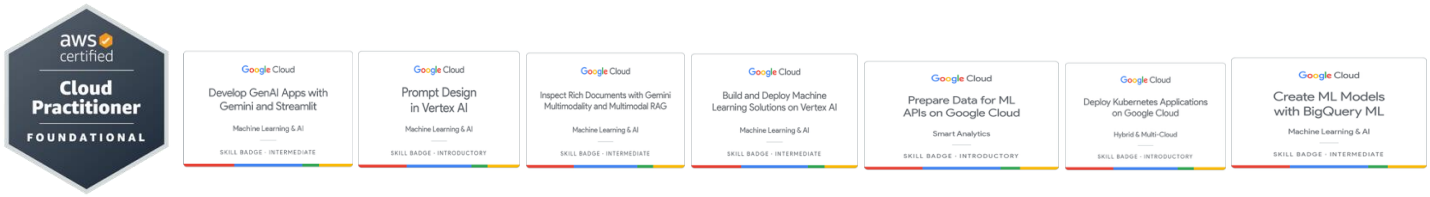
ANOOP REDDY YEDDULA

AI/ML Engineer | anoopy3001@gmail.com | +1-917-624-1889



SUMMARY

- Results-driven IT professional with 5 years of experience in developing and deploying data science, machine learning and NLP solutions using Python and PySpark.
- Expertise in the entire machine learning model development lifecycle, from data collection and preprocessing to model deployment and monitoring.
- Proficient in developing and implementing Generative AI solutions, particularly with Mistral, Llama, RAG and GPT technologies, as well as vector databases like Pinecone and Chroma.
- Experienced in building and deploying REST APIs using Azure API Gateway for seamless integration with existing systems and applications.
- Skilled in optimizing model performance using GPU-based processing and load balancing techniques on cloud platforms like Azure.
- Familiar with DevOps practices, CI/CD pipelines and automation tools in AI/ML contexts, ensuring efficient and reliable model deployment.
- Strong communication and collaboration skills, with a proven ability to work effectively with cross-functional teams and stakeholders to deliver high-impact solutions, including experience in high impact client communication.
- Certified in AWS Cloud Practitioner, Prompt Design and ML in Vertex AI, Develop GenAI apps and Gemini Multimodal Development, with a passion for solving complex problems through intelligent, explainable, and scalable systems.

Technical Skills

- **Programming Languages & Frameworks:** Python, Java, C/C++, SQL, R, React.js, HTML/CSS, JavaScript, Node.js, FastAPI, Flask, Langchain, LangGraph, LlamaIndex, CrewAI, AutoGen, TensorFlow, PyTorch, Dialogflow CX
- **Data Libraries & Scientific Computing:** Pandas, NumPy, SciPy, Plotly, Scikit-learn, Keras
- **AI/ML & NLP Techniques:** Supervised & Unsupervised Learning, Deep Learning, Anomaly Detection, Time Series Forecasting, Clustering, Regression, Classification, LLM Prompt Engineering, Text Classification, RAG, Knowledge Graphs, Named Entity Recognition, Sentiment Analysis
- **Agentic AI & Orchestration:** Agent Design & Implementation, Multi-Agent Systems, API Integration, Autonomous Workflow Automation using Langchain, LangGraph and CrewAI
- **Azure & DevOps:** ML Studio, Synapse, Data Factory, API Gateway, Azure Functions, Azure DevOps, Docker, Kubernetes, GitHub Actions, Jenkins, Terraform
- **Model Deployment & MLOps:** CI/CD Pipelines, Automated Model Monitoring, Real-Time Inference, GPU-based Optimization, Model Retraining, Azure ML Pipelines
- **Databases & Data Engineering:** PostgreSQL, MongoDB, SQL Server, NoSQL, Azure DataBricks, Azure Synapse Analytics, Data Warehousing, ETL Pipelines, Vector Search
- **Testing & Automation:** PyTest, Unit Testing, Model Evaluation Metrics, YAML Scripting, Shell/PowerShell Automation
- **Cloud Platforms:** AWS, GCP (Vertex AI, BigQuery, Cloud Storage, IAM)
- **Visualization & Reporting:** Tableau, Matplotlib, Seaborn, Streamlit Dashboards

