ANUSHA REDDY

LinkedIn **551-804-2870** **Anu2710r@gmail.com**

SUMAMRY

Accomplished Machine Learning and Generative AI Engineer with over 5 years of experience in developing and deploying end-to-end AI solutions across diverse domains. Proven expertise in building robust predictive models, from a high-accuracy pneumonia detection CNN (92% accuracy) to an NLP-powered recommendation engine that boosted user engagement by 15%. A seasoned Generative AI Engineer adept at designing and implementing Retrieval-Augmented Generation (RAG) architectures, successfully delivering a production-ready intranet chatbot for Microsoft leveraging Azure OpenAI (GPT-4o) and Azure Cognitive Search that achieved 95% response accuracy. Skilled in engineering, scalable MLOps pipelines on both GCP and Azure, including a real-time data drift monitoring and automated retraining system for Ford Connected Cars, and a production outage prediction model for Unilever. Proficient in a wide array of technologies, including Python, TensorFlow, Keras, PyTorch, BERT, PySpark, Azure Databricks, and Lang Chain, with a focus on delivering innovative and impactful solutions.

WORK EXPERIENCE

SPR Software systems **DEC 2023-JUNE 2025**

Generative AI Engineer

Project: Enterprise Intranet Chatbot for Microsoft powered by RAG & GenAI

- Developed comprehensive RAG system integrating Azure OpenAI GPT-4o with Azure Cognitive Search, achieving 95% response accuracy for enterprise knowledge queries
- Developed automated data ingestion pipeline processing 100+ SharePoint documents with real-time embedding generation and semantic chunking using Azure Function Apps
- Built production-ready GenAI solution handling 1000+ daily interactions with sub-2-second response times and multi-agent conversational AI capabilities
- Implemented enterprise AI governance framework ensuring GDPR compliance, content filtering, bias detection, and response validation for sensitive organizational data
- Conducted advanced prompt engineering, hyperparameter tuning, and retrieval optimization including semantic search, hybrid search, and re-ranking algorithms
- Established MLOps pipelines for GenAI model deployment, monitoring, and lifecycle management using Azure DevOps and MLflow for continuous improvement.

Tech Stack: Azure OpenAI (GPT-4o), RAG Architecture, LangChain, Azure Function Apps, Azure Cognitive Search, Azure ML, SharePoint API, Document Embeddings, Vector Databases, Prompt Engineering, Multi-Agent Systems, LLM Fine-tuning, Azure DevOps, MLflow, Git.

IBM **JULY 2021-AUG2023**

Machine learning Engineer

Project 1: Data Drift Monitoring & Automated Retraining Pipeline

- Engineered real-time data drift detection pipeline on Google Cloud Platform monitoring predictive model stability for connected automotive sensor data streams
- Configured automated retraining system triggered by predefined drift thresholds, ensuring continuous model accuracy and performance stability in production environment
- Analyzed drift patterns using statistical hypothesis testing and machine learning techniques to identify root causes and inform adaptive feature engineering strategies
- Implemented scalable drift monitoring pipeline using GCS buckets, Cloud Functions, and Big Query for high-volume automotive data processing and real-time alert generation.

Tech Stack: Google Cloud Platform (GCP), Custom Data Drift Algorithms, MLOps, Automated Retraining, Python, Machine Learning Libraries, GCS Buckets, Cloud Functions, Statistical Analysis, Git

Project 2: Production Outage Prediction Model

- Developed regression models using XGBoost and ensemble techniques to predict manufacturing production outages based on operational factors and historical maintenance data
- Processed large-scale manufacturing datasets using Azure Databricks and PySpark for scalable data transformation, feature engineering, and model training workflows
- Implemented comprehensive data governance framework using Unity Catalog ensuring data quality, lineage tracking, and compliance for sensitive operational datasets
- Trained and managed complete model lifecycle in Azure ML Studio including hyperparameter tuning, cross-validation, and automated model versioning for production deployment
- Deployed predictive model as REST API web application endpoint enabling real-time outage predictions and proactive maintenance scheduling for manufacturing operations.

Tech Stack: Azure Databricks, Azure ML Studio, Unity Catalog, PySpark, XGBoost, Regression Models, Web App Deployment, Python, REST API, Feature Engineering, Hyperparameter Tuning, Git.

