SAI CHANDU MACHAVARAPU

AI / Machine Learning Engineer

PROFESSIONAL SUMMARY

Results-driven AI & ML Engineer with 4+ years of hands-on experience designing, building, and deploying production-grade machine learning systems across banking and healthcare domains. Proven expertise in the full ML lifecycle — from raw data ingestion, ETL pipeline design, and feature engineering through model training, experiment tracking, and cloud-native deployment — using PyTorch, TensorFlow, HuggingFace Transformers, and MLflow. Deep specialization in Generative AI and LLM engineering including RAG pipeline architecture, LoRA/PEFT fine-tuning, Prompt Engineering, and agentic AI systems using LangChain, LlamaIndex, and vector databases. Experienced in building scalable MLOps infrastructure with Docker, Kubernetes, Kubeflow, and CI/CD automation across AWS (SageMaker, Lambda, EC2), GCP (Vertex AI), and Azure (Azure ML, Azure OpenAI). Skilled at partnering with cross-functional stakeholders to translate complex business problems into deployable AI solutions that measurably improve efficiency, accuracy, and business outcomes.

TECHNICAL SKILLS

Programming	Python, SQL, R, Scala, Java, C++, Bash/Shell Scripting
AI / ML	Machine Learning, Deep Learning, Supervised Learning, Unsupervised Learning, Reinforcement Learning, Feature Engineering, Data Preprocessing, Model Optimization, Hyperparameter Tuning, Time Series Forecasting, Recommendation Systems, Anomaly Detection
Generative AI & LLMs	Generative AI, LLMs, RAG, Prompt Engineering, LoRA, PEFT, OpenAI API, Azure OpenAI, Hugging Face, LangChain, LlamaIndex, Transformers, FAISS, Pinecone, ChromaDB
NLP & Computer Vision	Natural Language Processing (NLP), NER, Sentiment Analysis, Text Classification, Semantic Search, Conversational AI, Computer Vision, OCR, OpenCV
Frameworks & Libraries	TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, LightGBM, CatBoost, Pandas, NumPy, SciPy, spaCy, NLTK
MLOps & Deployment	MLflow, Kubeflow, Docker, Kubernetes, CI/CD Pipelines, FastAPI, Flask, REST APIs, Model Serving, TorchServe, ONNX Runtime
Cloud Platforms	AWS (SageMaker, S3, Lambda, EC2), GCP (Vertex AI, BigQuery), Azure (Azure ML, Azure OpenAI)
Data Engineering	Apache Spark, Kafka, Hadoop, Hive, ETL/ELT Pipelines, Data Pipelines, Data Warehousing, Apache Airflow, Big Data Technologies
Databases	MySQL, PostgreSQL, MongoDB, Cassandra, DynamoDB, NoSQL Databases, Vector Databases
DevOps & Tools	Git, GitHub, GitHub Copilot, GitLab, Jenkins, Terraform, Helm, Jupyter Notebook, VS Code, Databricks, Jira, Confluence
Monitoring & Visualization	Prometheus, Grafana, ELK Stack, Tableau, Power BI, Matplotlib, Seaborn, Plotly
Mathematics & Statistics	Probability, Statistics, Linear Algebra, Bayesian Modeling, Hypothesis Testing, A/B Testing

