Shiva Velichalamala

Senior AI & ML Engineer

Email: shivakumarreddy722@gmail.com | **Phone:** +1 (614) 706-7619| LinkedIn

---

PROFESSIONAL SUMMARY

- **AI/ML Engineer and a Senior Data Scientist** with **9+** years of experience **designing, deploying,** and **optimizing AI-driven workflow automation** systems across finance**,** health**,** and insurance domains**.**
- Focused on building **LLM-based RAG and agentic AI systems**, integrating **GPT-4, LangChain, and LangGraph** to enable enterprise-scale automation, intelligent decision-making, and workflow optimization.
- Proficient in **Python** with expertise across **PyTorch, TensorFlow, Keras**, and modern frameworks (**FastAPI, Flask, Django**).
- Strong background in **data science & analytics**, leveraging **Pandas, NumPy, Scikit-learn, Dask, Spark,** and **BigQuery** to perform **advanced data wrangling, feature engineering,** and **large-scale distributed processing**.
- Hands-on experience with **LangGraph, Model Context Protocol (MCP), Guardrails AI** and multi-agent coordination frameworks.
- Strong foundation in data-centric AI building, optimizing, and monitoring data pipelines with **Spark**, **Airflow**, and **Databricks** to ensure model reliability and traceability.
- Hands-on with model evaluation and governance using **SHAP**, **LIME**, **Guardrails AI**, and **Evidently AI** to ensure explainability, bias detection, and production transparency.
- Expert in **Generative AI** and **multimodal systems**, leveraging text, image, and audio models (**Whisper**, **Stable Diffusion**) for contextual understanding and automation.
- Proven success integrating **MLOps practices (CI/CD, Docker, Kubernetes, SageMaker)** to deliver secure, reproducible, and scalable AI deployments.
- Delivered measurable impact, 50% faster loan approvals, 30% improvement in insight **accuracy**, and 15% reduction in financial risk across **enterprise AI projects**.
- Experienced in building **end-to-end data pipelines, predictive models,** and **real-time analytics workflows** that deliver actionable insights and support enterprise decision-making.
- Expert in **LLM-based intelligent automation**, **LangGraph-orchestrated workflows**, and **RAG pipelines** using GPT-4, LangChain, Pinecone, and Weaviate to drive **FinTech back-office efficiency** and **regulatory compliance**.
- Skilled in **cloud & DevOps**, with expertise in **AWS SageMaker, Lambda, ECS, Azure Cognitive Services, Docker, Kubernetes, Helm, and Terraform** to dealiver scalable**,** secure**,** and production-ready **AI/ML** deployments.
- Expert in **relational and NoSQL databases** including **PostgreSQL, MySQL, MongoDB, Redis,** and **Elasticsearch**, with experience in **query optimization, data modeling,** and **high-performance pipelines**.
- Expertise in **Generative AI models** (**GPT-3/4/5, LLaMA 2, Stable Diffusion, DALL·E, Whisper**) and NLP frameworks (**Hugging Face Transformers, spaCy, NLTK**) for developing advanced conversational and text-generation solutions.
- Designed and implemented **RAG pipelines** using **LangChain, Pinecone, FAISS,** and **Elastic Search** to enhance retrieval efficiency, achieving a **30% improvement** in **accuracy of financial insights** through optimized information access.
- Proven ability to **deploy production-grade AI models** and backend Python microservices (FastAPI, Flask) integrated with **Airflow, Spark, and Kafka pipelines**, delivering compliant automation for loan and risk workflows.
- Developed **multimodal AI solutions** combining **text, image,** and **speech processing** for advanced real-world applications, utilizing **Stable Diffusion, Whisper, and custom LLMs** to deliver seamless cross-modal intelligence.
- Fine-tuned **Large Language Models (LLMs)** for domain-specific use cases across healthcare, finance, and retail, aligning with compliance standards to **improve accuracy**, **contextual understanding**, and **production-grade performance.**
- Designed and implemented **Agent-to-Agent (A2A)** communication protocols enabling autonomous collaboration, dynamic task delegation, and seamless coordination among LLM-based agents.
- Expertise in **cloud-native AI deployments** using **AWS SageMaker**, **Azure Cognitive Services**, and **GCP Vertex AI**, integrating **CI/CD automation** pipelines for secure, scalable, and efficient production workflows.
- Implemented **MLOps practices** including **CI/CD (GitHub Actions, Jenkins, GitLab CI)**, **containerization (Docker, Kubernetes, Helm)**, and **experiment tracking (MLflow, Weights & Biases)**.
- Delivered **predictive analytics** and **anomaly detection systems** that significantly reduced **fraud and financial risks**.
- Developed **interactive dashboards** and **visualization tools** with **Tableau, Power BI, Matplotlib, Plotly,** and **Seaborn** to track AI performance, enabling **real-time monitoring** and **data-driven decision-making**.

