Sai Sandeep Kethiboina

AI & Machine Learning Engineer

********** | **********

CAREER OBJECTIVE

AI/ML Engineer and Data Scientist with **5**+ years of experience designing, developing, and deploying scalable Artificial Intelligence, Machine Learning, Deep Learning, Predictive Analytics, and Generative AI solutions across Telecommunications, Banking, and Healthcare domains. Expertise in Python, SQL, TensorFlow, PyTorch, Scikit-learn, XGBoost, LightGBM, Apache Spark, Pandas, NumPy, and Statistical Modeling for building data-driven products and predictive systems. Hands-on experience developing LLMs, Generative AI, RAG, Agentic AI, AI Agents, NLP, Time Series Forecasting, Customer Analytics, Fraud Detection, Churn Prediction, Risk Modeling, and Recommendation Systems using OpenAI, Hugging Face, LangChain, LangGraph, FAISS, Pinecone, and ChromaDB. Skilled in building scalable Data Pipelines, ETL Workflows, REST APIs, Microservices, FastAPI, Flask, Docker, Kubernetes, MLflow, Kubeflow, Apache Airflow, and MLOps/LLMOps platforms. Proficient in AWS SageMaker, S3, Lambda, Vertex AI, Azure ML, Snowflake, Hadoop, Kafka, Data Warehousing, and Big Data Technologies for enterprise-scale analytics and AI deployments. Proven track record of delivering production-ready solutions that improve operational efficiency, customer experience, revenue growth, and business decision-making through AI Governance, Explainable AI (XAI), System Design, Solution Architecture, and Agile methodologies.

TECHNICAL SKILLS

- **Programming:** Python, SQL, R, Scala, Java, C++, Bash/Shell
- **AI/ML:** Machine Learning, Deep Learning, NLP, Computer Vision, Reinforcement Learning, Feature Engineering, Model Optimization, Time Series Forecasting
- **Generative AI & LLMs:** OpenAI, Azure OpenAI, Hugging Face, LangChain, LangGraph, LlamaIndex, RAG, Prompt Engineering, LoRA, PEFT, Fine-Tuning, Agentic AI, AI Agents, Multi-Agent Systems
- **Frameworks & Libraries:** TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, LightGBM, CatBoost, Pandas, NumPy, SciPy
- **Vector Databases:** FAISS, Pinecone, ChromaDB, Weaviate
- **APIs & Deployment:** FastAPI, Flask, REST APIs, TorchServe, ONNX Runtime
- **MLOps & DevOps:** MLflow, Kubeflow, Docker, Kubernetes, CI/CD, Jenkins, GitHub Copilot, Git, GitLab, Terraform, Helm
- **Cloud:** AWS SageMaker, S3, Lambda, EC2, Vertex AI, BigQuery, Azure ML, Azure OpenAI, Databricks
- **Data Engineering:** Apache Spark, Kafka, Hadoop, Hive, Apache Airflow, ETL/ELT, Data Warehousing
- **Databases:** MySQL, PostgreSQL, MongoDB, Cassandra, DynamoDB, Redis
- **Visualization:** Tableau, Power BI, Matplotlib, Seaborn, Plotly, Jupyter Notebook
- **Monitoring:** Prometheus, Grafana, ELK Stack
- **Methodologies:** MLOps, LLMOps, Agile, Scrum, SDLC, System Design, Solution Architecture

EDUCATION

- Masters in Computers & Information Science from ********** University – USA (XXXX - XXXX)

RELEVANT EXPERIENCE

Ai / Machine Learning Engineer **CVS Health, Texas, USA** **Feb 2025 - Current**

CVS Health is a leading healthcare services company that provides pharmacy services, health insurance, retail healthcare, and care management solutions across the United States. Through its integrated healthcare ecosystem, it helps improve access to care, reduce healthcare costs, and enhance patient outcomes using data-driven and technology-enabled services.