EDUCATION

M.S. Computer Science & Information Systems

Aug 2024 – May 2026

University of Texas at Tyler | Tyler, Texas, USA | GPA: 3.6 / 4.0

WORK EXPERIENCE

AI / Machine Learning Engineer | Comerica

Nov 2025 – Present

Dallas, Texas, USA

- Designed and implemented scalable end-to-end AI/ML pipelines — data ingestion, ETL, feature engineering, model training, validation, deployment, and monitoring — using Python, SQL, Apache Spark, and Apache Airflow for enterprise banking and financial analytics, reducing data processing latency by 35%.
- Developed and optimized ML and Deep Learning models using TensorFlow, PyTorch, Scikit-learn, XGBoost, and LightGBM for fraud detection, credit risk scoring, customer segmentation, anomaly detection, and time series forecasting, improving fraud detection precision by 22%.
- Built advanced Generative AI and LLM solutions using OpenAI API, Azure OpenAI, Hugging Face, LangChain, RAG architectures, Prompt Engineering, LoRA/PEFT fine-tuning, and vector databases (FAISS, Pinecone, ChromaDB) for intelligent document search, financial report analysis, and conversational AI systems.
- Implemented NLP pipelines using spaCy, NLTK, and HuggingFace Transformers for text classification, sentiment analysis, NER, semantic search, topic modeling, and automated financial document processing across regulatory compliance workflows.
- Architected enterprise-grade MLOps frameworks using MLflow, Kubeflow, Docker, Kubernetes, CI/CD pipelines, GitHub Copilot, Jenkins, and DVC to automate model training, experiment tracking, version control, deployment, and monitoring across production environments.
- Built scalable REST APIs and AI microservices using FastAPI and Flask integrated with AWS SageMaker, Lambda, S3, and GCP Vertex AI for real-time and batch inference at production scale.
- Implemented model governance and monitoring using SHAP, LIME, Prometheus, and Grafana with drift detection, explainability reports, and responsible AI practices aligned to regulated banking compliance standards.

Machine Learning Engineer | **Memorial Hermann Health System** *Nov 2024 – Oct 2025*

Houston, Texas, USA

- Designed end-to-end ML and AI solutions for healthcare analytics covering data preprocessing, feature engineering, model training, validation, deployment, and monitoring using Python, SQL, TensorFlow, PyTorch, and Scikit-learn.
- Built Generative AI and LLM applications using OpenAI, Hugging Face, LangChain, RAG pipelines, Prompt Engineering, FAISS, and Pinecone for intelligent medical document search, clinical note summarization, and conversational AI systems for care teams.
- Developed predictive Deep Learning models using CNNs, RNNs, LSTMs, Transformers, XGBoost, and LightGBM for patient risk stratification, 30-day readmission prediction, anomaly detection, and clinical outcome forecasting.
- Designed and deployed intelligent LangChain agents leveraging LLMs for automated document authoring, semantic search, and clinical recommendation tasks — reducing manual review workload by 40%.
- Implemented enterprise MLOps using MLflow, Kubeflow, Docker, Kubernetes, CI/CD, FastAPI, and Flask; deployed models on GCP Vertex AI and Azure Machine Learning for secure, HIPAA-aligned, scalable healthcare AI workloads.
- Built scalable ETL and real-time processing frameworks using Apache Spark, Kafka, Airflow, PostgreSQL, and MongoDB to support large-scale healthcare data analytics and reporting pipelines.

Data Scientist | **Equitas Small Finance Bank** *Apr 2022 – Jul 2024*

Chennai, India

- Developed and deployed predictive models using Python, Scikit-learn, TensorFlow, XGBoost, and LightGBM for credit risk scoring, fraud detection, customer segmentation, anomaly detection, and financial forecasting across retail and SME banking portfolios.
- Built scalable ETL pipelines and data processing workflows using Apache Spark, Pandas, PostgreSQL, and MongoDB to process high-volume structured and unstructured banking datasets for model training and reporting.
- Implemented NLP solutions using spaCy, NLTK, Transformers, and Hugging Face for customer feedback analysis, document classification, sentiment analysis, and automated loan document text processing.
- Deployed model services using FastAPI, Flask, Docker, Kubernetes, MLflow, and CI/CD pipelines on AWS SageMaker; built Tableau and Power BI dashboards for business intelligence and executive reporting.
- Performed in-depth statistical analysis, feature engineering, cross-validation, and hyperparameter tuning; applied A/B testing and causal inference methods to validate model performance in controlled experiments.