TECHNICAL SKILLS

Programming Languages	Python (Advanced), SQL, Scala, JavaScript (Basic), Shell Scripting
AI / Machine Learning Frameworks	TensorFlow, PyTorch, Keras, Scikit-learn, XGBoost, LightGBM, CatBoost, ONNX, Hugging Face Transformers
Generative AI & LLM Ecosystem	GPT-3/4/5, Claude, Gemini, LLaMA-2, Falcon, Mistral, LangChain, LangGraph, Model Context Protocol (MCP), Guardrails AI, PromptLayer, LoRA / QLoRA, PEFT, LangSmith, OpenAI API, Vertex AI Agent Builder, AWS Bedrock.
RAG & Vector Databases	Pinecone, FAISS, Weaviate, ChromaDB, ElasticSearch, Vector Embeddings, Retrieval Evaluation (BLEU, ROUGE, F1)
Data Science & Analytics	Pandas, NumPy, SciPy, Dask, Spark MLlib, Databricks, Delta Lake, Feature Engineering, SHAP, LIME, Evidently AI (Drift Detection), Time-Series Forecasting (Prophet), Clustering, PCA
Cloud & MLOps Platforms	AWS (SageMaker, Bedrock, Lambda, ECS, S3, Redshift), Azure (AI Studio, Cognitive Services, Data Factory, Synapse, AKS), GCP (Vertex AI, BigQuery), MLflow, Weights & Biases, Kubeflow, CI/CD (GitHub Actions, Jenkins, GitLab CI), Docker, Kubernetes, Helm, Terraform
Data Engineering & Pipelines	Apache Spark, PySpark, Airflow, dbt, Kafka, Delta Lake, ETL/ELT Design, Data Validation, Feature Store (Feast)
Visualization & Monitoring	Tableau, Power BI, Plotly, Matplotlib, Seaborn, Grafana, Prometheus, Streamlit, Dash
Databases	Relational (PostgreSQL, MySQL) and Non-relational (MongoDB, Elasticsearch, Redis) Snowflake, Hive, Oracle 11g.
Security & Privacy	HIPAA, SOC2, GDPR compliance in AI/ML workflows
Big-Data Framework	Hadoop Ecosystem 2.X (HDFS, MapReduce, Hbase 0.9), Spark Framework 2.X (Scala 2.X, Spark SQL, Pyspark, Spark, Mllib)
AI Infrastructure and Platforms	NVIDIA CUDA (A100, T4), TensorRT, Mixed-Precision Training, Model Quantization, GPU Workload Profiling

PROFESSIONAL EXPERIENCE

Client: Alaska Airlines, Seattle, WA
Role: Senior AI-ML Engineer **Jan 2025 - Present**
Responsibilities:

- Delivered a cloud-native **ML automation platform** using **Python**, **PyTorch**, **TensorFlow**, **Spark**, and **Azure ML**, improving model throughput by ~40% and enabling near real-time document intelligence across high-volume clinical and operational document workflows for multiple enterprise teams.
- Designed an enterprise-grade **microservices** and **event-driven** architecture with containerized ML services on **AKS**, integrating **vector indexing**, **feature stores**, and data lake storage on **Azure Data Lake** and **Blob Storage** to support scalable training, inference, and analytics workloads.
- Worked in **Agile/Scrum** with bi-weekly sprints, collaborating closely with **Product Owners**, solution architects, data scientists, and platform engineers to refine user stories, groom backlogs, and align ML initiatives with regulatory, operational, and business objectives.
- Built robust **data ingestion pipelines** to pull structured and unstructured data from **REST APIs**, **SQL databases**, **Blob/S3 storage**, and **Kafka** topics, standardizing schemas and metadata for downstream training, feature generation, and inference services.
- Developed scalable **data processing workflows** using **Pandas**, **PySpark**, and **Airflow**, performing cleaning, normalization, enrichment, and feature engineering to transform raw clinical and financial documents into ML-ready datasets with traceable lineage.
- Managed distributed data storage across **Azure Data Lake**, curated **Delta-style layers**, and **Snowflake** warehouses, enabling low-latency access patterns for ML training, analytics queries, and near-real-time scoring workflows.
- Implemented specialized **vector storage** solutions using **FAISS** and **Pinecone**, supporting large-scale **embedding generation**, **nearest-neighbor search**, and **semantic retrieval** for document similarity, classification, and content-based recommendations.
- Selected, trained, and productionized **ML models** using **PyTorch**, **TensorFlow**, and **scikit-learn** for document classification, entity extraction, and risk-related scoring, aligning model architectures with latency, accuracy, and interpretability requirements.
- Applied advanced techniques such as **retrieval-based scoring**, **vector similarity search**, mini-batch inference, and **streaming-style pipelines** to handle high-volume workloads while maintaining predictable response times and resource utilization.