PROFESSIONAL EXPERIENCE

AI Engineer at Ross Stores, Dublin CA, USA March 2024 – Present

- Developed an AI-powered chatbot using Rasa and Python to handle customer inquiries on products and services, enhancing service and reducing costs.
- Applied NLP techniques to accurately understand customer intent and deliver relevant, real-time responses.
- Integrated the chatbot with messaging platforms like WhatsApp and Facebook Messenger to extend accessibility.
- Implemented a Chroma-based knowledge base for efficient retrieval of banking-related information.
- Trained the chatbot on a large dataset of customer conversations and feedback, significantly improving accuracy and engagement.
- Achieved a 60% reduction in customer service response time and a 20% decrease in operational costs.
- Used analytics tools to monitor chatbot performance and identify areas for ongoing optimization.
- Documented chatbot architecture, functionalities, and usage guidelines to support easy adoption by other development teams.
- Designed and deployed an AI agent using Langchain to optimize supply chain operations, reduce costs, and improve efficiency.
- Enabled the agent to analyze shipping data and propose optimized routes and delivery methods.
- Integrated machine learning models to predict demand, optimize inventory, and improve transportation logistics.
- Reduced supply chain costs by 10% and improved on-time delivery rates by 15% through intelligent automation and optimization.

AI Engineer at Chevron, Houston TX, USA June 2023 - Feb 2024

- Developed a real-time fraud detection system for financial transactions using deep learning (TensorFlow, Keras), leveraging transaction history, user behavior, and device data.
- Improved model accuracy by 15% through advanced feature engineering and anomaly detection algorithms to flag suspicious transactions.
- Integrated Apache Kafka for real-time streaming of transaction data, enabling low-latency fraud detection and decoupled architecture.
- Implemented model monitoring tools to detect performance drift and ensure long-term system effectiveness.
- Designed an automated alert system that reduced fraud investigation response time by 50%.
- Deployed the fraud detection system on scalable cloud infrastructure using Kubernetes for high availability and performance.
- Built a credit risk assessment model with Python and PySpark to predict creditworthiness from historical data and macroeconomic indicators.
- Applied NLP techniques to analyze news and social media sentiment, identifying emerging risks impacting borrower profiles.
- Implemented a RAG system using Llama 2 and Pinecone to deliver real-time borrower insights, reducing manual review time by 40%.
- Developed REST APIs using Azure API Gateway to integrate real-time credit scoring with banking systems.
- Optimized inference performance with GPU-based processing and Azure-based load balancing, reducing prediction latency by 25%.
- Automated model deployment and monitoring via CI/CD pipelines in Azure DevOps, ensuring continuous delivery and reliability.

Data Scientist at Hudl, India Oct 2019 – Sept 2022

- Collaborated with a cross-functional team of data scientists, analysts, and business stakeholders to develop and implement data engineering solutions for advanced analytics.
- Implemented automation of well parameters data flow for over 70 workflows and 15k+ unique PID equipment and readings, by effectively managing a team of 4 contractors on PI OSIsoft framework with SQL.
- Analyzed data using hierarchies and divisions maintained over SAP and Excel to deliver meticulous and actionable insights via automated decks, reports, Tableau and PowerBI dashboards.
- Led data acquisition efforts, including data extraction, data cleaning, and data transformation to prepare raw data for analysis.
- Developed and maintained ETL processes using Python and SQL to efficiently process and integrate large datasets

from various sources.

- Implemented data quality checks and data validation processes to ensure data accuracy and integrity throughout the data pipeline.
- Optimized SQL queries for data extraction, transformation, and loading (ETL) processes to improve data processing efficiency and reduce processing time.
- Implemented data modelling and data warehousing techniques to support advanced analytics and reporting requirements.
- Developed and maintained data pipelines and workflows using distributed computing technologies such as Hadoop and Spark for processing large-scale data.
- Collaborated with the data visualization team to provide clean and structured data for creating visually powerful and actionable interactive reports and dashboards.
- Participated in data governance initiatives to ensure compliance with data privacy and security regulations.
- Worked with version control tools such as GitHub and Azure for managing codebase and ensuring code quality.
- Assisted in designing and implementing machine learning models for predictive analytics, including feature engineering, model evaluation and real-time KPI monitoring.

