Aayush Yadav

AI Engineer / Data Scientist / Machine Learning Engineer

Phone: 7021907344 | Email: aayushdyadav2003@gmail.com
GitHub: github.com/03ADY | Portfolio: aayush-yadav-portfolio.vercel.app/

Summary

Results-driven AI Engineer with expertise in building and deploying scalable machine learning solutions. Proficient in deep learning, MLOps, and full-stack development, with a proven track record of driving business impact through intelligent systems.

Skills

Programming & Core Technologies: Python, SQL, FastAPI, Flask, SQLAlchemy ORM, Docker, Uvicorn, REST APIs, GraphQL
Machine Learning & Deep Learning: Predictive Modeling, Natural Language Processing (NLP), Computer Vision, Reinforcement Learning, Generative AI, Large Language Models (LLMs), TensorFlow, PyTorch, Prophet, Random Forest, Neural Networks, SMOTE
Data Engineering & MLOps: Scalable AI Systems, Model Deployment, Real-time Data Pipelines, CI/CD, Containerization (Docker, Kubernetes exposure), Cloud Platforms (AWS, Azure, GCP concepts), Data Governance, Model Monitoring, Data Ingestion
Databases: PostgreSQL, SQLite, ChromaDB
Tools & Libraries: LangChain, Google Gemini 1.5 Flash, Google Generative AI Embeddings, pypdf, unstructured, python-docx, Plotly, Seaborn, Matplotlib, Whisper, YOLOv5, TensorRT, CUDA, WebSocket, JWT, SendGrid, Razorpay, SHAP, LIME, Git

Experience

Python Backend Developer Intern — NDT Reflection Engineers, Remote *October 2024 – March 2025*

- Developed scalable and robust backend services using Python, FastAPI, and Flask, supporting critical business operations.
- Architected and implemented secure, high-performance RESTful APIs for user authentication, efficient service booking, and automated report generation.
- Oversaw and optimized PostgreSQL database management utilizing SQLAlchemy ORM to ensure data integrity and high performance for application backend.
- Orchestrated end-to-end deployment of containerized applications using Docker across Render and Heroku platforms, ensuring continuous delivery and operational efficiency.

Education

A C Patil College of Engineering, Navi Mumbai *2021 – 2025*
B.Tech - AI & Data Science (Honors in Cybersecurity): GPA: 8.4
Key Courses: Deep Learning, Machine Learning, Data Structures, Databases, Cybersecurity

Projects

AI Assistant (Aura) - RAG Chatbot

- Architected and engineered a production-ready, full-stack Retrieval-Augmented Generation (RAG) system, "Aura," integrating FastAPI, LangChain, Google Gemini 1.5 Flash, and ChromaDB.
- Developed a robust data ingestion pipeline supporting multi-format documents (.txt, .pdf, .docx, .csv), featuring UI-driven file upload and automated, on-demand knowledge base re-indexing.
- Implemented a scalable conversational RAG architecture with efficient semantic retrieval (Google Generative AI Embeddings) and LLM-based generation, ensuring grounded, accurate, and context-aware responses.
- Engineered the high-performance backend API and orchestrated containerized deployment via Docker, establishing a portable foundation for MLOps practices.
- **Impact:** Streamlined enterprise knowledge access and significantly reduced information retrieval time, directly contributing to improved operational efficiency and data-driven decision-making through reliable AI-powered insights.

Multi-Modal Inference Engine

- Engineered a multi-modal ML platform integrating real-time speech-to-text (Whisper) and object detection (YOLOv5) for diverse applications.
- Optimized inference performance by leveraging TensorRT and CUDA, achieving up to 50% faster YOLOv5 processing.
- Developed a high-performance, asynchronous FastAPI backend that reduced API latency by 40%.
- Implemented real-time WebSocket support, enabling seamless audio transcription and live data streaming.
- Maximized GPU resource utilization with dynamic memory optimization and multi-GPU support, enhancing system efficiency and scalability.

AI-Powered Trading System with Risk Analytics

- Developed a deep learning-based market prediction system utilizing RNNs/Transformers for financial forecasting, achieving 18% improvement in prediction accuracy.
- Integrated advanced risk analytics (Sortino Ratio, VaR, Monte Carlo Simulations) into portfolio optimization, resulting in a 22% reduction in portfolio risk.
- Engineered robust, real-time data pipelines for financial data acquisition and preprocessing, enhancing decision-making efficiency by 30%.

Customer Churn Prediction System with API Deployment

- Engineered a robust customer churn prediction model (Random Forest, Neural Networks) with SMOTE for imbalanced data, achieving an 0.87 F1-score and boosting recall by 25%.
- Deployed a high-performance REST API using FastAPI and Uvicorn, enabling real-time predictions in under 100ms for proactive customer retention.
- Designed and implemented data preprocessing pipelines for feature engineering and model consumption.

Hybrid Predictive Maintenance System

- Developed a hybrid predictive maintenance system integrating Deep Learning (RNNs/LSTMs for time-series) and Reinforcement Learning for optimized failure detection.
- Improved failure detection by 30% and reduced downtime costs by 20% through predictive insights.
- Enhanced model interpretability using SHAP and LIME, providing actionable insights for maintenance engineers via an intuitive dashboard.