

# SHRIKANT SHARMA

774-438-9698 | [shrikant.sharma@myyahoo.com](mailto:shrikant.sharma@myyahoo.com) | [shrikant-sharma.github.io](https://shrikant-sharma.github.io) | Berlin, NH  
[linkedin.com/in/shrikant-sharma](https://linkedin.com/in/shrikant-sharma) | [github.com/Shrikant-Sharma](https://github.com/Shrikant-Sharma) | [public.tableau.com/app/profile/shrikant.sharma](https://public.tableau.com/app/profile/shrikant.sharma)

**Data Scientist | Pharma & Financial Services | Python, ML, NLP, GenAI, Agentic AI**

Data Scientist with ~8 years across pharma and financial services, building production ML and applied GenAI in regulated, high-liability environments where models must be interpretable and audit-defensible. Recent focus on agentic RAG with evaluation and guardrails, causal inference and experimentation, and anomaly detection at scale, demonstrated across 3 shipped, end-to-end portfolio projects.

## SKILLS

**Machine Learning & AI:** Generative AI, LLMs, Agentic AI, RAG, Causal Inference, Experiment Design, Survival Analysis, Anomaly Detection, NLP, Deep Learning, Classification, Regression, Statistical Modeling, Feature Engineering

**Languages & Libraries:** Python (Pandas, NumPy, Scikit-learn, XGBoost, PyTorch, Sentence Transformers, FAISS, SHAP, RAGAS, LangGraph, DoWhy, lifelines, PySpark, Matplotlib), SQL, R

**Cloud & Data Platforms:** AWS (S3, EC2, Lambda, ECR), Databricks, Snowflake

**Tools & Platforms:** Tableau, Power BI, Apache Spark, Git, Docker, FastAPI, Streamlit, MLflow

## EXPERIENCE

### Amgen | Data Scientist

Sept 2022 - May 2026

- Architected the Python ETL pipeline behind Amgen's CMS Open Payments submission, processing ~509K records (~\$630M) on AWS and Databricks with automated SQL validation that held the failure rate under 0.5%. Designed year-over-year Tableau dashboards and led process improvements that cut processing time from 2 weeks to 4 days over 4 annual submission cycles (2022 to 2026).
- Built a multivariate risk scoring system in Python (Pandas, NumPy, Scikit-learn) that ranked 65+ countries across 20+ compliance indicators using z-score normalization, with US scoring tailored at the product level. Rebalanced feature weights from 50/50 to 75/25 in favor of quantitative data and expanded the output from 3 risk tiers to 5, which surfaced 3 high-risk markets that had gone undetected. Presented findings to Directors and Executive Directors, informing resource allocation across 3 global regions.
- Developed a statistical flagging system in Python for compliance transactional monitoring across 5 payment processes (Speaker Programs, Advisory Boards, Donations, Fee-for-Service, Sponsorships), applying threshold-based detection on 50K+ records across the US, ELMAC, and JAPAC. Reduced manual review volume by 35%.
- Led an initiative to apply NLP to compliance document processing. Delivered a working text classification prototype using TF-IDF and Random Forest (Python, Scikit-learn) that cut manual categorization work by 25%. Also contributed to Amgen's enterprise AI direction through 2 internal GenAI hackathons.
- Delivered 3 concurrent projects on schedule while simultaneously owning end-to-end knowledge transfer of analytics workflows to incoming team members during organizational restructuring; mentored 3 junior data scientists on Python automation and Tableau development.

### Nugget.ai | Data Scientist Intern

Jan 2022 - May 2022

- Built models in Python (Scikit-learn) to predict candidate-role fit from behavioral assessment data, engineering 20+ features spanning structural, linguistic, behavioral, and content-based signals. Compared Logistic Regression, Random Forest, and Gradient Boosting with cross-validation, improving matching accuracy by 15% over the platform's existing baseline.

### Atos Syntel | Data Scientist | Pune, India | Client: American Express

Feb 2018 - Dec 2019

- Built risk scoring pipelines in Python and SQL, processing 10M+ transactions, engineering features around spending velocity, geographic patterns, and behavioral segmentation that fed downstream fraud and credit risk team reviews.
- Designed Tableau dashboards consolidating 5+ source systems into a single executive view used in the weekly credit risk review, supporting mitigation decisions totaling over \$2M and cutting reporting time by 30%.
- Implemented the team's A/B testing framework in Python, SQL, and Git for credit risk strategy experiments, handling experiment design, power-based sample sizing, and significance testing with multiple-testing correction across 12+ cycles per year.

### Atos Syntel | Data Analyst | Chennai, India | Client: American Express

Mar 2016 - Jan 2018

- Established the SQL workflows that processed 5M+ financial transaction records every month, with validation logic and quality checks that held accuracy at 99.5% for regulatory reporting.
- Automated 15+ recurring reports using SQL and Python, freeing up 30+ hours of manual work each month so the team could focus on analysis instead of data prep.

