

Srikar Reddy Nelavetla

Machine Learning Engineer

Seattle, WA | (720) 234-7493 | srikarreddy3@gmail.com | [LinkedIn](#) | [Portfolio](#)

SUMMARY

Machine Learning Engineer with **3+** years of experience designing, building, and deploying **scalable machine learning systems**, **LLM-powered applications**, and **real-time data pipelines** in production environments. Expertise in **Natural Language Processing (NLP)**, **Retrieval-Augmented Generation (RAG)**, **transformer model fine-tuning**, and **vector search (HNSW)**. Proven track record of improving **model accuracy**, **retrieval performance**, and **system latency** through hands-on experience with **Python**, **PyTorch**, **Hugging Face**, **Apache Spark**, **Elasticsearch**, and **AWS Bedrock**. Strong background in **MLOps**, **CI/CD**, and **cloud-based ML platforms**, leveraging tools such as **MLflow**, **Docker**, **Kubernetes**, **Airflow**, and **GitHub Actions** to deliver reliable and scalable ML solutions. Adept at collaborating with cross-functional teams to translate business requirements into **high-impact AI-driven products** that enhance performance, scalability, and user experience.

TECHNICAL SKILLS

Programming & Databases: Python, SQL, R, MySQL, PostgreSQL, MongoDB

Machine Learning & AI: Supervised & Unsupervised Learning, Deep Learning (CNN, RNN, LSTM), Natural Language Processing (NLP), Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), Reinforcement Learning, Time Series Forecasting, Feature Engineering, Hyperparameter Tuning, Prompt Engineering, AWS Bedrock

Frameworks & Libraries: PyTorch, TensorFlow, Scikit-learn, XGBoost, Hugging Face Transformers, BERT, TensorRT

Data Engineering & Big Data: Apache Spark, Apache Kafka, Apache Airflow, Databricks, ETL Pipelines

Cloud Platforms: AWS (SageMaker, EC2, S3, RDS, Redshift, Lambda), Snowflake

MLOps & Deployment: MLflow, Model Deployment, Model Monitoring, Model Versioning, CI/CD Pipelines, Experiment Tracking

DevOps & Infrastructure: Docker, Kubernetes, Jenkins, GitHub Actions, Terraform, Containerization, Infrastructure as Code (IaC)

Data Analysis & Visualization: Tableau, Power BI, Excel, Exploratory Data Analysis (EDA), A/B Testing, Statistical Modeling

PROFESSIONAL EXPERIENCE

Elastic | *Machine Learning Engineer*

CA | January 2025 – Present

- Designed and deployed **end-to-end machine learning pipelines** using **Elastic Stack (Elasticsearch, Kibana, Elastic ML)** for **real-time anomaly detection** and **log analytics**, supporting **enterprise-scale observability platforms**.
- Fine-tuned **transformer-based NLP models** using **PyTorch** and **Hugging Face Transformers** for **semantic search** and **relevance ranking**, improving retrieval accuracy by **18%** across large-scale document corpora.
- Built **LLM-powered Retrieval-Augmented Generation (RAG) systems** integrated with **AWS Bedrock (Claude, Titan)** and **Elasticsearch vector search (HNSW)**, improving **knowledge retrieval precision** by **22%** and enabling **scalable AI-driven search experiences**.
- Implemented and maintained **MLOps pipelines** using **MLflow**, **GitHub Actions**, and **Elasticsearch Model Management APIs**, enabling **model versioning**, **CI/CD automation**, and **monitoring**, reducing deployment cycle time by **25%**.
- Developed **data preprocessing** and **feature engineering pipelines** using **Pandas**, **NumPy**, and **Apache Spark**, processing **10M+ events per day** to power **real-time inference systems**.
- Optimized **search queries** and **ML inference performance** in **Elasticsearch clusters**, reducing **p99 latency** by **30%** and improving system scalability and responsiveness.

IBM | *Machine Learning Engineer*

India | May 2021 – May 2023

- Designed and implemented **scalable ETL pipelines** using **Python**, **Apache Spark**, and **SQL** on **IBM Cloud Pak for Data**, processing **terabytes of enterprise data** and reducing data preparation time by **25%**, accelerating **model training** and **deployment cycles**.
- Developed and deployed **machine learning models** using **XGBoost**, **LightGBM**, **K-Means**, and **DBSCAN** for **fraud detection**, **customer segmentation**, and **predictive maintenance**, improving **anomaly detection** accuracy by **30%** and reducing false positives by **22%**.
- Productionized ML solutions on **IBM Cloud (Watson Studio, Watson Machine Learning)** by building **real-time inference APIs (REST/gRPC)**, achieving **sub-150ms latency** for high-throughput business applications.
- Orchestrated **end-to-end ML workflows** using **Apache Airflow DAGs** for **data ingestion**, **feature engineering**, and **automated model retraining**, reducing manual intervention by **60%** and ensuring **continuous model freshness**.
- Built and optimized **CI/CD pipelines for machine learning** using **Docker**, **Kubernetes**, **Jenkins**, and **Git**, implementing **automated testing** and **canary deployments**, reducing deployment time by **40%** and improving release reliability.
- Collaborated in **Agile/Scrum environments** with **data scientists**, **engineers**, and **product stakeholders** to translate business requirements into scalable ML solutions, contributing to a **15%** increase in customer retention.

EDUCATION

Master's in Data Science | University of Colorado, Boulder, CO, USA.

