Prasanna Aleti

Phone: (940)-208-9668 | **Email:** prasannaaleti26@gmail.com | LinkedIn

SUMMARY

AI/ML Engineer with **4+** years of experience deploying LLM/RAG pipelines, predictive analytics, and MLOps solutions across AWS, Azure, and GCP. Delivered measurable business impact — 22% fewer fiber outages at Coforge and 70% faster audit reporting at Verizon. Skilled in Python, SQL, TensorFlow, PyTorch, Hugging Face, and Scikit-learn, with strong expertise in embeddings, explainability (SHAP/LIME), and scalable deployment (MLflow, Airflow, FastAPI, Docker, Kubernetes). Master's in Data Science (AI concentration), with additional experience mentoring and supporting healthcare predictive modeling research as a Graduate Teaching Assistant.

SKILLS

- **Programming Languages:** Python, SQL, Bash, Java, C#, R
- **ML & DL Frameworks:** TensorFlow, PyTorch, Scikit-learn, Keras, Hugging Face Transformers
- **AI/ML & GenAI Techniques:** Supervised Learning, Unsupervised Learning, Deep Learning, NLP, LLMs (GPT, Claude, LLaMA), Prompt Engineering, Few-shot Learning, Fine-tuning, Embeddings, RAG, Text Summarization, Conversational AI
- **MLOps & Deployment:** MLflow, Apache Airflow, FastAPI, Flask, Docker, Kubernetes, Model Monitoring, A/B Testing, Experiment Tracking
- **Cloud Platforms:** AWS (SageMaker, Bedrock, Lambda, S3), Azure (Azure ML, OpenAI), GCP (Vertex AI, PaLM 2)
- **Data Engineering:** PySpark, Apache Spark, Delta Lake, ETL Pipelines, Pandas, NumPy, Vector DBs (Pinecone, ChromaDB, FAISS)
- **Model Evaluation & Explainability:** SHAP, LIME, Regression, Clustering, BLEU, ROUGE, Hypothesis Testing
- **Version Control & CI/CD:** Git, GitHub Actions, Jenkins
- **APIs & Integration:** REST APIs, OpenAPI, Swagger, LangChain, Gradio, Streamlit
- **Agile Collaboration:** Jira, Confluence, Scrum Methodology, Technical Wikis
- **Model Serving & Inference:** SageMaker Endpoints, Azure ML Pipelines, OpenAI API, Docker + Kubernetes

EXPERIENCE

Verizon Wireless Systems **Jan 2024 – Current**

AI Engineer

- Developed and deployed a RAG pipeline with Claude Sonnet LLM on AWS Lambda, automating insurance audit reports and reducing manual effort by 70%, cutting report turnaround from days to hours.
- Implemented FAISS, LangChain, and SentenceTransformer libraries to enhance document retrieval precision, resulting in 50% faster report generation.
- Built a chatbot using BERT and Agentic AI for intent recognition, combined with RAG models and SQLAlchemy pipelines for context fetching, achieving 95% accuracy in handling queries and improving resolution rates by 60%.
- Fine-tuned a machine learning model on Amazon SageMaker, adjusting learning rate, batch size, and optimizer settings to improve training stability and performance.
- Contributed to a web-based platform for seamless interaction with models, increasing user engagement metrics by 30%.
- Used AWS EC2, Bedrock, S3, and Lambda for scalable computing, secure data storage, and serverless operations, enhancing claims management efficiency and cost-effectiveness.
- Assisted in migrating legacy database systems to PostgreSQL, improving data integrity, scalability, and operational efficiency across the enterprise.
- Improved system performance by 40% through code optimizations in React, making rendering faster and applications more responsive.

Coforge **Feb 2020 - Jul 2022**

Associate ML Engineer

- Contributed to the development of a Random Forest–based early warning system, reducing fiber outages by 22% in pilot rollout and cutting false alerts by 35% through model refinements.
- Built ETL pipelines with Apache Airflow + PySpark, processing millions of daily telemetry records and storing outputs in AWS S3 for scalable downstream analytics.
- Wrote SQL queries for data validation and feature engineering, improving the accuracy and quality of training datasets.
- Applied SHAP explainability in Python to highlight key features driving predictions, increasing trust among field engineers.
- Deployed results through a Plotly Dash dashboard (Python), giving non-technical stakeholders real-time visibility into fiber risk by region and route.
- Implemented a FastAPI REST API for real-time predictions with less than 200ms latency, enabling field teams to act quickly on risk alerts.
- Set up monitoring and alerting via Airflow SLA checks and Slack integration, improving response times to high-risk alerts.
- Collaborated with network engineers to refine model logic, reducing false positives by ~35% and improving signal-to-noise ratio in alerts.

EDUCATION

- **Master of Science: Data Science (**AI Concentration**)**
  University of North Texas
- **Bachelor of Technology: Electronics & Communication Engineering**
  VNRVJIET, Hyderabad, Telangana

TEACHING & RESEARCH EXPERIENCE

Graduate Teaching Assistant (*University of North Texas*)**.** **Aug 2022-Dec 2023**

- Assisted in a student research project on predictive modeling for healthcare outcomes, helping build ML models in Python (Scikit-learn, TensorFlow, XGBoost) that achieved ~85% accuracy in predicting chronic diseases.
- Supported AI-driven healthcare research with a focus on early intervention and preventive medicine, exploring how predictive modeling and explainability methods (SHAP, LIME) could be applied at scale to reduce chronic illness burden, lower healthcare costs and improve patient outcomes.