

SAI TEJA REDDY

saitejareddy97@gmail.com | 2488033210 | Edmond, OK | <https://www.linkedin.com/in/sai-teja-reddy-b5846322b/>

Professional Summary

- Experienced and detail-oriented AI/ML Engineer with 5 years of full-time industry experience delivering production-grade machine learning, AI, and data science solutions across highly regulated and data-intensive sectors including telecommunications, finance, and retail.
- Demonstrated expertise in designing robust ML pipelines, fine-tuning large language models (LLMs), architecting NLP systems, and integrating predictive models into scalable and cloud-native production environments.
- Deep understanding of transforming complex datasets into actionable business intelligence, with a strong track record of developing advanced AI solutions that support mission-critical business functions.
- Proficient in Python programming and deeply experienced with machine learning libraries such as TensorFlow, PyTorch, Hugging Face Transformers, Scikit-learn, and Spark MLlib.
- Successfully developed and deployed end-to-end AI models for a variety of use cases including intent classification, customer behavior prediction, credit scoring, and multilingual speech recognition.
- Skilled in deploying AI systems using cloud platforms like AWS SageMaker, GCP Vertex AI, and Azure ML; optimized for reliability, scalability, and cost-effectiveness.
- Well-versed in MLOps principles including model tracking, CI/CD integration, container orchestration (Docker), MLflow usage, and continuous monitoring for production models.
- Extensive experience working with big data ecosystems including Kafka, Hadoop, Hive, and Apache Spark to process and manage massive unstructured and structured datasets in real time.
- Known as a collaborative and goal-driven professional, capable of leading cross-functional AI initiatives and mentoring junior engineers on model best practices, code quality, and scalability.
- Strong communicator adept at translating complex ML outputs into strategic insights for both technical and non-technical stakeholders across product, engineering, and executive teams.
- Holds multiple industry-recognized certifications including Salesforce Certified AI Associate, Salesforce AI Specialist, and AWS Certified Developer – Associate.
- Currently pursuing a Master's in Computer Science with a focus on Artificial Intelligence; actively engaged in advancing knowledge in generative AI, multi-modal learning architectures, and ethical AI frameworks.

Core Competencies

- End-to-End Machine Learning Development & Deployment
- LLM Fine-Tuning, Prompt Engineering & Retrieval-Augmented Generation (RAG)
- Real-Time Predictive Analytics & Data Streaming (Kafka, Spark, Hive)
- NLP Systems for Intent Classification, Summarization, Named Entity Recognition
- Model Evaluation, Optimization & Explainability (F1, AUC, SHAP, LIME)
- Forecasting & Anomaly Detection in Time Series Data
- CI/CD Pipelines, Containerization, MLflow, FastAPI & DevOps Tools
- Cloud-Native AI on AWS, GCP & Azure
- Data Wrangling, Feature Engineering, SQL Automation & Airflow ETL
- Ethical AI, Data Governance & Privacy-Safe Model Design
- Agile Development, Sprint Planning, & Cross-Functional Team Collaboration
- Business Storytelling, Technical Documentation & Stakeholder Engagement
- Customer Enablement, AI Education Sessions & Technical Stakeholder Demos
- OpenAI, LangChain, RAG System Integration & Workflow Optimization

Skills

- **Programming Languages:** Python, R, SQL, Shell Scripting, JavaScript (basic), Bash
- **Libraries/Frameworks:** TensorFlow, PyTorch, Hugging Face, Keras, Scikit-learn, OpenCV, FastAI, Gensim, SpaCy, NLTK, XGBoost, LightGBM, OpenAI API, LangChain
- **Cloud Platforms:** AWS (SageMaker, Lambda, EC2, S3, CloudFormation), Google Cloud (Vertex AI, BigQuery), Microsoft Azure (Data Lake, ML Studio)
- **Big Data & ETL:** Hadoop, Spark, Hive, Kafka, MapReduce, Airflow, Snowflake, Sqoop, Apache Beam
- **Databases:** PostgreSQL, MySQL, MongoDB, Elasticsearch, SQLite, DynamoDB
- **Visualization & BI Tools:** Power BI, Tableau, Excel, Matplotlib, Seaborn
- **DevOps & MLOps:** Docker, GitHub Actions, CI/CD, Jenkins, MLflow, Flask, FastAPI, Swagger, RESTful API Development, Logging/Monitoring
- **Experiment Tracking Tools:** Weights & Biases, TensorBoard
- **Other Tools:** Jupyter Notebook, Google Colab, VS Code, Postman, Trello, Jira

