Sr AI/ML ENGINEER

ADITYA GUNTUPALLI

LinkedIn | 530-405-9360 | aditya.ai6676@gmail.com | GitHub

Professional Summary

- With **9+ years of experience as an AI/ML engineer**, developed and deployed **end-to-end AI/ML solutions**, leveraging frameworks like **TensorFlow, PyTorch, and Scikit-learn**, resulted in a 35% improvement in process automation and enhanced decision-making capabilities across diverse industries.
- Implemented end-to-end **fine-tuning** and **Agentic RAG of large language models (LLMs)** using **Transformers** and **VectorDB**, optimized model performance with custom enterprise datasets.
- Implemented **Model Context Protocol (MCP) servers** to standardize context-sharing between LLMs, APIs, and external data sources, enabling seamless orchestration of agent workflows across AWS services, Snowflake data, and domain-specific APIs.
- **Built production-ready Agentic AI systems on AWS** leveraging **Agent Core**, **Bedrock Agents,** Lambda, and DynamoDB, orchestrating multiple specialized sub-agents with **Guardrails** to deliver real-time insights with **Text2Sql** like tools.
- Extensive experience in cloud environments including **AWS, Azure, GCP and Palantir**, leveraging services such as **AWS SageMaker, AWS Bedrock, Vertex AI, AutoML, Azure AI & ML,** EC2, Lambda, BigQuery, and S3 for scalable training, deployment, and serving of machine learning models.
- Developed advanced **NLP** solutions leveraging **transformers, embeddings**, and **semantic similarity** techniques, resulted in 35% improvement in model accuracy and inference speed.
- Developed and fine-tuned **computer vision models using TensorFlow** for image classification achieving 95% accuracy.
- Leveraged **Claude Code CLI** as an agentic development accelerator across full-stack and AI/ML workflows, configuring custom **claude skills** for project context, MCP server integrations for AWS services, significantly reducing development cycle time for Lambda functions, Streamlit/React UIs, and multi-agent orchestration code.
- Developed and optimized deep learning models using **transfer learning** and **data augmentation**, achieving a **20% increase in model accuracy** and improved generalization on diverse datasets in **computer vision**.
- Implemented **data drift monitoring** and automated retraining pipelines using tools AWS CloudWatch, GCP Cloud monitoring and CloudTrail logs, triggered model retraining when drift thresholds were exceeded.
- Proficient in **evaluating RAG-based and GenAI systems** using frameworks such as **Ragas** and custom evaluation pipelines; applied precision, recall, F1-score, and AUC-ROC to ensure 95%+ accuracy and 90% precision across models deployed in production.
- Implemented **DevOps practices** to secure model pipelines following Data standards, using **IaC, encryption, and secret management** with tools like **Terraform**, **Docker** and **Kubernetes** for **finance, healthcare, and retail domains.**
- Built and deployed full-stack AI-powered applications using React JS with TailwindCSS, integrating FastAPI as backend with LLM-driven features, RESTful APIs, and real-time streaming interfaces.
- Built efficient **data ingestion** pipelines using **Airflow, PySpark, and DBT Cloud**, enabling processing of high-volume data from diverse sources.
- Developed machine learning algorithms such as **XGBoost, Random Forest, Support Vector Machines (SVM)**, Logistic regression and K-Means, utilizing techniques like **outlier detection**, **hyperparameter tuning**, **cross-validation**, and **feature engineering**, resulted in up to **15% improvement in model accuracy**.
- Conducted large-scale **model training** using **Vertex AI** and **SageMaker** across **GPU clusters**, using custom machine images to optimize training speed and scalability that **reduced training time by 20%**
- Proficient in MLOps tools like **MLflow, Kubeflow,** and TensorFlow Extended (TFX) for tracking experiments, managing model versions, and automating the entire ML lifecycle. Achieved a decrease in deployment **failures from 10% to just 3%**, ensuring reliable model performance in production.
- Expert in **Python**, leveraging **logging**, **error handling**, and tools like **Jupyter Notebook, Streamlit, FastAPI, and Flask** to prototype and deploy ML models efficiently. Containerized machine learning microservices using **Docker** with optimized dependencies and **CUDA support**, ensuring consistency in development and deployment across teams.

Skills

Programming Languages: Python, R, ReactJS, HTML, Shell Scripting, YAML.

Machine learning / Deep Learning: Supervised/Unsupervised Algorithms, ANN, CNN, Natural language processing(NLP), Computer Vision, GAN's, LSTM, Feature Engineering, Transfer Learning, ensemble models, Time series, Recommendation systems, Sentiment Analysis, Evaluation Metrics.