- Optimized ML pipelines with **hyperparameter tuning**, **model quantization**, **caching**, and dynamic batching strategies, improving end-to-end inference latency and reducing GPU/CPU costs across production environments.
- Used frameworks including **FastAPI**, **Spark MLlib**, and **Hugging Face** components to wrap models as reusable services, exposing standardized REST endpoints for internal consumers and downstream orchestration layers.
- Enforced strong **coding standards** through **modular Python**, **OOP-based ML components**, shared utility libraries, and code review practices, reducing technical debt and simplifying onboarding for new engineers and data scientists.
- Built comprehensive **evaluation workflows** with offline validation, **cross-validation**, **drift analysis**, and controlled **PoC benchmarking**, ensuring only rigorously tested models progressed into higher environments and production systems.
- Containerized all ML workloads using **Docker** and published images to **Azure Container Registry**, standardizing runtime environments and enabling consistent deployments across development, staging, and production clusters.
- Deployed and managed high-availability **inference endpoints** on **AKS** using **KServe**, horizontal **autoscaling**, and health probes, ensuring reliable performance under fluctuating traffic patterns and seasonal peaks.
- Implemented automated **CI/CD pipelines** in **GitHub Actions** for building, testing, containerizing, and deploying ML models, data pipelines, and microservices with minimal manual intervention and traceable change histories.
- Used **Terraform** and **Helm** to define and provision **AKS clusters**, networking, storage classes, secrets, and supporting cloud resources as code, ensuring reproducibility, auditability, and consistent infrastructure configurations.
- Established end-to-end **observability** with **Prometheus**, **Grafana**, and **Evidently AI**, monitoring latency, resource utilization, data drift, and prediction quality through custom dashboards and automated alerts.
- Implemented automated **unit tests**, **integration tests**, and **pipeline validation checks** using **PyTest** and **Airflow** test hooks, catching data and logic issues early before impacting downstream consumers and production workloads.
- Authored detailed **technical documentation**, **architecture diagrams**, **API specifications**, and internal **runbooks**, and led recurring **knowledge transfer sessions** for platform engineers, data scientists, and operations teams to ensure sustainable ownership.

Environment: Python, PyTorch, TensorFlow, scikit-learn, Spark, PySpark, Azure ML, Azure Data Lake, Azure Blob Storage, Snowflake, FAISS, Pinecone, FastAPI, Airflow, Docker, AKS, KServe, MLflow, DVC, Terraform, Helm, Prometheus, Grafana, GitHub Actions

Client: State of CA, SFO, CA
Role: Senior ML Engineer **Aug 2023 – Dec 2024**
Project: Real-Time Financial Risk Monitoring & Predictive Analytics Platform
Responsibilities:

- Delivered an end-to-end **financial risk ML platform** using **PyTorch**, **TensorFlow**, **scikit-learn**, and **Azure ML**, improving risk detection accuracy by ~35% and enabling earlier identification of anomalies across large-scale transactional data streams.
- Designed a modular **ML architecture** with containerized **microservices**, **event-driven retraining workflows**, centralized **model registry**, and integrated **feature store**, supporting governed model promotion and consistent features across multiple risk applications.
- Participated in **Agile/Scrum** sprints, collaborating with **risk**, **compliance**, **data**, and **product** teams to prioritize ML features, refine acceptance criteria, and align model behavior with regulatory and business expectations.
- Built resilient **ingestion pipelines** consuming **SQL warehouse** tables, **Azure Blob Storage**, **REST APIs**, and **Kafka** streams, normalizing disparate financial feeds into unified, versioned ML datasets.
- Engineered reusable **preprocessing and feature pipelines** using **Python**, **Pandas**, **scikit-learn**, and **PySpark**, ensuring consistent transformations between offline training workflows and online inference services.
- Managed curated ML data layers in **Azure Data Lake**, **Delta-style tables**, and **Snowflake**, tracked through **DVC** and **MLflow** for full lineage, experiment repeatability, and audit-ready history.
- Implemented and maintained a **feature store** using **Feast**, enabling centralized definition, low-latency serving, and cross-team reuse of critical risk and behavioral features.
- Developed **ML models** for **anomaly detection**, **fraud scoring**, and **time-series forecasting**, targeting transactional irregularities, late payments, and emerging risk patterns in financial portfolios.
- Applied advanced methods including **ensemble modeling**, **sliding-window forecasting**, **rare-event modeling**, and custom loss functions to improve performance on highly **imbalanced datasets** with limited labeled positives.
- Optimized models through systematic **hyperparameter tuning**, **model pruning**, **ONNX export**, and inference graph optimizations, reducing latency and compute costs while preserving required accuracy thresholds.

