

Sai Karan Reddy

Data Scientist

Lake Forest, CA | saikaranreddy99@gmail.com | (774) 519-1165 | [LinkedIn](#)

SUMMARY

- Data Scientist with **5+ years** of experience building and deploying **machine learning, deep learning, and generative AI (LLM/RAG)** solutions across **financial services** and **large-scale systems**. Strong expertise in **predictive modeling, anomaly detection, credit risk, and AML analytics**, with a focus on delivering production-ready solutions.
- Experienced in **MLOps** and **scalable ML systems** using **PySpark, MLflow, Apache Airflow, Docker, Kubernetes, and AWS**, with emphasis on **model deployment, monitoring, drift detection, and governance**.
- Hands-on experience in **Generative AI** and **LLMOps**, including **fine-tuning (LoRA, PEFT)**, **RAG pipelines (LangChain, FAISS)**, and **LLM evaluation**, with a strong focus on **model interpretability (SHAP)** and **reliable AI systems**.

WORK EXPERIENCE

AMD | CA | Data Scientist

September 2024 – Present

- Built and deployed **predictive models (LightGBM, PySpark)** on telemetry data from **18K+ production nodes**, achieving **AUROC 0.88** for 30-day GPU/CPU failure prediction and reducing unplanned downtime in hyperscale **data center environments**.
- Engineered scalable **time-series features (ECC errors, thermal signals, PCIe telemetry)** using **PySpark**, improving model performance by **6% AUROC** and enabling distributed feature pipelines.
- Applied **XGBoost + Isolation Forest anomaly detection** on internally provided manufacturing telemetry, improving defect detection by **18%** over rule-based systems and accelerating **root cause analysis** for the hardware reliability team.
- Established **MLflow model registry** with **PSI-based drift monitoring** integrated into the **Airflow scoring pipeline**, ensuring **model governance, monitoring, and production reliability**.
- Fine-tuned **LLMs (LoRA, PEFT, HuggingFace Transformers)** on hardware failure logs, achieving **89% classification accuracy** and automating ticket tagging across **10K+ monthly support cases**, eliminating manual triage.
- Designed and deployed **RAG pipelines (LangChain, FAISS, vector search)** over hardware documentation and failure reports, reducing diagnosis time by **32%** compared to traditional keyword-based systems.
- Built **LLM evaluation pipelines** and applied **SHAP-based explainability** to ensure **model interpretability, validation, and production readiness**, establishing reusable **GenAI evaluation standards**.

JPMorgan Chase | India | Data Scientist

June 2019 – July 2023

- Developed and deployed **credit risk models (XGBoost, Python, AWS SageMaker)** as part of a cross-functional model risk team, to score **1.5M+ loan applications**, improving **Gini coefficient from 0.52 to 0.61 (+9%)** while maintaining target NPL thresholds under the **RBI model governance framework**.
- Engineered high-impact **behavioral features using PySpark** (payroll stability, spending patterns, savings indicators) with **WOE encoding** and **IV-based feature selection**, and implemented **SHAP-based explainability** for **adverse action reporting** and **regulatory compliance**.
- Built **AML anomaly detection systems (Isolation Forest, Autoencoders, TensorFlow, LightGBM)** achieving **AUROC 0.84** and reducing **false positives by 40%**, significantly improving **investigator efficiency** and **compliance operations**.
- Engineered large-scale **graph features using GraphX (Apache Spark)** to identify **fraud networks, transaction linkages, and second-degree connections** across financial networks, enhancing AML detection coverage in collaboration with the compliance analytics team.
- Deployed scalable **ML pipelines on AWS (S3, EC2, SageMaker)** using **FastAPI, Docker, Kubernetes, Apache Airflow**, in collaboration with the platform engineering team enabling **real-time inference (under 300ms latency)** and batch scoring across **2M+ customers monthly**.
- Implemented **MLflow for model lifecycle management** and built **automated model monitoring (data drift, PSI, distribution tracking)** to ensure **production stability** and **governance compliance**.

TECHNICAL SKILLS

Programming: Python, SQL, C++

Machine Learning & AI: Scikit-learn, XGBoost, LightGBM, TensorFlow, PyTorch, Deep Learning, NLP (BERT, spaCy), Time Series Modeling, Feature Engineering, Hyperparameter Tuning, Model Evaluation

Generative AI & LLMs: LLMs, HuggingFace Transformers, LoRA, PEFT, Retrieval-Augmented Generation (RAG), LangChain, FAISS, Prompt Engineering, LLM Evaluation, Vector Search

Data Engineering & Big Data: PySpark, Apache Spark, Hadoop, Distributed Computing

MLOps & Deployment: MLflow, Docker, Kubernetes, CI/CD Pipelines, Model Monitoring, Drift Detection (PSI), Model Governance

Cloud Platforms: AWS (S3, EC2, SageMaker, Glue, Redshift), Azure (Azure ML)

Data Analysis & Experimentation: Statistical Analysis, EDA, A/B Testing, Dimensionality Reduction, Outlier Detection

Databases & Warehousing: MySQL, PostgreSQL, MongoDB, Snowflake, Redshift

Visualization & BI: Tableau, Power BI (DAX), Matplotlib, Seaborn

Tools & Environments: Git, GitHub, Linux, Jupyter Notebooks, VS Code, Agile Methodology

EDUCATION

Master's in Computer Science | Clark University, Worcester, Massachusetts, USA.

Bachelor's in Computer Science | GITAM University, Telangana, India.