Frameworks: Tensorflow, Keras, PyTorch, Scikit-Learn, openCV, NLTK, Pandas, Transformers, Flask, FastAPI, Streamlit, Gradio, Pickle, Pydantic, ONXX, JAX, Anaconda, Jupyter Notebook, Boto3.

Generative AI: Fine-tuning LLM's, Langchain, LangGraph, Llama index, RAG, AI Agents, Model Context Protocol(MCP), A2A, OpenClaw, CrewAI, Auto Gen, Prompt Engineering, Hugging face, Phi, OpenAI, Pinecone, Claude Code, Cursor

Statistics & Metrics: Statistical Analysis, Regression and Time series Analysis, Confusion Matrix, Matplotlib, SciPy, Probability Theory

MLOps Tools: Git, Github Actions, DVC, MLFlow, Data Build Tool, DagsHub, Kubeflow, Terraform, Ansible, CI/CD, Docker, Tensorflow serving, jenkins, CircleCI.

Cloud:

- **AWS:** Agent Core, Sagemaker, Bedrock, Lambda, Neputune, GNN, Lex, S3, EC2, IAM, Cloud watch, AMI's, Redshift ML, DynamoDB, Redis, Code Build, Code Deploy.
- **GCP:** Vertex AI, AutoML, cloud vision API, Dialogflow, NVIDIA GPUs, BigQuery ML, VM instance, VPC.
- **Azure:** Azure Machine learning, Azure AI & OpenAI, Blob storage, Azure Functions, Azure cognitive services.
- **Palantir:** Foundry, Ontology.

Databases: MySQL, PostgreSQL, MongoDB, Neo4j Graph DB, Vector DB, SupaBase.

Monitoring & Logging: MLFlow, Weights & Biases, Tensorboard, Prometheus, Grafana, Evidently AI

Certification

AWS Certified Machine Learning - Specialty - Link

Experience

Lead GenAI Engineer **BPX Energy** *Denver,CO* **02/2025- Present**

- Designed and implemented Agentic **Retrieval-Augmented Generation (RAG) pipelines on AWS** by integrating Amazon Bedrock LLMs with Amazon OpenSearch Service and a Snowflake database schema containing oilfield equipment data, reducing **query resolution time by 30%** and improving decision-making speed for field operators.
- Developed **custom Model Context Protocol (MCP) integrations,** allowing agents to securely access structured/unstructured data and system health metrics, which reduced duplication of logic, improved modularity, and accelerated onboarding of new agent capabilities by 40%, **lowering time-to-deployment** for new features.
- Integrated **AWS Neptune** with **AWS Lambda and API Gateway** to expose graph-based insights via REST APIs, enabling real-time relationship queries in scalable, serverless architectures.
- Deployed production AI agents on **Amazon Bedrock AgentCore** Runtime with complete session isolation and MCP-compatible tool integration via AgentCore Gateway, leveraging **AgentCore Memory** for episodic context retention, AgentCore Identity for secure OAuth-based access to AWS services and third-party tools, and AgentCore Observability with OpenTelemetry and CloudWatch dashboards to monitor token usage, latency, and goal success rates across multi-agent workflows.
- Developed a **Multi-Agentic architecture on AWS** featuring a central orchestrator agent that intelligently

coordinates multiple specialized sub-agents. The **Orchestrator Agent** dynamically routes user queries to the appropriate sub-agent, ensuring optimized workflows, **reducing latency by ~35%**, and increasing response accuracy by 25% in domain-specific scenarios.