## PROJECTS

**Clinical Trial Intelligence System** | Python, LangGraph, Sentence Transformers, PubMedBERT, Cross-Encoder, Groq Llama 3.3 70B, FAISS, RAGAS, Streamlit, Git **May 2026**  
[github.com/Shrikant-Sharma/clinical-trial-rag](https://github.com/Shrikant-Sharma/clinical-trial-rag) | [shrikant-clinical-rag.streamlit.app](https://shrikant-clinical-rag.streamlit.app)

- Built a clinical-trial search assistant over 484 ClinicalTrials.gov protocols spanning oncology and cardio-metabolic trials (3,264 chunks), so a research analyst could surface relevant studies through natural-language queries instead of manual abstract scanning.
- Compared 3 chunking strategies (500/1000/1500 chars) using RAGAS faithfulness and LLM-judged context precision, and separately validated system refusal behavior across a 25-query, six-category stress test.
- Architected agentic Corrective RAG (CRAG) upgrade with LangGraph adding LLM-as-judge document grading, query rewriting, and bounded retries. 3 independent refusal gates correctly refused adversarial queries scoring 0.91 cosine similarity that the baseline threshold missed.
- Added a cross-encoder reranker (MS-MARCO MiniLM-L-6-v2) for 2-stage retrieval before deduplication; halved in-corpus refusal rate (67% to 33%) on a 6-query LLM-judged eval by promoting previously missed trastuzumab-deruxtecan (T-DXd) chunks above generic trastuzumab studies on HER2-positive antibody-drug conjugate queries. Deployed live on Streamlit Community Cloud with graceful rate-limit handling.

**Patient Risk Stratification Pipeline** | Python (Scikit-learn, XGBoost, PyTorch, SHAP, DoWhy, lifelines, scikit-survival), MLflow, FastAPI, Docker, AWS (Lambda, ECR, Mangum), Git **May 2026**  
[github.com/Shrikant-Sharma/patient-risk-stratification](https://github.com/Shrikant-Sharma/patient-risk-stratification) | [Live API on AWS Lambda](#)

- Built an end-to-end ML pipeline for clinical heart disease risk stratification on UCI Heart Disease (303 patients, 13 clinical features); selected XGBoost from 5 model families via MLflow experiment tracking and 5-fold cross-validation, reaching 0.95 AUC-ROC and 0.93 recall (precision 0.81) at the clinical operating threshold, identifying 93 of 100 high-risk patients for early review.
- Added SHAP TreeExplainer for global and per-patient feature attributions critical for clinical decision support, validated probability calibration at Brier 0.092 with reliability curves, and tuned the decision threshold to prioritize recall while maintaining precision at or above 0.80.
- Extended the pipeline to causal inference and survival analysis on the NHEFS cohort (1,629 subjects); estimated treatment effects via propensity score matching and G-computation with DoWhy (3-way convergence within 0.2 kg of the IPW reference, 3 sensitivity tests passing) and modeled 10-year mortality with Cox PH and Random Survival Forest at 0.80 concordance, revealing that unadjusted log-rank significance ( $p=0.005$ ) was pure age confounding.
- Deployed live as a public REST API on AWS Lambda via Amazon ECR and Mangum (ASGI-to-Lambda adapter); multi-stage Docker build, EventBridge cold-start warmer holding warm latency at ~275ms, and FastAPI Swagger UI for interactive testing.

**Pharma Compliance Spend Analytics** | Python (Pandas, Scikit-learn, PySpark), Snowflake, Tableau, Git **May 2026**  
[github.com/Shrikant-Sharma/pharma-compliance-spend-analytics](https://github.com/Shrikant-Sharma/pharma-compliance-spend-analytics) | [public.tableau.com/app/profile/shrikant.sharma](https://public.tableau.com/app/profile/shrikant.sharma)

- Built a compliance spend analytics system on real CMS Open Payments data (16M+ records, \$13B) for federal payment-transparency review, engineering 5 HCP-level features across 989K sampled transactions and 289K providers.
- Applied within-specialty z-scores plus global IQR with a \$500 monetary floor and concentration logic; flagged the top 1.67% of providers as a HIGH-tier compliance triage queue.
- Layered Isolation Forest ML anomaly detection on top; reconciliation against rule-based flags surfaced 3 compliance archetypes: captured specialists (BOTH), captured generalists (rules-only), and industry consultants (ML-only), revealing patterns no single method catches alone.
- Implemented the same feature pipeline across 3 execution backends (Pandas, Snowflake, PySpark) with row-level equivalence verified; Snowflake SQL quantified that the top 4 device manufacturers (Arthrex, Stryker, Zimmer Biomet, Smith+Nephew) capture 59.7% of orthopedic surgery payments, visualized in 2 interactive Tableau Public dashboards.

## EDUCATION

**Clark University**, Worcester, MA | **Master of Science in Data Analytics**, GPA: 3.6/4.0 **May 2022**  
**Sant Gadge Baba Amravati University**, India | **B.E. in Electronics & Telecommunication**, CGPA: 7.4/10.0 **Feb 2016**

## ACHIEVEMENTS

**Acclaim Award (3-Star)** | **Amgen** | For compliance analytics work and cross-functional collaboration **Feb 2025**

## INTERESTS

Hiking state parks with my dog, skiing, trance music (Armin van Buuren), and strength training.