AWANTIKA SRIVASTAVA

AI/ML Engineer | GenAI, LLMs

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

PROFILE SUMMARY

AI/ML Engineer with **2+ years of hands-on experience** in designing, developing, and deploying **production-grade AI/ML solutions**. Strong expertise in **Python, Machine Learning, Deep Learning, LLMs, and GenAI workflows** with exposure to **cloud platforms (AWS), CI/CD pipelines, containerization, and scalable ML systems**. Experienced in building **end-to-end ML pipelines, REST APIs, automated testing, and performance-optimized models** aligned with enterprise and regulated environments.

CORE TECHNICAL SKILLS

- **Programming & Data Science:** Python (Advanced), Numpy, Pandas, Scikit-learn, FastAPI, C++, REST APIs.
- **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, TensorFlow, PyTorch, Keras.
- **GenAI & Agentic Systems:** LLMs, RAG, prompt engineering, Model Context Protocol (MCP)**,** LLM gateways, multi-agent architectures, context orchestration, tool calling, structured outputs
- **Cloud & DevOps:** AWS (EC2, S3, SageMaker), Docker, Kubernetes (basic), CI/CD pipelines, Deployment Automation
- **Databases:** SQL, PostgreSQL, MongoDB
- **Testing & Quality:** Unit Testing, Model Validation, Code Coverage, Test Automation
- **Tools:** Git, GitHub, Jupyter Notebook, Google Colab
- **Backend & APIs :** FastAPI, REST APIs, async processing, message-driven architectures, backend integration for AI systems

EXPERIENCE

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

- Worked on **Model Context Protocol (MCP) concepts** for managing LLM context, tool interactions, and agent workflows in GenAI applications.
- Designed and developed **end-to-end AI/ML solutions** covering data preprocessing, feature engineering, model training, evaluation, and deployment.
- Built and optimized **machine learning and deep learning models** using Python, Scikit-learn, TensorFlow, and PyTorch, improving model performance by **20–30%**.
- Developed **scalable ML APIs** using FastAPI for real-time inference and system integration.
- Implemented **LLM-based and GenAI workflows**, including prompt engineering and RAG pipelines for domain-specific use cases.
- Deployed AI solutions on **AWS cloud infrastructure**, ensuring scalability, security, and reliability.
- Applied **CI/CD pipelines, automated testing, and model validation** to support production-ready deployments.
- Collaborated with cross-functional teams and mentored junior engineers while maintaining coding standards and documentation.

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)