SREE KRISHNA SURESH

Dallas, TX • +1 (857) 318-6995 • sreekrishnadav@gmail.com • LinkedIn • Github • Portfolio

5+ years of experience | Award-winning ML solutions | $1M+ profit impact | Gen AI

TECHNICAL SKILLS

Programming and Database: Python, SQL, R, Java, Vector Databases (Faiss, Pinecone, Azure Cosmos DB), Elasticsearch
Data Science and ML: NLP (LLMs, RAG, AI Agents), Regression Models, Time Series Forecasting, Computer Vision, A/B Testing
Frameworks and Tools: PyTorch, TensorFlow, Langchain, LlamaIndex, pandas, scikit-learn, Hugging Face Transformers, Git
Cloud and DevOps: AWS (Bedrock, SageMaker, EC2, Airflow, S3), Azure, MLflow, Docker, Jenkins, Terraform, Kubernetes
Certifications: Azure Data Science Certified, PCAP (Python), Machine Learning Specialization - Coursera

PROFESSIONAL EXPERIENCE

Infosys, Richardson, TX **Dec 2024 – Present**
AI Engineer

- Led team at Everest Insurance global hackathon, winning runner-up and securing $200K in funding for developing a no-code, drag-and-drop AI Agent architecture orchestration platform, enabling rapid prototyping of AI-driven solutions.
- Citigroup - NLP Engine for Trade Management
  - Automated extraction and structuring of complex trader data by developing software with fine-tuned entity recognition LLMs and custom-built Deep learning models.
  - Retrained sentence transformer model (Hugging Face) and implemented advanced prompt engineering strategies for efficient data summarization and semantic search/ranking, eliminating external APIs usage and reduced operational costs by 70%.
- Exxon Mobil - Advanced chat Interface and real-time inventory interaction platform
  - Designed and built multimodal RAG application with Azure AI Search, Python SDK and AI Studio, and deployed on Bot Service contributing to successful client acquisition.
  - Orchestrated Microsoft Fabric data pipelines, developed and integrated SQL-based Agent Chatbot deployed on Azure cloud as web application.
- Microsoft - Outlook Mail automation for Q&A forum - PoC
  - Developed and deployed real-time email response system using Microsoft Graph API, webhooks and GPT-4o, created LLM-optimized datasets for NLP feature engineering, processing over 10,000 emails daily.
  - Engineered scalable Flask application with secure OAuth 2.0 and API integration, demonstrating microservices architecture for large-scale ML model serving and integration.
- Designed and implemented a scalable and modular AI-powered internal talent management platform leveraging multi-agent based architecture (Langchain, LangGraph) and RAG over Graph DB accelerating time to hire by 40%.

Abecedarian LLC, Boston, MA **Feb 2024 – Oct 2024**
Machine Learning Engineer

- Architected and operationalized a full-stack GenAI application for intelligent candidate shortlisting, integrating RAG-LLM, BERT embeddings, and GPT-based models with Faiss vector search. Orchestrated CI/CD via Jenkins, GitHub Actions, and Docker, driving an 80% increase in user engagement.
- Refactored and wrote 400 lines of code in Python for an AI-powered job ranking application using LlamaIndex, OpenAI LLMs and reranking transformer models. Integrated with responsible AI monitoring from Azure services.
- Led data-driven decisions among stakeholders by creating Tableau dashboards for trend analysis and performing RCA using PowerBI.
- Decreased model deployment time by 50% across 5 product teams by spearheading a cross-functional MLOps initiative, implementing robust data drift detection and mitigation strategies, and using MLflow for model tracking and versioning.
- Prototyped the integration of fine-tuned LLaMA models into enterprise ML platforms using QLoRA and PEFT techniques. Conducted iterative experimentation on representative datasets, optimizing performance metrics such as perplexity, ROUGE, and F1-score.

Lightforce Orthodontics, Boston, MA **Jan 2023 – July 2023**
Machine Learning Engineer

- Automated processing and analysis of 5 TB of point cloud data monthly by designing a scalable data pipeline on AWS using Airflow.
- Reduced product design costs by $50K annually with a novel 3D CNN PointNet solution built using PyTorch on AWS SageMaker.
- Enhanced 3D object segmentation model accuracy by 25% via feature engineering, hyperparameter tuning, and testing loss functions.
- Streamlined CI/CD workflows and deployment using Terraform and TeamCity, enabling faster updates and releases.

Northeastern University, Boston, MA **Sep 2022 – Dec 2023**
Graduate Teaching Assistant - Machine Learning

- Improved grades of 40 students by mentoring projects and tutoring on topics including supervised machine learning and data mining.
- Increased student engagement by 20% through management of course logistics, assignments, and in-class activities along with faculty.

Wipro, India **May 2018 – Aug 2021**
Data Engineer - Machine Learning

- Optimized telecommunication network equipment distribution by developing and deploying predictive models (XGBoost, LightGBM) in Python, accelerating network infrastructure rollouts by 20% and reducing logistics overhead by 10%.
- Improved telecommunication equipment inventory management using time-series forecasting (ARIMA, LSTM), minimizing service disruptions from part stockouts and decreasing holding costs by 15%.
- Streamlined large-scale data processing for ML workflows, optimizing manual data preparation time by 60% for 100 million records through GraphQL API implementation and SQL script optimization.
- Boosted server response time by 9 seconds and accelerated data retrieval by 25% for ML model inference by designing a custom ETL pipeline between PostgreSQL and Elasticsearch, resulting in $100K in annual operational cost savings.
- Led cross-functional initiatives by conducting agile sprint meetings and data architecture workshops, and mentored junior team members, resulting in a 30% increase in project adoption for new ML solutions.

EDUCATION

Northeastern University, Khoury College of Computer Science, Boston, MA **Sep 2021 - Dec 2023**
Master of Science in Data Science
Relevant Coursework: Statistical Modelling, Machine Learning, Deep Learning, Data Mining, NLP, Big Data Analytics

PROJECTS & AWARDS

Engineered a multimodal AI system for advanced financial market analysis and investment strategy, integrating a fine-tuned GPT-3.5 LLM (via OpenAI API) with advanced CNN-Transformer architectures, Azure Document Intelligence, and BERT for comprehensive data extraction, orchestrated via LangChain. Won ”The Novel Model Architecture” award and ranked in top 5 among 2300 teams at the Cloudera Hackathon.

Crop Production Data Analytics and Yield Prediction with Remote Sensing Data (link)- Built a custom ML model from scratch

Recognized with the Panache Award - outstanding performance and high learning ability at Wipro (2020)