NAVIN KUMAR M

Chennai - 600053, Tamil Nadu, India
PhoneNo | Email | LinkedIn | Github | Website

Machine Learning Engineer with 2+ years of experience developing scalable, production-grade AI systems, including credit risk models, large language models, training infrastructure, and high-performance inference systems. A pragmatic problem solver with a strong passion for applying AI to deliver impactful solutions in fintech and enterprise domains, seeking opportunities to further advance trustworthy and large-scale AI applications.

WORK EXPERIENCE

Branch International Remote, India
Software Engineer – Machine Learning B1 Jun 2024 – Present

- Driving an ambitious neural modeling effort to extract credit signals from SMS-based financial behavior.
- Accelerated feature fetch 8× through a distributed head–worker system, fine-tuned service worker and thread-pool configurations, optimized SQL queries, and batched I/O- & CPU-intensive requests.
- Developed a credit model for premium user segment, achieving a delinquency rate of 2%, at 15% APR reduction and ensuring model stability.
- Built fully automated training infrastructure on AWS SageMaker, reducing model training time and costs by 5×.
- Overhauled the hiring process, thereby broadening the talent funnel and guiding cross-team execution of critical, high-impact initiatives.
- Refined training sample queries and introduced shorter window targets, enabling the use of recent data and improving data confidence & generalizability across all user segments.

Software Engineer - Machine Learning Intern Jan 2024 – May 2024

- Developed cashflow and balance classification & extraction SMS model with AUC of 0.99, increasing coverage by 3x.
- Designed product-specific features from credit bureau data, resulting in a 6% improvement in the AUC of the model.
- Ensured 99.5% data consistency through rigorous QA & QC testing of credit bureau data in multiple formats.

Vellore Institute of Technology Chennai, India
Research Intern Nov 2023 – Apr 2024

- Built a distributed training and inference workspace on IBM PowerPC (ppc64le) using Ray, which allowed 15+ concurrent model experiments for student and professor projects
- Deployed "TutorAI," a distributed learning AI platform mentoring 50+ university students, enhancing coursework guidance and academic engagement.

Gutsy Innovation Remote, India
Machine Learning Engineer Intern May 2023 – July 2023

- Implemented YOLOv7 and ArcFace for real-time face detection & recognition in live video streams.
- Integrated NVIDIA DeepStream SDK with Azure IoT Hub and Qdrant to process 5+ concurrent camera feeds with vector-based face matching.

EDUCATION

Vellore Institute of Technology Sep 2020 - Apr 2024
B.Tech in Computer Science & Engineering, Specialization in Artificial Intelligence & Machine Learning
CGPA : 8.92 / 10

Relevant Coursework: Machine Learning, Deep Learning, Data Structures, Algorithms, Operating System, Security, Networks, DBMS, Linear Algebra, Statistics, Software Engineering, Product Development & Entrepreneurship.

PROJECTS

AI Learning Platform (github)

- Fine-tuned LLaMA-7B models using SFT and RLHF on a synthetic student Q&A dataset, enabling adaptation for domain focused tasks.
- Designed and deployed an LLM-RAG pipeline that utilizes vLLM for high-throughput inference, integrated with Qdrant and LangChain to provide contextual real-time doubt resolution.
- Built a production-ready backend with Django, storing user profiles in PostgreSQL and chat/feedback data in Cassandra, ensuring scalability for multi-user learning environments.

AI Malware System (github)

- Developed a Cross-Platform Application that uploads files to an AI-driven backend for identifying malware with an Azure PowerBI dashboard to visualize the insights in real-time.
- Employed CoAtNet Transformer model pre-trained on 2 million malware binaries, and conducted advanced feature analysis using LSTM and LightGBM based on logs from Cuckoo Sandbox and malware header information.

Benetech - Making Graphs Accessible (github)

- Visual Reasoning from charts and plots to make them accessible for visually impaired people
- Utilized EfficientNetv2, Matcha (Pix2Struct) for analyzing bar, line, and dot plots, and Faster-RCNN for scatter plot derendering.

SKILLS

Programming Languages:	Python, SQL, C++, Go, Rust
Frameworks & Infrastructure:	PyTorch, Ray, AWS Cloud, Git, Kubernetes, Qdrant, FastAPI, Django, Snowflake, Metabase, SageMaker, Airflow, RabbitMQ
Technical Expertise:	Credit Risk Modeling, Natural Language Processing, Large Language Models, Time-Series Forecasting, Linux, Cloud, DevOps, High- & Low-Level Design, Training & Inference Infrastructure

ACHIEVEMENTS

- Runner, Branch Hackathon – For Building a GenAI Credit Risk Model Jan 2025
- Winner, Best Final Year Project, VIT University, for developing an on-premise AI learning platform Apr 2024
- Winner, Branch Hackathon – For developing an AI Agent to assist engineers with documentation Jan 2024