- Leveraged **PyTorch Lightning**, **TensorFlow**, **Spark ML**, and **FastAPI** to standardize training pipelines and expose production-ready inference endpoints integrated with upstream and downstream risk systems.
- Established robust **coding standards**, including **modular architecture**, **configuration-driven pipelines**, shared utility layers, and **code reviews**, increasing maintainability and reducing regression risk across ML repositories.
- Built reusable **evaluation frameworks** using **cross-validation**, **ROC/AUC**, **precision/recall**, **KS statistics**, and **challenger-versus-champion** comparisons to support regulatory model validation and governance.
- Containerized ML services with **Docker** and published images to **Azure Container Registry**, standardizing runtime environments and simplifying promotion between dev, test, and production clusters.
- Deployed models onto **AKS** using **GPU-backed nodes**, **HPA-based autoscaling**, and **blue/green** rollout strategies, minimizing downtime and enabling safe, incremental model upgrades.
- Implemented end-to-end **CI/CD** using **GitHub Actions**, **Azure DevOps**, and **Argo Workflows** for **continuous training (CT)** and **continuous deployment (CD)** of ML models, pipelines, and infrastructure components.
- Provisioned and managed cloud **infrastructure** via **Terraform**, covering **AKS clusters**, virtual networks, storage accounts, secrets management, and supporting PaaS resources with environment-specific configurations.
- Set up comprehensive **monitoring** using **Azure Monitor**, **Prometheus**, and **Grafana**, tracking pipeline health, **data drift**, **model performance**, and infrastructure utilization through dashboards and alert rules.
- Built automated **unit tests**, **integration tests**, and **UAT support** scripts leveraging **PyTest** and custom validation utilities, ensuring that new model versions and pipeline changes met reliability and compliance standards.
- Documented end-to-end **ML workflows**, **feature dictionaries**, **deployment runbooks**, and **troubleshooting guides**, and led **onboarding workshops** to enable analysts, engineers, and risk stakeholders to adopt and extend the ML platform.

Environment: Python, PyTorch, TensorFlow, scikit-learn, XGBoost, Spark, Feast, MLflow, DVC, Airflow, FastAPI, Docker, Kubernetes/AKS, Helm, Terraform, Azure ML, Azure Data Lake, Azure Blob Storage, Azure Monitor, Azure DevOps, Prometheus, Grafana, GitHub Actions

Client: Elevance Health, Indianapolis, IN.
Role: ML Engineer **Mar 2021 – Jul 2023**
Project: Enterprise Predictive Analytics & Operational Risk Intelligence Platform
Responsibilities:

- Delivered an enterprise-wide **ML platform** for **predictive analytics** and **operational risk** using **PyTorch**, **TensorFlow**, **XGBoost**, and **Azure ML**, improving risk signal detection and decision-making speed.
- Designed a **modular ML architecture** with containerized **microservices**, a central **model registry (MLflow)**, and **batch/real-time scoring** flows on **AKS**.
- Worked in **Agile/Scrum** with 2-week sprints, collaborating with **Product Owners**, **data scientists**, and **engineers** to refine ML use cases and acceptance criteria.
- Built **ingestion pipelines** using **Azure Functions**, **REST APIs**, **Azure Blob Storage**, and database connectors to onboard financial and operational data.
- Engineered **preprocessing and feature pipelines** using **Python**, **Pandas**, **scikit-learn**, and **Azure ML Pipelines** to standardize transformations across training and inference.
- Managed **storage** across **Azure Data Lake** and curated **Delta-style layers** feeding analytics and ML workloads.
- Implemented **data quality checks**, **schema validation**, and **drift-aware** input monitoring to protect model stability in production.
- Trained models (**XGBoost**, **neural networks** in **PyTorch/TensorFlow**) for **risk scoring**, **anomaly detection**, and **operational forecasting**.
- Applied **ensemble modeling**, **time-series analysis**, and **imbalance handling** to improve recall on rare risk events.
- Optimized performance via **hyperparameter tuning**, **cross-validation**, and **model compression** techniques.
- Used **TensorFlow Serving**, **TorchServe**, **FastAPI**, and **MLflow** to package, register, and deploy ML models at scale.
- Enforced **modular Python code**, **reusable ML components**, and **configuration-driven pipelines** with rigorous **code reviews**.
- Built **evaluation workflows** with **MLflow experiments** and standardized metrics to select and promote models to production.
- Containerized ML workloads with **Docker** and stored images in **Azure Container Registry** for consistent deployment across environments.
- Deployed and orchestrated models on **AKS** using **Helm** and **Terraform**, integrating **Prometheus**, **Grafana**, and **Azure Monitor** for end-to-end observability.

