SREE YUKTHA SUNKARA
Phone: 940-758-4601
Email: sunkarasreeyuktha99@gmail.com
LinkedIn: https://www.linkedin.com/in/sreeyukthasunkara/

PROFESSIONAL SUMMARY:

- AI/ML Engineer with 4+ years of experience in data analysis, building and deploying end-to-end ML pipelines, fine-tuning large language models (LLMs), and implementing deep learning models.
- Proficient in implementing MLOps pipelines covering the entire ML lifecycle: data preprocessing, model training, version control, experiment tracking, model deployment, and monitoring in production environments.
- Proficient in end-to-end model building processes including feature engineering, model selection, training, hyperparameter tuning, validation, and performance evaluation using tools and libraries such as Python, Scikit-learn, TensorFlow, PyTorch, and Optuna. Experienced with a wide range of algorithms—supervised, unsupervised, and reinforcement learning—and advanced architectures such as transformers and CNNs.
- Skilled in building scalable AI/ML workflows using Python and SQL for handling large datasets, leveraging distributed processing with Spark, orchestration with Kubeflow, and deployment on AWS cloud infrastructure.
- Focused on ensuring data integrity through comprehensive data quality checks, reconciliation processes, and implementation of robust operational controls across ML pipelines.
- Developed AI solutions across diverse domains: voice assistants, chatbots, image classification, computer vision, NLP, recommendation systems, sentiment analysis, virtual reality, autonomous systems, healthcare applications, and predictive modeling.
- Proven ability to design and deploy real-time AI/ML solutions on AWS and Google Cloud Platform (GCP), ensuring scalability and high availability.
- Optimized model performance and reduced training time using mixed precision training, distributed training, and hyperparameter tuning.
- Built and fine-tuned transformer-based architectures such as NanoGPT, BERT, ViT, and LLaMA for NLP and computer vision tasks using PyTorch.
- Created interactive dashboards with Tableau, Power BI, and D3JS to visualize KPIs and drive business insights.
- Developed and implemented reinforcement learning algorithms to enable autonomous agents to learn optimal policies.
- Passionate about open-source AI tools, continuously exploring advancements in ML, LLMs, and multimodal learning.
- Designed and deployed advanced agentic workflows using LangChain, AutoGen, and LangGraph for dynamic multi-agent reasoning and task execution.
- Implemented Retrieval-Augmented Generation (RAG) pipelines with LlamaIndex and integrated vector databases (FAISS, Pinecone, Weaviate) for semantic search.
- Fine-tuned and applied LLMs (GPT-3/4, Claude, Llama) for high-accuracy, real-world AI solutions.
- Skilled in prompt engineering, designing and optimizing prompts for context-aware generative AI outputs.
- Architected and deployed graph-based workflows using LangGraph, integrating APIs and external data sources for adaptive, stateful conversational agents.
- Leveraged LangSmith for tracing, observability, and performance evaluation of LLM applications.
- Experienced in API testing using Postman and JavaScript; developed and implemented RESTful GET and POST APIs using Express.js.
- Skilled in creating and managing collaborative projects within Dataiku, including environment setup, package and plugin installation, and performance optimization.
- Experienced with data lakehouse architecture, integrating AWS services (S3, RDS, Redshift) and configuring database connections for large datasets.
- Experienced in leveraging Databricks to design and deploy scalable, Spark-based data pipelines and machine learning workflows, enabling efficient processing of large datasets and seamless integration with cloud storage and ML lifecycle tools like MLflow.

TECHNICAL SKILLS:

- **Programming Languages:** Python, R, SQL, Java, JavaScript (Node.js, React, Express), C++
- **ML & DL Frameworks:** TensorFlow, PyTorch, Keras, Scikit-learn, XGBoost, LightGBM, CatBoost, JAX
- **Computer Vision:** OpenCV, YOLO (You Only Look Once)
- **Natural Language Processing (NLP):** HuggingFace Transformers, spaCy, NLTK, Gensim, Prompt Engineering
- **Data Processing & Distributed Computing:** PySpark, Databricks
- **LLM & Agent Frameworks:** LangChain, LangGraph, CrewAI, LlamaIndex, HuggingFace, AutoGen, Sematic Kernel
- **MLOps & Model Deployment:** Docker, Kubernetes, MLflow, FastAPI, Flask, Django
- **Vector Databases & RAG:** FAISS, Pinecone, Weaviate, OpenSearch, pgVector, FAISS, Qdrant Milvus, ChromaDB
- **Cloud Platforms:** AWS (Bedrock, SageMaker, EC2, S3,), Google Cloud Platform (Vertex AI, BigQuery, Cloud Run), Azure ML
- **Data Analysis & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Plotly, Tableau, Power BI, Dataiku, Streamlit
- **Databases & Data Engineering:** MySQL, PostgreSQL, MongoDB, Snowflake
- **Version Control & Collaboration Tools:** Git, GitHub, JIRA, VS Code, Jupyter Notebooks, Pytest

