Rakshith Roy Gantagogula

AI/ML Engineer

+1 (972) 730-4157 | rgantagogula@gmail.com | LinkedIn

SUMMARY

- AI/ML Engineer with 3+ years of experience delivering end-to-end machine learning solutions in finance and technology.
- Proficient in *Python*, TensorFlow, PyTorch, HuggingFace, and NLP for LLM platforms and GPT-powered automation.
- Skilled in MLOps (CI/CD, Docker, Kubernetes, Pulumi) for scalable, production-grade deployments.
- Experienced in cloud platforms (AWS: SageMaker, Bedrock, ECS, Lambda, Redshift; *Azure* ML) and data engineering (*SQL*, Spark, ETL pipelines).
- Strong background in predictive analytics, statistical modeling, and BI tools (*Tableau*, *Power BI*, AWS QuickSight) to optimize business performance.

SKILLS

Programming Languages: Python, TypeScript, JavaScript (ES6+), Java, C++, SQL

Machine Learning & AI: PyTorch, TensorFlow, Scikit-learn, XGBoost, HuggingFace, BERT, Pandas, NumPy, OpenAI GPT, Claude, LangChain, LangGraph, RAG, Onnx, DeepSpeed, ElevenLabs

MLOps & Deployment: Flask, RESTful APIs, Docker, Kubernetes, AWS ECS, Serverless (AWS Lambda), GitHub Actions (CI/CD), Pulumi

Full-Stack & API Development: Node.js, React.js, Next.js, NestJS

Cloud & Infrastructure: AWS (EC2, ECS, Lambda, SageMaker, Bedrock, CloudWatch), Azure (Functions, ML)

Data Engineering: PostgreSQL, DynamoDB, Apache Spark, ETL Pipelines, Kinesis, Redshift, Amazon RDS

Visualization & Monitoring: Tableau, Power BI, AWS QuickSight, Matplotlib, Seaborn

EXPERIENCE

AI/ML Engineer | CGI, USA **Jul 2024 – Present**

- Architected and deployed a distributed AI/ML platform for multi-billion parameter LLMs, enabling 100+ production AI agents across business units.
- Built a GPT-4/Claude-powered knowledge assistant that automated internal queries, reducing manual workload by ~3 hours/day per department and scaling GenAI adoption from 1K → 4.5K employees.
- Implemented an AI-driven code review agent integrated with *GitHub* PRs, cutting review cycle time by 30% and lowering post-deployment defects by 15%.
- Designed an ETL pipeline (DynamoDB → PostgreSQL) with a ReactJS dashboard, improving monitoring efficiency and reducing retrieval latency by 40%.
- Developed reusable *Python*/Node.js libraries for ML workflows, reducing cloud infrastructure costs by 20%.
- Automated infrastructure provisioning and CI/CD deployments with Pulumi, Docker, and *GitHub* Actions, ensuring consistent multi-environment rollouts.*GitHub* Actions, ens
- Enhanced observability by delivering CloudWatch and QuickSight dashboards for real-time AI/ML pipeline monitoring.

ML Engineer | Hexaware Technologies, India **Jan 2020 – Jul 2022**

- Designed and launched a churn prediction system using Logistic Regression, Decision Trees, and Gradient Boosting, improving customer retention by 25% within six months.
- Built a BERT-based sentiment analysis pipeline with 90% accuracy, including automated retraining for evolving customer data.
- Engineered advanced feature transformations on large-scale banking datasets (transactions, demographics), improving model robustness and accuracy.
- Developed RESTful APIs with Flask to integrate churn prediction models into CRM, enabling real-time retention alerts.
- Collaborated with marketing teams on A/B testing of targeted campaigns, reducing churn by 15% in intervention groups.
- Created real-time portfolio risk dashboards using Matplotlib, Seaborn, and AWS QuickSight, enabling proactive risk management.
- Containerized and deployed ML models on AWS EC2 with Docker and Git, ensuring secure and scalable production deployments.

EDUCATION

Masters in Information Technology and Management, The University of Texas at Dallas, Texas Aug 2022 – May 2024

CERTIFICATIONS

- AWS Certified Solution Associate
- HashiCorp Terraform Associate (003)
- Databricks Generative AI Fundamentals