Neural Networks, Transformers

Programming & Data: Python (Pandas, NumPy, Scikit-learn, Statsmodels, FastAPI, Flask), SQL, SQL Server, Cosmos DB, MongoDB

Cloud & MLOps: Azure (Functions, ML), AWS (SageMaker, S3), Docker, Git, MLflow, CI/CD, Model Drift Detection, Model Monitoring & Observability.

Visualization & Domain: Power BI, Streamlit, Excel; Credit Risk Modeling, Fraud Detection, Model Risk Management (MRM)

PROFESSIONAL EXPERIENCE

7-Eleven | AI / ML Developer

Oct 2025 – Present

- **Designed and deployed a conversational-commerce chatbot** embedded in the 7-Eleven app, letting customers place orders, update item quantities, manage their cart, and ask product questions with retrieval-augmented (RAG) answers built with LangChain-orchestrated AI agents and tool-calling over the ordering and catalog APIs.
- **Built and evaluated agentic AI systems** using MCP and Claude to orchestrate multi-step reasoning, tool invocation, and autonomous task execution across ordering and operations workflows.
- Built and automated AWS S3 data pipelines to ingest and stage data feeding the AI services.
- Established MLOps and LLMOps for the GenAI stack, including containerized deployment, prompt and model versioning, automated evaluation, and production monitoring for reliable, scalable performance.

Tricolor Holdings | Senior Data Scientist

Jan 2024 – Sep 2025

- **Designed and deployed a Probability of Default (PD) model** with advanced feature engineering and calibration, improving risk-prediction accuracy to 78%.
- **Reduced loan delinquency by 32%** with models that proactively identify first-payment defaulters.
- **Automated core credit-risk processes with Azure Functions**, driving a 95% gain in operational efficiency while integrating model outputs into regulatory-compliant credit frameworks.

Tricolor Holdings | Data Scientist

Jun 2022 – Dec 2023

- **Built a Retrieval-Augmented Generation (RAG) chatbot** with LangChain over SQL Server and Excel, giving executives instant self-service access to portfolio data.
- **Engineered "Crossfire," a real-time fraud-detection platform** (FastAPI + fuzzy matching) to flag repeat customers and straw purchasers, improving fraud-detection accuracy by 58%.
- **Designed an NPV model for loan-portfolio valuation** supporting successful securitization pitches; lifted campaign performance 84% via logistic regression, and A/B testing, with model drift monitored via PSI under MRM standards.

Cognizant Technology Solutions | Programmer Analyst / Data Scientist

Jan 2019 – Dec 2020

- **Forecasted new-product sales at 79% accuracy** with ML and data-mining techniques (deployed via Flask and Amazon SageMaker), and built a supply-chain optimization model to forecast demand and optimize warehouse placement, cutting transportation and storage costs.

EDUCATION

M.S., Computer Science | University of Texas at Arlington, TX | GPA 3.9

B.E., Computer Science & Engineering | Anna University, India

PROJECTS & CERTIFICATIONS

- **Claude Certified Architect - Foundations (CCAF):** Anthropic certification in designing and building agentic, LLM-powered applications.
- **Dictify (open-source macOS app):** real-time voice-to-text with LLM text refinement, powered by the Groq API (Whisper for speech-to-text, gpt-oss for refinement) for low-latency inference.
- **datanerd & cmdline-utils** (PyPI): published Python libraries streamlining data-analysis and CLI developer workflows.
- **Technical Blogging:** 250K+ views on DEV.to and Hashnode covering data science, Python, and automation.