- Developed a **Text-to-SQL Tool** integrated with the Well Agent, enabling natural language queries to be converted into optimized SQL statements. The Lambda securely connects to Snowflake to retrieve well operation and telemetry data, streamlining field engineers' access to real-time insights and **reducing manual query writing effort by 60%.**
- Designed a routing agent using the AWS multi-agent framework to generate optimal service routes for field technicians dynamically. The system utilized real-time location data, equipment status, and priority metrics which **reduced travel time by ~20%** and enhanced overall operational efficiency, saving an estimated 15+ technician hours per week.
- Integrated **Amazon API Gateway** with configured methods to securely activate the main orchestrator agent, enabling seamless integration of **multi-agent workflows** from external applications and user interfaces.
- **Reduced operator report preparation time** by deploying our agentic AI application, leading to a measurable **10% boost in production efficiency** through streamlined data retrieval, analysis, and reporting workflows.
- Engineered agentic AI workflows leveraging **AWS Nova Sonic conversational models**, reducing end-to-end multi-agent orchestration latency by ~40% and enabling seamless real-time dialogue experiences.
- Adopted and implemented **Machine Learning and Generative AI use cases within the Palantir Foundry** platform, enabling data-driven decision-making, automated insights generation, and intelligent workflows across enterprise operations.
- Developed **Snowflake-based data products** and implemented automated data pipeline refresh schedules using dbt, ensuring timely, accurate, and analytics-ready datasets for downstream AI/ML workflows.
- Designed a **health monitoring dashboard** by integrating **Amazon CloudWatch logs and alarms**, configured to automatically trigger alerts whenever a Lambda function or system component experiences failures, enabling proactive incident response and minimizing downtime.
- Deployed agentic AI application code using **Azure DevOps (ADO) CI/CD pipelines**, automating build, testing, and deployment to ensure fast, reliable, and consistent releases across AWS environments.
- Conducted **A/B testing across multiple Claude models** and integrated **Claude 3.5 Sonnet**, achieving significant improvements in precision and recall for domain-specific query handling and response generation.

AI/ML/GenAI Engineer **OPTUM** *Remote* **01/2023- 12/2024**

- Developed and **fine-tuned deep learning Computer vision models** using **TensorFlow and PyTorch**, implemented advanced architectures such as CNNs and Transformers and **achieved high performance in image classification** and natural language processing tasks.
- Conducted model evaluations utilizing precision, recall, and F1-score, leading to a **30% reduction in false positives**, thereby significantly enhanced model reliability for accurate predictions.
- Implemented **hyperparameter tuning techniques, Data Augmentation** and optimized deep learning models, resulted in a 20% increase in model accuracy and improved generalization to unseen data.
- Utilized **transfer learning from pre-trained CV models** in TensorFlow to develop an efficient image classification pipeline for tumor detection in healthcare.
- Built and deployed custom LLM solution using **HuggingFace's Llama-3.2** model and fine-tuned on 50K domain-specific examples, integrated with **LangChain for RAG** capabilities, resulted in 50% reduction in operational costs versus commercial APIs.
- Implemented a **vector database** for knowledge base of enterprise data, utilizing **cosine similarity** to efficiently retrieve and rank contextually relevant documents based on user queries.
- Created detailed **model performance dashboards** using **MLflow** and visualization tools, allowing stakeholders to monitor key metrics over time and providing insights into **model drift** and performance.
- **Improved model accuracy from 82% to 94%** by implementing **transfer learning** techniques and fine-tuning a pre-trained **ResNet50** architecture on a custom dataset of 50,000 images.
- Managed **EC2 Nvidia GPU instances** (such as P3 and G4) for deep learning model training, optimized resource allocation and reduced training time by 35% through efficient use of **NVIDIA CUDA** and distributed training techniques.
- Optimized deep learning pipelines for **online and batch predictions**, ensured **SLA requirements** by

achieving real-time inference with response latency under 200ms for online predictions and completing batch predictions within the 20-minute SLA.

AI/ML Engineer **First Citizens Bank** *Remote* **01/2022- 11/2022**

- Designed and executed scalable machine learning **model training workflows using Amazon SageMaker**, optimized training performance through distributed processing.
- managed a centralized **Feature Store in Amazon SageMaker**, enabling real-time feature retrieval and reuse across ML models ensuring consistent feature availability with 99.9% uptime.
- Experimented with multiple machine learning algorithms, including XGBoost, Random Forest, and Logistic Regression, to evaluate performance on fraud detection and identified **XGBoost** as the most effective, **achieving a 95% AUC and a 20% reduction in false positives**.
- Optimized model training workflows by leveraging **Amazon EC2 Spot Instances** for high-performance compute tasks, **reduced training costs by 40%** while maintaining model accuracy and scalability.
- Utilized **AWS Lambda to automate preprocessing** of incoming data for machine learning models, enabling real-time feature extraction and reduced data pipeline latency by 30%.
- Implemented feature engineering techniques such as **Label encoding, feature scaling**, and feature transformation to create **optimized training datasets.**
- Implemented **outlier detection** techniques such as **Z-Score and IQR** methods to improve data quality, and applied **dimensionality reduction using PCA** to reduce feature space by 40%, enhancing model efficiency and interpretability.
- Implemented Amazon SageMaker **Model Monitor to detect data drift** and model performance degradation, leveraging evaluation metrics such as accuracy, F1-score, and feature distribution analysis, ensuring a 99% model reliability in production.
- Utilized Amazon **SageMaker Experiments to track**, organize, and compare machine learning training runs, enabling efficient hyperparameter tuning and achieving a 15% improvement in model accuracy by identifying optimal configurations.
- **Containerized** machine learning workflows using **Docker**, ensuring seamless integration of dependencies, libraries, and runtime environments, enabling consistent deployment across development, testing, and production.