RMSI **AUG 2020-JULY 2021**

Software Engineer

Project: Full-Stack Machine Learning & MLOps Pipeline

- Engineered an end-to-end machine learning solution covering the entire lifecycle from data analysis to production monitoring.
- Performed extensive Exploratory Data Analysis (EDA) on datasets exceeding 50 records and implemented text normalization to improve data quality.
- Developed a high-performance predictive model by fine-tuning BERT to achieve a 95% accuracy, resulting in a 20% reduction in user search time and a 15% increase in click-through rates (CTR).
- Established robust MLOps practices by creating an automated deployment pipeline on Azure ML, including model versioning and a continuous monitoring system for data drift detection.
- Ensured model reliability through cross-validation and optimized performance with hyperparameter tuning using techniques like Grid Search.

Tech Stack: Python, Azure ML, BERT, PyTorch, Scikit-learn, Pandas, NumPy, MLOps, Data Analysis, Data Preprocessing, Hyperparameter Tuning, Monitoring.

SMARTBRIDGE **APRIL 2018-AUG 2018**

- Developed a Convolutional Neural Network (CNN) using TensorFlow/Keras to classify chest X-ray images for pneumonia detection, achieving 92% validation accuracy.
- Preprocessed and augmented medical image data (resizing, normalization, rotation, flipping) to improve model generalization and reduce overfitting.
- Utilized transfer learning with pretrained models like VGG16 and ResNet50 to boost model performance and reduce training time.
- Performed model evaluation using metrics such as accuracy, precision, recall, F1-score, and confusion matrix, ensuring clinical relevance and reliability.

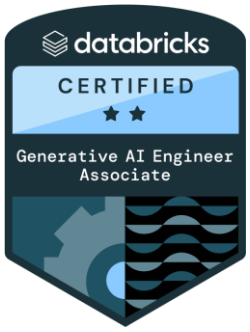
Tech Stack: TensorFlow, Keres, scikit-learn, CNN, Python

TECHNICAL SKILLS

- **Programming Languages:** Java, Python, PySpark.
- **Generative AI & Large Language Models:** OpenAI GPT Models, ChatGPT, OpenAI API, DALL-E, Retrieval Augmented Generation (RAG), Lang Chain, LangGraph, Lang Smith, Llama Index, Prompt Engineering, AI Agents, Multi-Agent Systems, Autonomous AI Agents, MCP Agents, Hugging Face Transformers.
- **Machine Learning & Deep Learning:** Machine Learning, Deep Learning, Natural Language Processing (NLP), Computer Vision, Image Processing, Text Classification, Text Generation, Classification, Regression, Recommendation Systems, XGBoost, Scikit-learn, Keras, TensorFlow.
- **Cloud AI/ML Platforms** Bedrock, AWS SageMaker, Google Vertex AI, Azure OpenAI, Azure Machine Learning, Google Cloud AI Platform, MLflow
- **Data Science Libraries & Frameworks:** NumPy, Pandas, Matplotlib, Seaborn, Fast API, Flask, Hugging Face, TensorFlow, Keras.
- **Cloud Technologies:** Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure.
- **Databases & Data Warehousing:** Amazon Redshift, Snowflake, PostgreSQL, MySQL.
- **DevOps & Version Control:** Git, GitHub, GitLab, Bitbucket, Azure DevOps, Docker, Kubernetes, CI/CD Pipelines, SVN.
- **Operating Systems:** Windows, Linux, UNIX, macOS.
- **Specialized AI/ML Applications:** Chatbot Development, Conversational AI, Text Mining, Sentiment Analysis, Predictive Modeling, Neural Networks, Feature Engineering, Model Deployment, Model Monitoring.

CERTIFICATIONS

DATABRICKS GENERATIVE AI ENGINEER ASSOCIATE
https://credentials.databricks.com/51b62d5e-6e08-493a-9d20-8e4180fc167a



OCI GENERATIVE AI PROFESSIONAL
https://catalog-education.oracle.com/pls/certview/sharebadge?id=4DC19148F918FCF88283909A5C951C205BA73B0C77EA5FFDA209777A036987DD