Environment: Python, PyTorch, TensorFlow, XGBoost, scikit-learn, FastAPI, Docker, Kubernetes/AKS, Helm, Terraform, Airflow, MLflow, DVC, Azure ML, Azure Functions, Azure Blob Storage, Azure Container Registry, Azure Key Vault, Azure Monitor, Azure DevOps, TensorFlow Serving, TorchServe, Prometheus, Grafana, GitHub Actions

Client: Global Atlantic financial group, New York, NY
Role: Data Scientist **Aug 2018 – Jan 2020**
Responsibilities:

- Developed **customer churn** and **credit risk prediction** solutions using **XGBoost**, **LightGBM**, and **scikit-learn**, enabling data-driven retention and risk strategies.
- Designed **analytics workflows** combining **batch scoring**, **model explanations**, and **BI reporting** to deliver actionable insights to stakeholders.
- Worked with **product**, **risk**, and **marketing** teams to refine features, thresholds, and rules based on model outputs.
- Built **feature pipelines** using **Pandas**, **NumPy**, and **SQL** to process customer demographics, transaction histories, and behavioral signals.
- Integrated data from **MySQL**, **PostgreSQL**, and flat-file sources with robust cleaning, imputation, and normalization steps.
- Managed **curated datasets** in **relational stores** optimized for model training, analytics, and reporting.
- Implemented **data quality checks** and validations to ensure stable model inputs and trustworthy insights.
- Selected algorithms such as **XGBoost**, **LightGBM**, **logistic regression**, and **tree ensembles** for **binary classification** and **risk scoring**.
- Used **feature importance analysis**, **class-imbalance handling**, and **probability calibration** to align outputs with real-world risk behavior.
- Ran **grid search/random search** with **cross-validation** to maximize **AUC**, **recall**, and business KPIs.
- Deployed models via **Flask-based APIs**, containerized with **Docker**, and integrated them into internal decision-support tools.
- Built **EDA reports** and **visualization dashboards** using **Seaborn** and **Plotly**, explaining key churn and risk drivers to non-technical audiences.

Environment: Python, scikit-learn, XGBoost, LightGBM, Pandas, NumPy, Flask, Docker, spaCy, NLTK, Seaborn, Plotly, MySQL, PostgreSQL, Git, GitHub

Client: Trigent Software – Bangalore, India
Role: Data Engineer **Mar 2016 – Jun 2018**
Responsibilities:

- Engineered **data and analytics automation** solutions using **Python**, **Pandas**, **NumPy**, **SQL**, and **Tableau** to streamline BI workflows and predictive reporting.
- Designed **ETL architectures** to feed downstream **analytics** and **ML models** supporting sales forecasting, segmentation, and performance tracking.
- Built **ingestion jobs** extracting data from **relational databases** and **CSV/flat files** into centralized analytical schemas.
- Cleaned, transformed, and aggregated **raw data** using **Pandas/NumPy**, standardizing metrics and dimensions for dashboards.
- Managed storage in **SQL-based warehouses** and optimized query performance using **indexing**, **partitioning**, and efficient schema design.
- Implemented **data validation**, **anomaly detection scripts**, and **reconciliation checks** to ensure high-quality datasets for reporting and analytics.
- Developed baseline **predictive models** using **scikit-learn** (regression, classification, clustering) to support **revenue forecasting** and **customer segmentation**.
- Automated recurring **ETL** and **reporting workflows** using **Python scripts** and schedulers, reducing manual report generation.
- Built **Tableau dashboards** for **KPI** and **trend analysis**, integrating ML outputs to enhance decision support.
- Standardized development practices with **Git-based version control** and **modular Python code** for maintainability and reuse.

Environment: Python, Pandas, NumPy, scikit-learn, SQL, Tableau, Git

Eduction :
Sathyabama University in Computer science 2015