Machine Learning Engineer **L L Bean** *Remote* **12/2019- 11/2021**

- Developed and deployed machine learning models to predict customer churn which enabled proactive retention strategies that **reduced churn rates by 15%**.
- Utilized **Azure Blob Storage to process large-scale retail customer data** and performed advanced feature engineering to optimize datasets for machine learning model development.
- Implemented **hyperparameter optimization using Azure Machine Learning**'s built-in HyperDrive, which improved the F1 score of the classification model by 12%.
- Used Amazon **SageMaker notebooks** for scalable model development and experimentation, streamlining data preprocessing, training, and evaluation workflows.
- Developed a **recommendation system using collaborative filtering** and content-based algorithms, which increased customer engagement and boosted **product sales by 25%.**
- Used **confusion matrix and classification metrics** to evaluate model performance and identified key areas of misclassification and improved predictive accuracy.
- Integrated **Azure Cognitive Services** to implement natural language processing for **sentiment analysis**, which improved customer feedback classification accuracy by 20% and enhanced decision-making for targeted marketing strategies.
- Pushed code to **Git repositories** using Git Bash, including **model dependencies and pickle files** enabling efficient version control and streamlined deployment workflows.
- Developed and maintained **YAML configuration files** to streamline workflow automation which improved environment consistency and enabled better deployment and pipeline management.
- Worked with Data Scientists and developed predictive models and analytics solutions using Python, contributing to data driven decision-making processes.

Machine Learning Engineer **Aspire Systems** *Hyderabad(India)* **11/2016- 10/2019**

- Performed **data cleaning, normalization, and feature engineering** on large datasets using Python, SQL, and PySpark to enhance model performance and scalability.
- Optimized model performance through **hyperparameter tuning, cross-validation**, and feature selection to achieve target metrics such as reduced latency and higher accuracy.
- Created comprehensive documentation for data workflows, model architectures, and deployment pipelines to facilitate team knowledge sharing and reproducibility.
- Implemented Continuous integration and delivery (CI/CD) pipelines using **Git, Docker and Jenkins** to streamline model deployment.
- Effectively communicated AI/ML techniques and their business impact to non-technical stakeholders, enabling data-driven decision-making and alignment across teams.
- Implemented and compared various regression algorithms by training models using **train-test-split**, optimizing for accuracy and predictive performance.

Python Developer **Parallel wireless** *Banglore(India)* **06/2015- 11/2016**

- Automated repetitive tasks and processes using **Python scripting**, saving 10+ hours of manual effort per week and improved operational efficiency.
- Integrated RESTful APIs into Python applications, streamlining third-party data access and ensuring secure communication between services.
- Developed and executed **unit tests using Pytest**, ensuring code quality and reducing production bugs by 25%.
- Developed and deployed **RESTful web services using Flask**, enabling seamless integration with front-end applications and improved API response times.
- Designed and implemented Python scripts to process and analyze **large-scale telecom data**, which improved network performance monitoring and reduced downtime by identifying anomalies in real-time.
- Developed Python scripts to interact with **SQL databases**, performing data extraction, transformation, and loading (ETL) operations, which improved query efficiency and streamlined data analysis workflows.

Projects

Generative AI-Powered Contextual Q&A RAG System for Research Papers: LINK

- Integrated LangChain with **Groq and Llama3** models to build a Generative AI-powered Retrieval Augmented Generation(RAG), enabling accurate retrieval and response generation based on research paper content.
- Leveraged **FIASS VectorDB** for vector embeddings to create a knowledge base from research documents, enhancing the Q&A system's ability to provide precise, context-aware answers based on **co-sine similarity** and implemented **prompting techniques** like zero shot and few shot learning using prompt templates.

Interests

Enthusiastic about learning and exploring innovative concepts in data science and AI, while also keeping up with new advancements through reading research papers.
Actively engaged in Conference, webinars, boot camps, Kaggle competitions, and enjoy listening to Lex podcasts.

Education

Bachelors in Computer Science - Jntuk(AndhraPradesh) - 2011 to 2015