

AWANTIKA SRIVASTAVA

AI/ML Engineer | GenAI | MLOps

+91-8920482037 | sawantika81@gmail.com | [LINKEDIN](#) | [Github](#)

PROFILE SUMMARY

AI Engineer with **2+ years of hands-on experience** in building, optimizing, and deploying **machine learning, deep learning, and generative AI solutions** at scale. Strong expertise in **LLMs, AI-driven APIs, edge AI, model optimization, MLOps pipelines, and cloud-based AI systems**. Experienced with **large datasets, feature engineering, distributed training concepts, automation, monitoring, and performance tuning**. Proven ability to work across **data engineering, AI infrastructure, and production ML systems**.

CORE TECHNICAL SKILLS

- **Programming & Backend:** Python, C++, SQL, Numpy, Pandas, Scikit-learn, REST APIs, FastAPI
- **Statistics & Mathematics:** EDA, Statistical Modeling, Descriptive Statistics, Hypothesis Testing, Probability, Sampling, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Reinforcement Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Hyperparameter Tuning.
- **Deep Learning:** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), LLMs, RAG, Generative AI, Model Fine-tuning, TensorFlow, PyTorch, Keras
- **NLP:** Text preprocessing, Tokenization, Chunking, Sentiment Analysis, Transformer-based models, NLTK, Spacy
- **Computer Vision:** Image Classification & Preprocessing, Object detection (YOLO, SSD, MobileNet, ResNet), Video Analytics, OpenCV
- **Edge AI & Model Optimization:** TensorFlow Lite, Model Quantization, Low-Latency Inference, Edge Deployment
- **AI-Driven APIs & Integrations:** AI SDKs & APIs, Partner Integrations, Scalable AI Services
- **MLOps / DevOps:** CI/CD for AI, ML Pipelines, Docker (Exposure), Kubernetes (Conceptual), Automation Tools
- **Databases & Cloud:** SQL, NoSQL, AWS EC2, AWS Sagemaker, GCP
- **AI Analytics, Monitoring & Explainability:** Model Performance Monitoring, AI Observability (Conceptual), Model Explainability & Performance Tuning

EXPERIENCE

AI/ML Engineer | PPS International Pvt. Ltd.

January 2024-Present

- Designed and deployed **end-to-end AI and ML systems** from data ingestion to production deployment.
- Trained and optimized **deep learning and large-scale ML models** using TensorFlow and PyTorch.
- Built **AI-driven APIs** for integration with internal and partner systems.
- Implemented **edge AI solutions** with optimized inference for low-latency environments.
- Worked with **large datasets**, performing feature engineering and model training at scale.
- Contributed to **CI/CD-style ML pipelines** for repeatable and automated deployments.
- Monitored **model performance and reliability**, improving inference stability.
- Collaborated with **data engineers, platform engineers, and product teams**.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- Designed and deployed a real-time computer vision-based ML system to detect unsafe driver behaviors from continuous video streams.
- Trained and optimized CNN-based object detection models (SSD MobileNet architecture) to perform real-time inference on video data.
- Implemented end-to-end ML pipelines covering data ingestion, preprocessing, model training, evaluation, and production inference.
- Achieved 20–25 FPS real-time processing with <150 ms inference latency by optimizing models for deployment.
- Built and integrated a Flask-based web interface to visualize detections and automatically recorded 30-second event clips, reducing manual review effort.

Amazon Stock Price prediction | Applied ML Project

- Developed an LSTM-based time series forecasting model using historical data of 50K+ records to predict future trends.
- Applied data normalization, sequence modeling, and feature selection, improving forecast accuracy by 18–22%.
- Optimized model performance through hyperparameter tuning, reducing validation loss by 20%.
- Evaluated models using RMSE and MAE, ensuring reliable performance on unseen data.
- Built reusable Python ML pipelines for training, evaluation, and inference.

LLM and RAG Based Chatbot | Applied ML Project

- Designed and deployed a Retrieval-Augmented Generation (RAG) based chatbot to answer resume-specific queries with 90%+ response relevance.
- Implemented document chunking, embeddings, and vector similarity search, reducing incorrect responses by 35%.
- Integrated LLM APIs with Python backend services, enabling real-time, context-aware question answering.
- Optimized prompt engineering and retrieval logic, improving answer precision by 25%.
- Built a scalable architecture suitable for production-ready GenAI applications.

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)