Experience

- 03/2022 - 07/2024
Ericsson Canada
Ottawa, Ontario,
Canada
Machine Learning Engineer
 - Designed, deployed, and maintained large-scale AI applications focused on real-time decision systems for telecom networks, supporting 100M+ customer events daily.
 - Fine-tuned transformer-based LLMs for conversational bots handling network diagnostics using domain-specific corpora, reducing customer query resolution time by 41%.
 - Developed custom tokenizers and vector stores to support semantic search on customer knowledge bases, improving top-3 hit accuracy to 92%.
 - Led integration of retrieval-augmented generation (RAG) with LangChain into Ericsson's internal knowledge platform using Pinecone and FAISS.
 - Delivered predictive modeling pipelines in AWS SageMaker to forecast radio signal degradation using 6+ months of real-time telemetry logs.
 - Automated daily ML workflows via Airflow and integrated CI/CD using MLflow + Docker + GitHub Actions. Reduced model deployment lag from 5 days to under 6 hours.
 - Conducted SHAP-based explainability audits for regulatory review of AI models in telecom automation. Shared best practices across international teams.
 - Partnered with security teams to implement model guardrails and monitor data drift, ensuring continued fairness and robustness.
 - Managed stakeholder demos, internal training workshops, and AI ethics briefings. Recognized as the go-to technical resource for all AI/ML initiatives within the North America division.
 - Presented research outcomes at internal innovation day and received cross-team recognition for model reproducibility and deployment standards.
 - Collaborated with client-facing product teams and internal stakeholders to align GenAI model deployment with telecom business KPIs and technical architecture requirements.
 - Led internal enablement sessions on LangChain-based RAG pipelines, prompt tuning workflows, and deployment best practices across multiple AI teams.
 - Delivered technical deep-dives and strategy reviews to engineering and executive audiences, supporting adoption of AI-driven automation platforms.
 - Consolidated user feedback on retrieval-based QA systems and contributed actionable insights to internal platform roadmap discussions.
- 06/2021 - 02/2022
Cash4You Inc.
Cambridge, Ontario,
Data Scientist
 - Developed predictive credit scoring models using XGBoost and LightGBM that outperformed legacy logistic regression baselines, leading to a 23% reduction in loan default rates.

Canada

- Engineered over 70 features from transactional, demographic, and behavioral datasets to improve customer segmentation and targeting.
- Created real-time API for fraud detection scoring using Flask and integrated with internal loan processing application, reducing application fraud by 18%.
- Partnered with business analysts and compliance officers to translate regulatory requirements into AI-friendly audit logs and data handling pipelines.
- Built dynamic dashboards in Tableau and Power BI for ongoing loan performance monitoring and risk forecasting.
- Automated ETL jobs using Python and Airflow to maintain pipeline reliability across multiple third-party data ingestion sources.
- Designed and conducted experiment-based uplift modeling to identify optimal email reminders for late payment reductions, increasing recovery rates by 12%.
- Received internal spotlight award for designing the AI-driven fraud engine adopted across three branches.

09/2018 - 11/2019

Data Analyst

SanSah Innovations
Hyderabad,
Telangana, India

- Collaborated with cross-functional teams to collect, validate, clean, and model customer sales and operational data, identifying revenue leakages.
- Developed reports and dashboards using Power BI and Excel to track key performance indicators, supporting monthly executive strategy meetings.
- Built customer lifetime value (CLTV) models using regression techniques and helped marketing teams run targeted loyalty campaigns.
- Cleaned datasets from CRM, surveys, and point-of-sale systems using advanced SQL joins and Python scripts.
- Created Excel-based simulators to forecast costs and returns of marketing campaigns under various budget scenarios.
- Trained junior analysts on data modeling best practices and developed SOPs for internal reporting workflows.
- Supported cross-functional process automation initiative resulting in ~8% time savings per quarter in reporting operations.

Projects

LLM Classification Fine-Tuning

- Designed and implemented a complete fine-tuning pipeline for binary classification of instruction prompt pairs using Hugging Face Transformers with distilbert-base-uncased.
- Preprocessed prompt-completion pairs into balanced comparison samples using pairwise formatting, truncation, and token-level encoding with attention masks.
- Utilized Hugging Face Datasets library to load, tokenize, and prepare PyTorch-ready input batches; enabled max-length padding and truncation for uniform sequence handling.
- Trained DistilBERT on a subset of 5,000 Kaggle samples to optimize GPU memory usage and experiment speed, leveraging the Trainer API and callback-based evaluation.
- Monitored training performance using validation loss and accuracy metrics to assess convergence and model generalization.
- Developed a modular and reusable codebase in Google Colab for iterative experimentation, hyperparameter tuning, and fast prototyping.
- Ensured full integration of dataset loading, tokenizer configuration, and training logic to create a reproducible and scalable experimentation pipeline.

HCU Speech Recognition

- Developed a multilingual speech recognition and translation pipeline using Hugging Face Transformers, combining `airesearch/wav2vec2-large-xlsr-53-th` for Thai ASR and `Helsinki-NLP/opus-mt-th-en` for Thai-to-English translation.
- Preprocessed audio with `librosa` and passed waveform data to `Wav2Vec2Processor` for tokenization and attention mask generation.
- Integrated ASR and NMT logic into a single function, combining Connectionist Temporal Classification (CTC)-based decoding and MarianMT translation in a real-time pipeline.
- Structured input datasets from audio folders and dynamically mapped predictions to filenames and transcripts for batch evaluation.
- Built the entire system in Google Colab with modular, reproducible cells to enable rapid experimentation, translation comparison, and visualization.
- Validated system outputs manually across dialect-influenced audio samples and tone-heavy pronunciations to assess robustness.
- Demonstrated applied understanding of multilingual ASR + NMT architecture and Hugging Face pipeline orchestration for speech translation tasks.

Certifications

- Salesforce Certified AI Associate
- Salesforce Certified AI Specialist
- AWS Certified Developer – Associate

Education

- 08/2024 - 08/2025 **Master of Science** in Computer Science – Artificial Intelligence
Edmond, OK Oklahoma Christian University
- 01/2020 - 05/2021 **PG Diploma** in IT Business Analysis
Doon, Ontario Conestoga College
- 09/2014 - 06/2018 **B.Tech** in Electronics & Communication Engineering
Hyderabad, JNTU Hyderabad
Telangana