NAGA SAI

AI Engineer/ Gen AI Engineer/ Data Scientist

7816025839
nagasai161141@gmail.com
www.linkedin.com/in/nagasai44

OBJECTIVE

Experienced Data Scientist and MLOps Specialist with 4.11 years of expertise in delivering scalable AI solutions and deploying ML models on AWS and Azure. Skilled in developing and fine-tuning Generative AI models and AI agents using LangChain and advanced GenAI frameworks. Proven ability to lead cross-functional teams to build innovative AI/ML-driven systems across various domains. Strong background in end-to-end data science workflows, cloud deployments, and CI/CD pipelines. Adept at driving business value through high-impact machine learning use cases and clear communication of data insights to stakeholders.

TECHNICAL SKILLS

- Generative AI: LLMs, Diffusion models, Multimodal models, LLaMA,AI Agents
- Programming & Tools: Python, SQL, JSON, Git, GitHub, Power BI, Tableau
- Machine Learning & Deep Learning: TensorFlow, PyTorch, scikit-learn
- Cloud Platforms: AWS (SageMaker, Bedrock, EC2, S3, Lambda, Glue, CloudFormation), Azure (AI Studio, Data Factory, Databricks, Cognitive service, Pyspark, Blobstorage)
- MLOps: Model monitoring, Automated Deployment, Model versioning, MLflow, GitOps, Airflow
- Frameworks: FastAPI, Flask, Django
- CI/CD Automation: Jenkins, Terraform, Docker, Kubernetes, ArgoCD

EXPERIENCE

Data Scientist- HCL Technologies(Vijayawada) **June 2021- Present**

- Designed and implemented multimodal Generative AI solutions by combining vision and text models to elevate customer experience.
- Built comprehensive machine learning pipelines leveraging AWS SageMaker and Azure Databricks for streamlined model development and deployment.
- Created a conversational chatbot using Retrieval-Augmented Generation (RAG), utilizing OpenAI embeddings and managing vector stores with ChromaDB, Pinecone, FAISS, and PEFT techniques like LoRA and QLoRA.
- Spearheaded the deployment of a text-to-image diffusion model for marketing campaigns, cutting down campaign creation time by 30%.
- Deployed custom Llama-based language models to automate text summarization and customer support workflows.
- Integrated LangChain to facilitate advanced chaining of language models, improving multi-turn interaction and contextual understanding across AI applications.
- Architected serverless AI solutions with AWS Lambda and API Gateway to ensure scalable and cost-effective deployments.
- Improved customer support efficiency by reducing response time by 40% through AI-powered automated summaries.
- Managed MLflow experiments to track model training, tuning, and deployment metrics, ensuring reproducibility and model version control.
- Automated model lifecycle management using MLflow, enhancing continuous integration and delivery of machine learning models in production.

Data Scientist- Accenture(Mumbai) **Dec 2018- June 2019**

- Deployed and monitored large-scale machine learning systems on Azure infrastructure, ensuring high availability and scalability.
- Implemented continuous integration/continuous deployment (CI/CD) pipelines for machine learning workflows using Jenkins and GitLab.
- Leveraged Terraform for infrastructure automation, managing Azure resources such as Azure Virtual Machines, Blob Storage, and Azure Functions.
- Built and automated CI/CD pipelines for machine learning models using Azure DevOps Pipelines and Azure Machine Learning.
- Integrated Docker for containerized model deployments to enable consistent environment replication across development, staging, and production.
- Automated model versioning and rollback processes using Azure Machine Learning services and Azure Functions.
- Conducted continuous model monitoring and logging with Azure Monitor and Application Insights to ensure sustained model performance.
- Implemented data preprocessing, feature engineering, and model selection using Python (Pandas, Scikit-learn).
- Applied various Artificial Intelligence (AI)/machine learning algorithms and statistical modeling techniques including decision trees, text analytics, natural language processing (NLP), supervised and unsupervised learning, and regression models.
- Built predictive models to forecast sales trends, achieving a 95% accuracy rate.

Project 1: Generative AI for Text Summarization

- Implemented NLP models using Hugging Face Transformers and Azure AI Studio.
- Deployed models using AWS Lambda and Azure Data Factory (ADF) workflows.

Project 2: MLOps Workflow Automation with AWS

- Built CI/CD pipelines for ML models using Jenkins and Docker.
- Deployed Utilized AWS SageMaker and Azure Databricks for model training and deployment.

Project 3: Machine Learning for Customer Churn Prediction

- Engineered features from customer interaction data.
- Trained and fine-tuned classification models using Scikit-learn and XGBoost.

EDUCATION

Jawaharlal Nehru Technological University (JNTU), Kakinada , India **Sept 2014- Apr 2018**
Bachelor of Technology (B.Tech) – Electronics and Communication Engineering

CERTIFICATIONS

- Microsoft Azure AI Fundamentals – AI900
- Oracle cloud Infrastructure 2024 AI Foundations Associate
- Microsoft Azure Data Engineer Associate – DP 203