AWANTIKA SRIVASTAVA

Machine Learning Engineer | GenAI, LLMs

+91-8920482037 |sawantika81@gmail.com| LINKEDIN |Github

PROFILE SUMMARY

AI/ML Engineer with 2+ years of experience in building, deploying, and scaling machine learning and GenAI solutions, specializing in LLMs, Retrieval Augmented Generation (RAG), AWS cloud services, and end-to-end AI pipelines with strong exposure to real-world production systems.

CORE TECHNICAL SKILLS

- **Programming & Data Science:** Python (Advanced), Numpy, Pandas, Scikit-learn, FastAPI, C++, REST APIs, Jupyter Notebook.
- **Statistics & Mathematics:** Statistical Modeling, Hypothesis Testing, Confidence Intervals, A/B Testing, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Model Evaluation Metrics (Precision, Recall, F1-score, ROC-AUC).
- **Deep Learning :** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Transfer Learning, Model Fine-tuning, Computer Vision.
- **Frameworks:** TensorFlow, keras, Pytorch, TensorFlow Lite, LangChain, Hugging Face Transformers.
- **GenAI & LLM :** LLMs, RAG, prompt engineering, prompt tuning, embeddings, semantic search, structured outputs, prompt guardrails
- **Databases:** SQL, NoSQL (MongoDB)
- **Cloud & AWS :** AWS S3, Lambda, EC2, ECS/EKS, PostgreSQL, OpenSearch (pgvector), SQS, Secrets Manager
- **MLOps & Deployment :** Docker, Kubernetes, CI/CD pipelines, model versioning, monitoring, logging, scalable AI services
- **Backend & APIs :** FastAPI, REST APIs, async processing, message-driven architectures, backend integration for AI systems

EXPERIENCE

Machine Learning / GenAI Engineer | PPS International Pvt. Ltd. **January 2024-Present**

- Designed and developed **end-to-end machine learning pipelines**, covering data ingestion, preprocessing, model training, evaluation, and deployment.
- Built and optimized **ML models** using Python, scikit-learn, **TensorFlow**, and **PyTorch** to solve real-world business problems.
- Developed **RESTful ML services using FastAPI** for model inference and integration with downstream systems.
- Implemented **GenAI and LLM-based workflows**, including prompt engineering and retrieval-based augmentation.
- Deployed and tested ML solutions on **cloud environments (AWS)**, ensuring scalability, reliability, and performance.
- Applied **MLOps best practices**, including **CI/CD pipelines**, model versioning, monitoring, and automated validation checks.
- Worked closely with global engineering and product teams to translate business requirements into robust technical implementations.
- Supported **peer reviews** and ensured analytical outputs met quality and documentation standards required in regulated environments.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- **Designed** and deployed a real-time **computer vision** ML system for unsafe driver behavior detection using **CNN-based** models (**SSD MobileNet with TensorFlow**).
- Built **end-to-end ML pipelines** including data ingestion, preprocessing, feature extraction, model training, evaluation, and deployment.
- Optimized models using **TensorFlow Lite** to achieve **20–25 FPS real-time inference** with **<150 ms latency** on production/edge environments.
- Exposed model inference through **FastAPI-based REST services** for integration with downstream systems and monitoring tools.
- Performed **error analysis, bias checks, and performance monitoring** to improve model robustness and operational reliability.

Chatbot Using LLM & RAG | Applied ML Project

- Built a Lightweight **LLM-Powered chatbot** using **TinyLLaMA** to answer user contextual questions over structured resume data.
- Implemented a **Retrieval-Augmented Generation (RAG) pipeline**, including document indexing and similarity-based retrieval, to ground LLM responses.
- Applied **prompt engineering and response evaluation techniques** to improve answer relevance, consistency, and factual accuracy.
- Developed **FastAPI-based inference endpoints** to serve the LLM pipeline in a scalable manner.
- Deployed the application using **Streamlit**, focusing on low memory usage, fast inference, and user-friendly interaction.

YouTube Comments Sentiment Analyzer | link - https://youtube-ai-analyzer-ndzqo6r2mepjrtsdjmwaxl.streamlit.app/

- Built an end-to-end **NLP pipeline** for sentiment analysis using **transformer-based models (BERT/DistilBERT)**.
- Performed text preprocessing, feature extraction, model training, and evaluation on large volumes of user-generated data.
- Evaluated model performance using **precision, recall, and F1-score**, and iteratively improved results through error analysis.
- Served the trained model via **FastAPI APIs** for real-time inference and integration.
- Analyzed sentiment trends to generate insights that can inform **product and user-experience decisions**.

CERTIFICATION

- IBM Data Science & AI Certification
- AWS Generative AI with Large Language Models
- OpenCV Computer Vision Certification

EDUCATION

IMS Engineering College, Ghaziabad **September - 2020**

Bachelor of Technology (Electrical and electronics engineering)