PROFISSIONAL EXPERIENCE:

Cigna Healthcare Inc   **Mar 2025 to Present**
Title: Gen-AI ML Engineer

Responsibilities:

- Designed and developed a personalized healthcare assistant that provides users with tailored insights into their health plans, medical records, and wellness recommendations using LLMs and RAG pipelines.
- Ingested and unified heterogeneous healthcare datasets (EHR, lab reports, clinical notes, insurance plans) using PySpark, SQL, and Dataproc pipelines on GCP.
- Processed and stored large-scale clinical data securely in BigQuery and Cloud Storage, ensuring compliance with HIPAA and data governance standards.
- Built a document chunking system that segments long clinical notes and summaries into semantically meaningful sections for accurate retrieval.
- Generated dense embeddings for medical text using Vertex AI Embeddings API and domain-tuned Sentence Transformers, improving similarity matching across diverse note types.
- Indexed and searched clinical vectors using FAISS and Vertex Matching Engine, enabling low-latency retrieval across millions of records.
- Implemented hybrid retrieval (semantic + keyword search) using FAISS and BM25, increasing recall and precision of retrieved documents by over 25%.
- Designed a Retrieval-Augmented Generation (RAG) pipeline using LangChain to combine retrieved patient context with LLM reasoning for more factual responses.
- Fine-tuned domain-specific LLMs (BioBERT, ClinicalBERT, Clinical-LLaMA, GPT-3.5) using Supervised Fine-Tuning (SFT) with instruction-formatted clinical dialogues.
- Trained models on Vertex AI GPU instances (A100) and Databricks GPU clusters, using distributed training strategies with DeepSpeed and accelerate.
- Integrated Bedrock APIs with existing RAG workflows, enabling seamless orchestration of domain-specific LLMs with patient data for enhanced clinical decision support.
- Built instruction datasets from clinician-annotated summaries and medical recommendations, ensuring fine-tuned models understood context and terminology.
- Implemented few-shot and instruction-based prompting techniques to boost LLM response accuracy on rare or ambiguous medical queries.
- Created multi-agent workflows where specialized agents (retriever, summarizer, planner, recommender) collaborate dynamically to answer user queries.
- Used LangChain, AutoGen, and Semantic Kernel frameworks to orchestrate reasoning-based multi-agent pipelines for personalized health recommendations.
- Deployed fine-tuned LLMs as microservices using FastAPI, Docker, and Cloud Run, ensuring scalable, low-latency inference across healthcare use cases.

- Developed and maintained automated unit and integration tests using pytest to validate data processing and ML model components.
- Designed and implemented an end-to-end CI/CD pipeline on Azure using Jenkins, automating build, test, and deployment workflows for ML and web applications.
- Built explainability dashboards using Tableau and Plotly Dash to visualize retrieved evidence, model predictions, and confidence scores for clinician validation.
- Designed evaluation framework for LLM outputs using healthcare-specific metrics (Precision, Recall, F1, BLEU, ROUGE, and clinician-verified relevance).
- Collaborated with clinicians and data scientists to validate model outputs, improve interpretability, and achieve production-grade reliability for healthcare decision support.

University of North Texas/UNT   **August 2023 - December 2024**
Denton, Texas
Title: Research Student

Responsibilities:

- Developed an intelligent automation system using Large Language Models (LLMs) to extract key information from previously submitted documents and populate new digital forms with high accuracy.
- Designed context-aware prompt engineering strategies to improve the model's understanding of diverse form structures, terminologies, and user inputs.
- Built a robust pipeline to handle unstructured and semi-structured data, enabling consistent extraction of names, dates, identifiers, and contextual responses.
- Implemented validation logic and confidence-scoring mechanisms to ensure extracted fields met quality standards before being populated into new forms.
- Experimented with Python libraries such as PyPDF2, PDFPlumber, and PyMuPDF to preprocess, split, merge, and text-mine PDFs, improving the reliability of the extraction workflow.
- Developed an immersive Unity 3D virtual reality (VR) simulation designed to visualise emergency exit routes in real time for enhanced safety training.
- Integrated realistic physics, lighting, navigation meshes, and environment modelling to create a highly interactive and accurate building-evacuation experience.
- Programmed custom C# scripts to dynamically render exit paths and guide users using markers, animations, and audio cues.
- Conducted usability testing with target users to assess clarity of exit instructions, navigation flow, and overall training effectiveness.
- Packaged the simulation for deployment on VR headsets, supporting environments such as Oculus/Meta Quest, enabling scalable safety-preparedness training.