Responsibilities:

- Designed and implemented end-to-end AI/ML pipelines using **Python, SQL, Apache Spark, Apache Airflow, Pandas, NumPy, ETL, Feature Engineering, Model Training, Model Deployment, and Model Monitoring** to process over 50M+ healthcare records, improving data availability by 40% and accelerating predictive analytics initiatives.
- Developed and deployed **Machine Learning, Deep Learning, TensorFlow, PyTorch, Scikit-learn, XGBoost, LightGBM, Supervised Learning, Unsupervised Learning, and Time Series Forecasting** models for patient risk

stratification, hospital readmission prediction, disease progression forecasting, and claims fraud detection, improving model accuracy by 25%.

- Built enterprise **Generative AI (GenAI)** solutions leveraging **OpenAI, Hugging Face, LangChain, Transformers, Prompt Engineering, Fine-Tuning, LoRA, Quantization, LLMOps, and Large Language Models (LLMs)** to automate clinical documentation, medical coding, and provider support workflows, reducing manual effort by 35%.
- Engineered **LLM-based Healthcare Assistants** using **Retrieval-Augmented Generation (RAG), Semantic Search, Vector Search, Knowledge Graphs, AI Agents, Agentic AI, Multi-Agent Systems, FAISS, Pinecone, and Vector Databases**, enabling secure retrieval of clinical guidelines, policy documents, and patient care information with over 90% response relevance.
- Implemented advanced **Natural Language Processing (NLP), Computer Vision, Transformers, Hugging Face, OpenCV, OCR, Named Entity Recognition (NER), Sentiment Analysis, Text Classification, and Document Intelligence** solutions to extract insights from EHRs, physician notes, claims data, and unstructured healthcare documents, improving processing efficiency by 45%.
- Architected scalable **Microservices, REST APIs, Flask, Docker, Kubernetes, Real-Time Inference, Batch Inference, and Event-Driven Architecture** platforms to support production AI workloads, achieving 99.9% application availability and low-latency model serving.
- Established enterprise **MLOps, MLflow, Kubeflow, CI/CD Pipelines, Model Registry, Feature Store, Data Versioning, Model Monitoring, A/B Testing, and Distributed Training** practices, reducing model deployment time by 60% while improving model governance and reliability.
- Deployed cloud-native AI solutions using **AWS SageMaker, Amazon S3, AWS Lambda, Google Cloud Vertex AI, Cloud Computing, Hybrid Cloud Architecture, and Model Serving**, enabling scalable training, deployment, and monitoring of healthcare analytics applications.
- Built large-scale **Data Engineering, Apache Spark, Apache Airflow, Data Pipelines, ETL Pipelines, Data Warehousing, PostgreSQL, MongoDB, NoSQL Databases, and Big Data Processing** solutions integrating EHR, EMR, claims, pharmacy, provider, and member datasets, improving processing performance by 50%.
- Led AI solution delivery using **System Design, Solution Architecture, Explainable AI (XAI), Responsible AI, AI Governance, Agile/Scrum, Stakeholder Management, Power BI, Matplotlib, Seaborn, and Jupyter Notebook**, enabling data-driven clinical and operational decision-making across healthcare organizations.

Machine Learning Engineer **Capital One, Texas, USA** **Feb 2024 - Dec 2024**

Capital One is a leading U.S.-based financial services company that provides credit cards, consumer banking, commercial banking, and lending solutions to millions of customers. It leverages advanced analytics, cloud technologies, and AI-driven platforms to deliver secure, personalized financial products and digital banking experiences.

Responsibilities:

