Sai Nikhil Mattapalli

AI/ML ENGINEER

NJ, USA | 5185521672 | ms1104n@gmail.com | LinkedIn

Professional Summary

AI/ML Engineer with 5 years of experience building and deploying scalable Machine Learning and Generative AI solutions across healthcare and IT domains. Proficient in Python, SQL, Scikit-learn, XGBoost, PyTorch, and TensorFlow for end-to-end ML pipelines. Experienced in LLMs, RAG, LangChain, and vector databases, with strong skills in REST APIs (FastAPI, Node.js), Docker, AWS, and data visualization using Power BI and Tableau.

Skills

Language:	Python, React, Node.js, JavaScript, TypeScript, C++, Java, SQL (Postgres), HTML/CSS, R
Data & Statistics:	Panda s, NumPy, Feature Engineering, Hypothesis Testing, A/B Testing, Regression, Classification, Clustering
ML/AI:	Scikit-learn, XGBoost / LightGBM, PyTorch, TensorFlow / Keras, Hugging Face Transformers, Model evaluation (ROC-AUC, F1, Precision-Recall, SHAP, bias testing), LLM APIs (OpenAI, Anthropic, LLaMA-based models), RAG (Retrieval-Augmented Generation), LangChain, Azure OpenAI (GPT-4o), serpapi, LlamaIndex, CrewAI (multi-agent workflows), Prompt Engineering & Evaluation, Fine-tuning (LoRA / PEFT), Embedding pipelines, Token optimization & cost monitoring
Cloud & Databases:	SQL(Postgres), NoSQL (MongoDB, Firebase), VectorDB (Pinecone, ChromaDB, FAISS), AWS, Vercel
Developer Tools:	Git, Docker, FastAPI, Flask, Kubernetes, CI/CD, MLflow, n8n (workflow automation / orchestration), Weights & Biases, Tensor Board, VS Code
Backend APIs:	FastAPI, Node.js, REST APIs

Work Experience

AI/ML Engineer | Molina Heath, USA — Aug 2023 - Present

- Designed and implemented end-to-end ML pipelines using Python (Pandas, NumPy, Scikit-learn) for member risk stratification and cost prediction, improving care management decision-making for Medicaid and Medicare populations.
- Processed large-scale **claims, eligibility, and provider datasets** using PySpark, enhancing data pipeline scalability and transformation efficiency across payer systems.
- Developed predictive models using XGBoost, Random Forest, and TensorFlow to **identify high-risk members, predict hospital readmissions, and detect care gaps**, enabling proactive intervention strategies.
- Optimized model performance using hyperparameter tuning, improving accuracy by 15% and enhancing AUC and F1-score for **risk adjustment and utilization prediction models**.
- Designed scalable data architecture using AWS S3 and Snowflake for **claims and member data ingestion**, reducing data retrieval time by 30% and improving reporting efficiency.
- Built **RAG pipelines with LangChain** to generate explainable summaries of member health risk, claims history, and recommendations using LLMs.
- Improved inference performance by building low-latency RAG systems, reducing latency by 22% and improving response efficiency by 12% for **real-time care management insights**.
- Designed AI-powered decision support systems combining ML predictions with contextual insights to support **care coordinators, case managers, and provider engagement workflows**.
- Developed Power BI dashboards to visualize **member risk scores, HEDIS quality metrics, utilization trends, and cost KPIs**, enabling data-driven decisions across clinical and operational teams.

Software Engineer | Cognizant, India — Feb 2020 - Aug 2022

- Automated reporting workflows using **Tableau extracts and scheduled refreshes**, streamlining data pipelines and reducing manual reporting effort by **30%**, improving data availability and reporting efficiency for stakeholders.
- Generated high-quality embeddings using **Hugging Face and OpenAI models**, and stored them in **vector databases (Pinecone, FAISS, ChromaDB)** to enable efficient semantic search and retrieval in RAG pipelines.
- Conducted exploratory data analysis (EDA) and statistical testing using **SciPy and Statsmodels**, identifying patterns and anomalies, improving feature engineering effectiveness by **20%** and model accuracy by **12%**.
- Integrated **LLM APIs (OpenAI, LLaMA-based models)** with advanced **prompt engineering techniques** to generate domain-specific, context-aware responses for intelligent analytics applications.
- Developed and deployed **machine learning models (classification, regression, clustering)** using **LightGBM and PyTorch**, enabling accurate predictions and supporting data-driven decision-making across business use cases.
- Built scalable and high-performance **REST APIs using FastAPI and Flask**, enabling real-time ML and GenAI inference, improving response time by **45%** and system integration efficiency by **60%**.
- Implemented **custom prompt templates using LangChain**, enabling, accurate, and explainable responses within RAG-based AI systems.
- Deployed and managed applications using **Docker and AWS**, ensuring scalability, reliability, and production readiness with resource utilization.

Education

Master of Science in Computer Science – Suny Albany at Albany, USA
Bachelor of Technology in Computer Science – Bharath University, India