CERTIFICATIONS

- AWS CERTIFIED CLOUD PRACTITIONER
- Prompt Design in Vertex AI
- Inspect Rich Documents with Gemini Multimodality and Multimodal RAG
- Develop GenAI Apps with Gemini and Streamlit
- Build and Deploy Machine Learning Solutions on Vertex AI

EDUCATION

- **Master's in computer science** at Portland State University, Portland, Oregon, USA
- **Bachelor's in computer science & engineering** at Dayananda Sagar University

PROJECTS

Customer Churn Prediction using LLM and ML:

Problem Statement: Predict and explain potential customer churn using a hybrid of machine learning models and large language models (LLMs) for actionable retention strategies.
Responsibilities:
• Trained a churn prediction model using Scikit-learn and the BankChurners dataset, improving prediction accuracy by 30%.
• Integrated Hugging Face's Flan-T5 LLM with LangChain to generate natural language explanations for predicted churn cases.
• Deployed a full-stack churn prediction pipeline to provide real-time business insights and retention recommendations.
Challenges: Combining explainability with prediction accuracy and integrating traditional ML with LLM-based workflows.
Tools Used: Python, Scikit-learn, Hugging Face, LangChain, Flan-T5, Pandas, NumPy

Automated Insurance Claim Validation System:

Problem Statement: Validate insurance claims efficiently by automating text extraction and document classification across varied formats.
Responsibilities:
• Extracted structured data using EasyOCR from image-based documents and multi-page PDFs.
• Applied BERT-based classification for document validation, improving accuracy by 35% and reducing manual review time by 40%.
• Built an intuitive Gradio interface for real-time claim review with support for multiple file formats.
Challenges: Handling OCR noise from scanned documents and ensuring high classification accuracy across diverse formats.
Tools Used: Python, EasyOCR, Hugging Face BERT, Gradio, Pandas, OpenCV

Real-Time Inventory Advisor:

Problem Statement: Optimize inventory management by predicting product restocking needs based on sales trends and AI-generated forecasts.

Responsibilities:

• Engineered a predictive model using Scikit-learn to forecast inventory demands with 95% accuracy.
• Augmented predictive insights using Hugging Face Transformers to generate restocking advice using generative AI.
• Deployed a real-time advisory interface using Gradio, reducing stockouts by 25%.

Challenges: Integrating time series forecasting with generative responses while maintaining real-time responsiveness.

Tools Used: Python, Scikit-learn, Hugging Face Transformers, Gradio, Pandas

Credit Score Type Prediction:

Problem Statement: Classify customers into credit score categories based on real-time financial inputs.

Responsibilities:

• Engineered and trained a Random Forest model to classify customers as Poor, Average, or Good in terms of credit health.
• Developed a Streamlit application allowing users to input real-time data for instant predictions.
• Optimized feature selection and model parameters to maximize classification accuracy.

Challenges: Managing real-time user inputs and balancing prediction granularity across credit score types.

Tools Used: Python, Scikit-learn, Random Forest, Streamlit, Pandas, NumPy

Drug Effectiveness Prediction:

Problem Statement: Assist healthcare providers by predicting the effectiveness of drugs for specific medical conditions using clinical data.

Responsibilities:

• Used logistic regression with TF-IDF vectorization on drug-condition text pairs to achieve 88% accuracy.
• Built an interactive Gradio app for real-time input of drug names and conditions, delivering instant predictions.
• Enabled data-driven support for prescribing effective treatments.

Challenges: Representing unstructured medical input data effectively for classification models.

Tools Used: Python, Logistic Regression, TF-IDF, Gradio, Scikit-learn, Pandas

Decoding Facial Recognition using CNN:

Problem Statement: Develop a facial recognition system capable of high-accuracy identification using custom datasets.

Responsibilities:

• Designed CNN architecture in TensorFlow, achieving 95% accuracy on facial recognition tasks.
• Preprocessed image data and tuned training parameters for better generalization and model performance.
• Visualized metrics and conducted iterative evaluations to refine recognition results.

Challenges: Ensuring robustness across varying lighting conditions and facial angles.

Tools Used: Python, TensorFlow, Keras, OpenCV, NumPy, Matplotlib