- Developed and deployed end-to-end **Machine Learning**, **Deep Learning**, and **Predictive Analytics** solutions using **Python, SQL, TensorFlow, PyTorch, Scikit-learn, XGBoost, and LightGBM** to support credit risk assessment, fraud detection, customer segmentation, and loan default prediction, improving model accuracy by 30%.
- Built scalable **ETL Pipelines**, **Data Pipelines**, and real-time data processing frameworks using **Apache Spark, Apache Airflow, PostgreSQL, MongoDB, and NoSQL Databases** to integrate transaction, customer, credit card, and banking datasets, reducing data processing time by 45%.
- Designed and implemented enterprise **Generative AI (GenAI)** applications using **OpenAI, Hugging Face, LangChain, Prompt Engineering, Fine-Tuning, LoRA, and LLMOps** to automate customer service operations, financial document analysis, and regulatory reporting, reducing manual effort by 40%.
- Engineered **LLM-based** intelligent banking assistants leveraging **RAG Architecture, Semantic Search, Vector Search, FAISS, Pinecone, AI Agents, Agentic AI, and Vector Databases** to enable secure retrieval of financial policies, compliance documents, and customer knowledge bases with 90%+ response accuracy.
- Developed advanced **NLP**, **Transformers**, **Named Entity Recognition (NER)**, **Text Classification**, and **Document Intelligence** solutions to extract insights from loan applications, customer communications, KYC records, and regulatory documents, improving operational efficiency by 35%.

- Implemented **Time Series Forecasting**, **Anomaly Detection**, and **Predictive Modeling** frameworks for transaction monitoring, liquidity forecasting, revenue prediction, and fraud prevention, enabling proactive business decision-making and reducing financial risk exposure.
- Architected scalable **REST APIs**, **Microservices**, **Flask**, **FastAPI**, **Docker**, **Kubernetes**, **Real-Time Inference**, and **Batch Inference** platforms to support mission-critical AI applications processing millions of financial transactions daily with 99.9% system availability.
- Established enterprise-grade **MLOps** and **LLMOps** practices using **MLflow, Kubeflow, CI/CD Pipelines, Model Registry, Feature Store, Data Versioning, Model Monitoring, A/B Testing, and Automated Retraining**, reducing model deployment cycles by 60% and improving governance compliance.
- Leveraged **AWS SageMaker, Amazon S3, AWS Lambda, Vertex AI, Cloud Computing, and Distributed Training** to build scalable cloud-native AI solutions supporting customer analytics, risk modeling, fraud detection, and personalized banking recommendations.

Data Scientist **Jio, Hyderabad, India** **Jun 2019 - Nov 2022**

Jio is an Indian telecommunications company and a subsidiary of Jio Platforms. Applied the techniques like predictive modeling, time-series forecasting, natural language processing (NLP), and computer vision to high-impact projects. Ensured the data quality, governance, and compliance with relevant industry and regional regulations.

Responsibilities:

- Designed and deployed end-to-end **Data Science**, **Machine Learning**, and **Predictive Analytics** solutions using **Python, SQL, Scikit-learn, XGBoost, TensorFlow, Spark MLlib, Apache Spark, and PySpark** to analyze telecom subscriber behavior, predict customer churn, optimize recharge campaigns, and improve customer retention by 25%.
- Built scalable **ETL Pipelines**, **Data Pipelines**, and **Data Lake** solutions using **Apache Airflow, Apache Spark, Hadoop, Hive, Snowflake, PostgreSQL, MongoDB, Cassandra, HBase, Redshift, and NoSQL Databases** to process over 100M+ customer, billing, network, and usage records, reducing data processing time by 40%.
- Developed **Time Series Forecasting**, **Customer Segmentation**, **Demand Forecasting**, **Anomaly Detection**, **Classification**, **Regression**, and **Dimensionality Reduction** models to improve network capacity planning, subscriber growth forecasting, service quality monitoring, and operational efficiency.
- Processed real-time telecom streaming data using **Kafka, Kinesis, Spark Streaming, AWS, Hadoop, and Big Data Technologies** to monitor network performance, identify service disruptions, and generate actionable insights for network operations teams, reducing incident resolution time by 30%.
- Created executive dashboards and self-service analytics solutions using **Power BI, Tableau, Matplotlib, Seaborn, and Jupyter Notebook**, enabling stakeholders to track ARPU, customer lifetime value, churn trends, network utilization, and business KPIs, improving decision-making efficiency by 35%.
- Collaborated with business, product, marketing, and network engineering teams in **Agile/Scrum** environments while leveraging **MLflow, Statistical Analysis, Data Modeling, Feature Engineering, AWS Cloud Services, and Advanced Analytics** to deliver data-driven solutions that increased customer engagement by 30% and supported strategic telecom growth initiatives.