Data Scientist | **MRF Tyres** *Mar 2020 – Mar 2022*

Chennai, India

- Developed predictive analytics and machine learning models using Python, Scikit-learn, TensorFlow, PySpark, and Spark MLlib for demand forecasting, quality defect prediction, anomaly detection, and operational optimization across tyre manufacturing.
- Built scalable big data pipelines using Apache Spark, Kafka, Azure Databricks, Hive, and Hadoop to ingest and process large-scale manufacturing sensor and operational datasets in near-real-time.
- Applied feature engineering, PCA, cross-validation, and hyperparameter tuning with ROC-AUC and precision-recall evaluation; performed customer segmentation and time-series forecasting to optimize inventory and supply chain planning.
- Delivered interactive dashboards using Tableau, Power BI, Matplotlib, and Seaborn for production KPIs, model performance tracking, and cross-functional stakeholder reporting.

FEATURED PROJECTS

RAG-Powered Document Intelligence Platform — *LangChain · OpenAI API · FAISS · Pinecone · FastAPI · ChromaDB · Azure OpenAI*

- Architected a production-grade RAG system for financial and healthcare document Q&A; ingested PDFs, reports, and regulatory filings via a chunking and embedding pipeline using OpenAI embeddings stored in FAISS and Pinecone vector databases, achieving sub-200ms query response times.
- Built a multi-turn conversational AI interface with LangChain agents, tool-calling, and memory management; integrated with Azure OpenAI and Hugging Face models with Prompt Engineering and guardrails — reducing manual document review time by over 60% in production.

Healthcare Patient Risk Prediction System — *PyTorch · XGBoost · SageMaker · MLflow · FastAPI · SHAP · Evidently AI*

- Built an end-to-end clinical risk prediction platform for 30-day hospital readmission and patient deterioration using ensemble models (XGBoost, LightGBM) and a custom PyTorch Transformer architecture trained on structured EHR data — achieving AUC-ROC of 0.91 on held-out validation sets.
- Deployed HIPAA-aligned inference service via FastAPI on AWS SageMaker with real-time SHAP explainability outputs per prediction; integrated Evidently AI for continuous data drift monitoring with automated retraining triggers on feature distribution shifts.

Real-Time Fraud Detection & Anomaly Pipeline — *XGBoost · Isolation Forest · Apache Kafka · PySpark · AWS Lambda · MLflow*

- Designed a streaming fraud detection system processing 50,000+ transactions/minute using Apache Kafka for event ingestion, PySpark for feature computation, and an ensemble of XGBoost and Isolation Forest models for real-time anomaly scoring — achieving 94% precision with sub-50ms inference latency.
- Automated full ML lifecycle using MLflow for experiment tracking and model registry, AWS Lambda for trigger-based retraining on data drift signals, and CloudWatch for latency, throughput, and alert routing to on-call engineers via SNS.

Demand Forecasting & Time Series Intelligence Platform — *LSTM · Prophet · Apache Airflow · SageMaker · Plotly · MLflow*

- Built a multi-horizon demand forecasting system combining Meta Prophet for trend/seasonality decomposition and a stacked LSTM network for residual modelling across 200+ SKUs; outperformed ARIMA baseline by 31% on MAPE across 90-day forecasting windows.
- Orchestrated automated daily retraining pipelines using Apache Airflow with SageMaker training jobs; delivered an interactive Plotly dashboard with forecast confidence intervals, anomaly flags, and model performance history — integrated directly into supply chain planning workflows.

CERTIFICATIONS

▪ **Machine Learning Specialization**	DeepLearning.AI	*2025*
▪ **Generative AI with LLMs**	DeepLearning.AI	*2025*
▪ **AI for Medicine Specialization**	DeepLearning.AI	*2025*
▪ **AI Engineer for Developers Associate**	DataCamp	*2025*
▪ **Oracle Certified AI Foundations Associate**	Oracle	*2024*
▪ **Oracle AI Vector Search Certified Professional**	Oracle	*2024*
▪ **Oracle Database@AWS Certified Architect Pro**	Oracle	*2024*
▪ **GitHub Copilot Certified**	GitHub / Microsoft	*2024*
▪ **Azure Administrator Associate**	Microsoft	*2023*
▪ **Azure AI Fundamentals**	Microsoft	*2023*