PRANSHU GOYAL

(571) 618-4567 • NY • pg2592@nyu.edu • LinkedIn • GitHub • Portfolio

Machine Learning Engineer with 2+ years of experience building scalable ML systems and real-time data pipelines. Specialized in fine-tuning LLMs, deploying RAG architectures, and optimizing inference on AWS, Docker, and Kubernetes.

EDUCATION

New York University **Sept 2023 - May 2025**
Masters in Computer Science (Recipient of Merit-based scholarship) 3.77/4.00
Thapar Institute of Engineering and Technology, India **July 2018 - June 2022**
Bachelors in Computer Engineering

WORK EXPERIENCE

Machine Learning Graduate Assistant, NYU, NY **Sep 2024 - Present**

- Supported implementation of ML models (logistic regression, Support Vector, Random Forest, XGBoost) using Python.
- Evaluated models using F1-score, ROC, AUC and cross-validation to reinforce best practices in model development.

Machine Learning Intern, LOCOMeX, NY **June 2024 - Aug 2024**

- Implemented a **GenAI RAG pipeline (LangChain + OpenAI GPT-3)** automating ESG data extraction from contracts and improving data processing efficiency by 60%.
- Trained and deployed XGBoost and Random Forest models on **SageMaker & EKS, achieving 80% predictive accuracy.**
- Built **ETL pipelines using AWS** (S3+Lambda+API Gateway) improving data reliability & **reducing processing time by 35%.**

Software Development Engineer, ION Trading, India **Jan 2022 - June 2023**

- Delivered **5+ clients requested paid enhancements** as part of an **Agile team** by contributing across the SDLC - requirement analysis, system design, development, and test automation (unit tests, ATDD).
- Led development of **15+ features for global banking clients** (UBS, TD Bank, Scotia) using **Java, C++, TypeScript** in a microservices based architecture.
- Prototyped and evaluated **ML based sales trade approval** prioritization using historical data, **reducing decision latency.**
- Designed **Java REST APIs** ensuring high-throughput trade booking across **distributed systems.**
- **Reduced release time by 12%** via Docker based builds and **Jenkins CI/CD optimization.**
- Resolved high priority client issues by analyzing logs and deploying fast, stable, tested fixes to maintain 99% uptime.

Machine Learning Research Assistant, TIET, India **Oct 2021 - April 2022**

- **Led a team of 4** to design and train CNN and SOTA Deep Learning models for fabric classification and defect detection, achieving **90%+ accuracy and enhancing quality** control in textile manufacturing. **[Springer Paper]**
- Built a robust ensemble model (XGBoost + CatBoost + RF) for music genre classification, improving **classification accuracy by 15%** and contributing to research in **audio based ML**. **[Springer Paper]**

Machine Learning Intern, PAYTM, India **July 2021 - Aug 2021**

- Improved liveness detection accuracy to **93.4%** by integrating MobileNetV2 and **self engineered CNN** models.
- Developed Flask based OCR application using Tesseract for text extraction; deployed **fine-tuned BERT** for named entity recognition on **Kubernetes to streamline online KYC processing.**
- Set up **real time model observability** using NewRelic, Prometheus & Grafana, in compliance with **MLOps best practices.**

TECHNICAL SKILLS

Machine Learning and AI : PyTorch, Tensorflow, LangChain, Hugging Face, Unsloth, Onnx, QLoRA, CUDA
Big Data and Analytics : PySpark, Scala, Vector Database, Grafana, Kafka, PowerBI, Dask, Tableau, Airflow
Programming Languages and Frameworks : C, C++, Java, Python, R, SQL, Angular, Flask, Typescript, PostgreSQL
Cloud and CI/CD : AWS, Docker, Kubernetes, Jenkins, Azure, Google Cloud Platform
Relevant Coursework: Reinforcement Learning, Natural Language Processing, Machine Learning, Big Data, Deep Learning, Predictive Analytics using Statistics, Optimization Techniques, Probability and Statistics, Computer Vision

PROJECTS

Real Time Forex Arbitrage and Price Prediction System [code]

- Built a **real-time forex trading system** using **Spark** for arbitrage detection and a bi-directional **LSTM** for price prediction. Integrated Flask APIs for serving, automated data workflows with **Airflow**, and enabled live monitoring via Looker Studio.

NYUAssistant [code]

- Built an **GenAI query system** for the NYU community using a **RAG** pipeline with instructor embeddings, **FAISS**, and **FlashRank**. Integrated **LangChain** with **LLaMA3-8B** for answer generation and deployed a Streamlit web app.

Sentiment Insight for Finance [code]

- Fine-tuned LLaMA3-8B and Mistral-7B using **LoRA** and **4-bit quantization** for financial sentiment analysis, boosting accuracy to 84.3% (Mistral) and 80.6% (LLaMA), with Mistral outperforming across key metrics.