Carpus Consulting Services/ DTCC   **June 2022 - June 2023**
Hyderabad, India
Title: Machine Learning Data Scientist

Responsibilities:

- Collected and unified structured and unstructured data from multiple sources (APIs, databases, streaming platforms, logs).
- Store raw and processed data securely in Amazon S3 as the data lake.
- Use Dataiku for visual data preparation, cleansing, and normalization via drag-and-drop workflows, enabling faster collaboration with business users.
- Perform complex feature engineering and transformations in Databricks with PySpark for large-scale processing.
- Apply data validation and quality checks with Great Expectations integrated in Dataiku and Databricks environments.
- Conduct EDA in Dataiku and Databricks Notebooks to identify trends, outliers, and correlations.
- Leverage Dataiku's built-in visualizations and Databricks SQL Analytics for deep insights.
- Developed machine learning models using frameworks such as Scikit-learn, TensorFlow, and XGBoost on Amazon SageMaker's managed infrastructure, enabling scalable training and efficient resource utilization. Conducted hyperparameter tuning using SageMaker Automatic Model Tuning and custom grid/random search methods.

- Tracked experiments, hyperparameters, model versions, and performance metrics using MLflow, integrated seamlessly across SageMaker and Databricks environments to ensure reproducibility and effective model management.
- Automate data pipelines with Apache Airflow or AWS Step Functions, coordinating data ingestion, feature updates, model training, and validation.
- Schedule retraining workflows triggered by new data arrival or model performance degradation.
- Use Amazon SageMaker to train models on managed GPU/CPU instances with distributed training support.
- Use Spot Instances and hyperparameter tuning jobs for cost-effective, optimized model development.
- Deploy trained models as scalable, real-time inference endpoints using SageMaker Endpoints or batch jobs via SageMaker Batch Transform.
- Containerize models with Docker and serve via AWS Lambda + API Gateway or microservices built with FastAPI and orchestrated on Amazon ECS/Fargate or Kubernetes.
- Monitor model accuracy, latency, and data/model drift using SageMaker Model Monitor and custom logging with Amazon CloudWatch.
- Set alerts on performance degradation and automate retraining pipelines to maintain production model health.
- Create dashboards in Dataiku, Power BI, and Tableau to communicate model insights and business KPIs.
- Enable self-service analytics for business users via Dataiku's visual tools.
- Enforce data and model access controls using AWS IAM, encryption with AWS KMS, and data cataloging/governance via Databricks Unity Catalog.
- Ensure compliance with regulatory standards by auditing data usage and model lifecycle.
- Use SageMaker Studio and Databricks Workspaces for collaborative notebooks, code sharing, and version control integration (Git).
- Document workflows, architecture, and decisions using Confluence, diagrams, or README files to enable reproducibility and knowledge sharing.

Carpus Consulting Services/ DTCC   **June 2021 - June 2022**
Hyderabad, India
Title: Data Scientist Intern

Responsibilities:

- Built and maintained end-to-end Dataiku workflows for data ingestion, transformation, and model development, automating recurring finance processes such as monthly reporting, budgeting, and forecasting using Dataiku's scenario engine and visual recipes.
- Converted legacy SQL and Alteryx scripts into scalable Dataiku workflows, improving efficiency, automation, and maintainability of finance-related ETL pipelines.
- Cleaned, transformed, and validated financial datasets using Pyspark SQL, Python, and Dataiku, ensuring high-quality inputs for analytics, reporting, and machine-learning models. Implemented data quality checks, validation rules, and metadata management across pipelines.
- Designed financial data models, KPIs, and dashboards in collaboration with FP&A and finance stakeholders, providing actionable insights to support data-driven decision-making. Created interactive visualizations in Dataiku dashboards and Tableau.
- Implemented machine learning and time-series models for forecasting, budgeting, and financial planning using Python, Scikit-learn, Statsmodels, Prophet, and XGBoost, leveraging Dataiku and Azure (Azure ML, Databricks, Synapse) for scalable data processing, model training, deployment, and monitoring.
- Productionized and monitored finance workflows and ML models in Azure environments, including automation via Dataiku scenarios, containerized application Docker, and Databricks notebooks to ensure reliability and scalability.
- Collaborated with cross-functional teams to document data pipelines, feature engineering logic, experiments, and results, ensuring reproducibility, maintainability, and compliance with data governance standards.

EDUCATION:

Master of Science: Artificial Intelligence **December 2024**
University of North Texas, Denton, Tx
- Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing, Feature Engineering, Empirical Analysis, Scientific Data Visualization.

Bachelor of Technology: Computer Science **May 2022**
Hindustan Institute of Technology and Science, Chennai, India
- Relevant Coursework: C, C++, Java, Python, Dats Structures and Algorithms, Web technologies, Computer Vision, Database Management Systems.

CERTIFICATIONS:

Machine Learning Professional Certification - Altair RapidMiner

Cloud Machine Learning Engineering and MLOps – Coursera, Duke University