Yeswanth Sai Yerramasu

AI/ML Engineer

Email: yeswanthsai873@gmail.com| **Tel:** (443) 850 7652

Professional Summary

AI/ML Engineer with **3+ years of experience** developing and deploying **machine learning**, **deep learning**, and **NLP** solutions across diverse domains. Skilled in **Python**, **PyTorch**, **TensorFlow**, and **Scikit-learn**, with expertise in **LLMs** (BERT, GPT), **RAG pipelines**, **recommendation systems**, and **computer vision**. Proficient in building scalable **data pipelines** with **Apache Spark** and **Airflow**, and implementing **MLOps** workflows using **Docker**, **Kubernetes**, **AWS SageMaker**, and **Vertex AI**. Adept at **feature engineering**, **model optimization**, and real-time inference. Strong track record of translating business problems into AI-driven solutions that deliver measurable impact.

Skills

- **Machine Learning & AI:** Supervised/Unsupervised Learning, NLP, BERT, GPT, LLMs, Prompt Engineering, RAG, Image Classification, Segmentation, Metric Learning
- **Data Science & Analytics:** Predictive Modeling, Time Series Forecasting, Statistical Modeling, Hypothesis Testing, EDA, SHAP, LIME
- **Programming & Development:** Python, SQL, PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, OpenCV, Pandas, NumPy, R, C, C++.
- **Big Data & Distributed Computing:** Apache Spark, Hadoop, Data Pipelines, Feature Engineering, Airflow
- **MLOps & Deployment:** Docker, Kubernetes, MLflow, Kubeflow, FastAPI, Flask, CI/CD, Model Monitoring, Quantization, Distillation
- **Cloud Platforms:** AWS(SageMaker), Azure (Vertex AI)
- **Vector Search & Databases:** FAISS, Pinecone, Weaviate, ElasticSearch, PostgreSQL, MongoDB
- **Mathematics & Statistics:** Linear Algebra, Probability, Optimization, Bayesian Methods, Regression Analysis, Experimental Design
- **Visualization & BI Tools:** Tableau, Power BI, Matplotlib, Seaborn, Plotly

Experience

Saigon Technology **Jan 2025 – Current**

AI/ML Engineer

- Designed and fine-tuned **retrieval-augmented generation (RAG) pipelines** using **PyTorch** and **Hugging Face** to deliver context-aware responses from internal knowledge bases, improving response relevance and accuracy.
- Implemented **FAISS-based semantic search** for sub-300 ms document retrieval, enabling **GPT-powered models** to deliver precise, context-driven answers at scale.
- Developed and deployed the chatbot backend with **FastAPI**, containerized via **Docker**, and orchestrated on **AWS** with **CI/CD pipelines** for consistent, production-grade releases.
- Created automated validation scripts to monitor **response accuracy**, **latency**, and **prompt drift**, ensuring conversational quality remained above 90% satisfaction.
- Analyzed user interaction logs and feedback to iteratively refine **prompts** and **fine-tuning datasets**, increasing accuracy from ~78% to over **92%** and reducing response time by nearly half.
- Partnered with product and UX teams to translate customer pain points into technical improvements, resulting in measurable gains in engagement and retention.

Coforge **Feb 2021 - Jul 2023**

Data Scientist / ML Engineer

- Conducted in-depth analysis of historical **transaction data** to identify fraud patterns and engineered **behavioral** and **temporal features** for model training.
- Developed an **XGBoost-based classification model** augmented with **anomaly detection** techniques to flag suspicious activities in near real-time.
- Built a **Kafka–Spark streaming pipeline** to process and score live transactions within milliseconds, enabling proactive fraud prevention.
- Applied **dataset balancing** and **Grid Search hyperparameter tuning** to reduce false positives from ~30% to **8%**, significantly improving investigation efficiency.
- Designed and implemented interactive **Tableau dashboards** to monitor **model performance**, visualize fraud trends, and support investigative decision-making.
- Established an automated **model retraining schedule** leveraging fresh **transaction data**, maintaining fraud detection recall rates above **95%** despite evolving attack patterns.

PROJECTS

Multi-Agent Generative Model for Personal Finance & Investment Guidance

- Built a multi-agent LLM system where specialized agents performed budgeting, stock trend analysis, sentiment extraction, and government data parsing.
- Implemented a central manager agent to coordinate agent outputs and deliver coherent, actionable investment recommendations.
- Trained models with custom financial datasets, integrated RAG pipelines using Hugging Face Transformers, and deployed via Streamlit for interactive use.

Research Paper Summarization Model

- Designed a hybrid classification + generative model to summarize academic papers based on tone, depth, and audience.
- Applied LLMs (GPT, Gemini) with prompt tuning, semantic filtering, and embedding layers for tailored summarization.
- Delivered a customizable UI allowing users to adjust summary goals in real time, backed by vector embeddings and document classification.

Ai - restaurant desk attendant (Voice to voice model project)

- Developed a real-time LLM-based voice system trained on menus from 400+ restaurants to handle automated customer calls.
- Converted spoken orders into structured text, generating receipts, kitchen notifications, and bills with integrated order tracking.
- Enabled AI voice interactions that responded naturally to customer queries, improving order accuracy and reducing manual desk attendant workload.

EDUCATION

- **Masters of Professional Studies in Data Science**
  University of Maryland Baltimore County (UMBC) Maryland, USA
- **Bachelor of Engineering in Computer Science**
  R.V.R & J.C College of Engineering, India

CERTIFICATIONS

- **Data analytics by python**
- **Problem solving through programming in C**
- **Data analysis with R**
- **Google Analytics**