Deepak Rambarki

AI/ML Engineer

rambarkideepak@gmail.com | +1 (940) 843-7994 | LinkedIn |

SUMMARY

AI/ML Engineer with 5+ years of experience building and deploying production machine learning systems for fraud detection and personalization at scale (3M+ daily transactions). Delivered measurable business impact including 19% fraud reduction and 28% revenue uplift through real-time inference, advanced modeling, and MLOps on AWS. Hands-on expertise in Python, deep learning, feature engineering, and scalable cloud-native architectures. Experienced in integrating Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and agentic workflows into intelligent decision-support systems. Strong ownership across the full ML lifecycle from model development to deployment, monitoring, and governance.

SKILLS

Programming: Python (Advanced), SQL, R, Java, C++, OOP, Data Structures & Algorithms

Machine Learning: Classification, Regression, Anomaly Detection, Time Series Forecasting, XGBoost, LightGBM, TensorFlow, PyTorch, CNNs, LSTM/BiLSTM, Transformers, Transfer Learning, Hyperparameter Optimization (Optuna, Bayesian)

Generative AI & NLP: Large Language Models (LLMs), GPT, BERT, T5, Retrieval-Augmented Generation (RAG), Prompt Engineering, Embeddings, Semantic Search, Hugging Face, spaCy

Data & Big Data: Apache Spark (PySpark), Kafka, ETL Pipelines, Feature Engineering, Snowflake, Redshift, MongoDB, Cassandra

MLOps & Deployment: AWS SageMaker, MLflow, CI/CD for ML, Model Monitoring, Data Drift Detection, Docker, Kubernetes, Real-Time Inference APIs

Cloud Platforms: AWS, Azure, GCP, AWS Lambda

Evaluation & Responsible AI: AUC-ROC, F1-Score, Precision@K, SHAP, Model Interpretability, AI Governance, GDPR/HIPAA Compliance

EXPERIENCE

AI Engineer, Brex **Dec 2024 – Present**

- Designed and deployed a real-time fraud detection system processing **3M+ daily transactions**, reducing fraud by **19%** using BiLSTM-based behavioral sequence modeling.
- Architected end-to-end ML pipelines from data ingestion to production inference using **AWS SageMaker, Snowflake, and microservices architecture**.
- Built scalable feature engineering pipelines with **Python, SQL, and PySpark**, processing 10M+ records and managing 15TB+ structured data.
- Improved model performance to **0.87 F1-score / 0.93 AUC-ROC** through hyperparameter optimization, ensemble modeling, and feature selection.
- Explored **agentic AI frameworks for automated case triaging**, enabling multi-step reasoning with tool-calling and structured outputs.
- Automated retraining, model versioning, and monitoring using **MLflow, CI/CD, and data drift detection**, ensuring production reliability and model freshness.
- Deployed containerized ML services with **Docker, Kubernetes, and AWS Lambda**, achieving scalable low-latency real-time inference.
- Implemented **SHAP-based explainability and AI governance controls**, supporting audit readiness and regulatory compliance standards.

Machine Learning Engineer, Zebronics | India **Jan 2020 – June 2023**

- Developed a hybrid recommendation engine (collaborative + content-based filtering) increasing cross-sell revenue by **28%** across 3M+ users.
- Designed personalized ranking algorithms leveraging behavioral data and embeddings to improve recommendation accuracy by **20% over baseline**.
- Built demand forecasting models using **XGBoost, LSTM, ARIMA, and Prophet**, reducing overstock by **18%** and stockouts by **22%**.
- Designed ML-powered REST APIs integrated into web and mobile platforms, improving conversion rates by **10%** and session duration by **12%**.
- Conducted structured **A/B testing and experiment design**, driving 6% retention growth and measurable revenue uplift.
- Developed real-time ML pipelines using **Azure ML, Kafka, Spark Streaming, Docker, and Kubernetes**, enabling scalable and automated model deployment.
- Optimized model inference performance and resource utilization, reducing infrastructure costs while maintaining high availability.
- Partnered with product and business stakeholders to translate KPIs into measurable ML objectives, aligning AI initiatives with revenue growth strategy.

EDUCATION

The University of North Texas | Denton, TX, USA **Aug 2023 – May 2025**

Master in Data Analytics

PROJECTS

MEDCOMPARE – AI Powered Medication Data Benchmarking Tool

- Built a real-time Streamlit tool benchmarking medication data from LLMs using semantic similarity, fuzzy matching, and biostatistics;aligned outputs with FDA and RxNorm standards; supported batch evaluation, scoring, and FHIR exports for clinical review.

Classification with XGBoost + ONNX Deployment

- Developed an XGBoost classifier on 1.2M entries, outperforming baselines; applied scaling, outlier handling, and data prep; converted model to ONNX for real-time edge/cloud deployment, showcasing scalable predictive analytics in IoT and embedded environments.