Nitish Kumar Manthri

• nitish.m@savemymails.com • +1(901)659-4445 • Open to Relocate• LinkedIn

SUMMARY

AI/ML Engineer with around 4 years of experience delivering production-grade machine learning and generative AI systems across healthcare, hospital billing, and insurance domains. Skilled in ML, deep learning, LLMs, cloud platforms, MLOps, data pipelines, and scalable AI systems. Successfully developed predictive patient payment and risk models, reducing operational costs $250K/year and resolving 6,800 delayed accounts monthly. Experienced with Generative AI, RAG pipelines, Transformers, Hugging Face, CNN/LSTM multimodal models, and semantic search, delivering high-performance AI solutions with measurable clinical and business impact.

SKILLS

Machine Learning & Predictive Modeling: LightGBM, CatBoost, Random Forest, XGBoost, Gradient Boosting, Regression, Decision Trees, Support Vector Machines, Prophet, Temporal Fusion
Deep Learning & AI: PyTorch, TensorFlow, Keras, CNNs, LSTM, Transformers, GPT-4, LLaMA, CLIP, Multimodal Vision-Language Models
Generative AI & NLP: Hugging Face Transformers, LangChain, Pinecone, FAISS, Weaviate, RAG Pipelines, Chatbots, Prompt Engineering
Programming & Data: Python, R, Julia, NumPy, Pandas, SQL, PostgreSQL, NoSQL (MongoDB, DynamoDB), ETL, Feature Engineering, PySpark
MLOps & Deployment: MLflow, Weights & Biases (W&B), Docker, Kubernetes, REST APIs, FastAPI, CI/CD Model Serving
Cloud & Platforms: AWS (SageMaker, Lambda, EC2, S3), Azure ML
Data Visualization: Power BI, Tableau, Plotly, Seaborn, Matplotlib, Advanced Excel, Interactive Dashboards

EXPERIENCE

CVS Health | Data Scientist — Jun2024 – Present | USA

- Built patient risk stratification models using Python, CatBoost, Pandas, and SQL on claims, pharmacy, and clinical datasets, improving high-risk case handling and supporting timely care interventions for 15,000+ members annually, achieving 85% accuracy.
- Developed medication adherence prediction system using ensemble learning, temporal feature engineering, and semantic search with FAISS and vector embeddings, enabling care teams to act on 2,100 non-adherent patients monthly, increasing adherence interventions 25%.
- Designed end-to-end healthcare analytics pipelines on AWS (S3, Glue, Athena, Redshift), processing large-scale clinical and claims datasets, reducing reporting turnaround from 7-8 hours to <2 hours, cutting operational costs $250K/year.
- Implemented NLP pipelines with Hugging Face Transformers, clinical BERT, and SpaCy to extract insights from physician notes, discharge summaries, and care plans, accelerating utilization review and case resolution workflows, reducing review time 40%.
- Established multimodal disease progression models combining LSTM for temporal data and CNNs for imaging/document data, supporting early-stage risk identification for diabetes and cardiovascular cohorts, achieving AUC 0.88–0.91 and enabling personalized care pathways.
- Created internal knowledge retrieval and semantic search platform using FAISS, Weaviate, embeddings, and Elasticsearch to query clinical policies and research documents, reducing clinician search time 2–3 hours per workflow.
- Deployed ML services with FastAPI, Docker, MLflow, and CI/CD, supporting internal care and analytics applications with automated monitoring, versioning, and low-latency inference, achieving 99% uptime and <200ms latency.

KPIT | Data Scientist — Jun 2020– Aug 2022 | India

- Engineered patient payment risk models using TensorFlow, Scikit-learn, Pandas, and NumPy on hospital billing and claims datasets, identifying 6,800 delayed accounts monthly and enabling timely finance team interventions.
- Constructed predictive payment adherence system with Gradient Boosting and CatBoost, leveraging semantic embeddings for document search, reducing follow-up call time by 1,200 hours per quarter.
- Orchestrated ETL and feature engineering pipelines using PySpark, Dask, and Airflow, processing 1.5 TB of transactional and demographic data monthly, reducing preprocessing time from 3 days to under 12 hours.
- Applied NLP pipelines with Spark NLP to extract insights from patient billing notes and insurance documents, automating document review and enabling finance teams to resolve 95% of flagged cases within 48 hours.
- Architected anomaly detection frameworks using Isolation Forest and Gradient Boosting to flag unusual billing trends, generating alerts that helped prevent 220 potential revenue losses per month.
- Formed patient segmentation models with K-Means clustering and PCA on hospital accounts, producing 12 actionable clusters to optimize billing outreach strategies.
- Crafted interpretable AI pipelines using SHAP and LIME to explain predictions for finance and care teams, increasing confidence in risk scoring.
- Productionized ML pipelines with TensorFlow, Databricks, and Kubernetes, supporting scalable inference and retraining across billing workflows.
- Devised interactive dashboards in Power BI, pulling data from Snowflake, visualizing billing trends, high-risk accounts, and workflow efficiency for finance managers, improving decision speed and accuracy.

ACADEMIC PROJECT

Generative AI & Clinical Decision Support Assistant

Skills Used: *Python, Transformers, Multimodal AI, NLP, GPT-4, LLaMA, BLIP-2, RAG, Semantic Search, Vector Databases, FAISS, Pinecone, LangChain, FastAPI, Docker, Power BI, Model Monitoring*

- Programmed multimodal clinical assistant (GPT-4, LLaMA, BLIP-2) for patient records, medical images, and clinical notes, delivering real-time diagnostic insights and reducing clinician review time by 45%.
- Operationalized RAG pipelines with LangChain and vector stores (FAISS, Pinecone) for accurate retrieval of medical history and guidelines.
- Configured APIs & dashboards using FastAPI, Docker, and Power BI for real-time monitoring of model performance and clinical risk signals.

EDUCATION

Master of Science in Computer Science | University of Memphis, Memphis, Tennessee — **May 2024**