Nikhil Naik

Austin, TX 78717 | nnaik2@stevens.edu | LinkedIn | GitHub | Medium | +1 (917)282-6968

SUMMARY

Machine Learning Engineer with 3+ years of specialized experience in Generative AI, large language models, and software development. Proven track record in fine-tuning LLMs, developing RAG systems, and building scalable cloud-based solutions. Successfully led cross-functional teams, driving projects that helped secure multi-million-dollar contracts. AWS Certified Solutions Architect with extensive deployment experience on Nvidia Triton Inference Server.

EDUCATION

Stevens Institute of Technology; Hoboken, New Jersey **May 2024**
Masters of Science in Machine Learning CGPA: 3.96 / 4.0
Courses: Deep Learning, Big Data Tech, 3D & 2D CV, NLP, Knowledge Discovery and Data Mining

PROFESSIONAL EXPERIENCE

Machine Learning Engineer – GenAI Intern -*Medium Article*
LifeSage Inc; Austin, USA **July 2024 - Present**

- Developed a Healthcare LLM assistant using PEFT fine-tuning on Llama3 and Mistral MOE with LlamaFactory
- Achieved 3x faster fine-tuning with FlashAttention and Unsloth for quantized LORA
- Tested model deployment using TensorRT and Triton Inference Server, handling 40 concurrent async responses
- Enhancing performance and reliability with a LangChain-powered Agentic RAG system

Software Developer & Machine Learning Engineer
Gajshield Infotech Firewalls; Mumbai, India **Sept 2018 – Mar 2021**

- Developed in-house malicious email detecting system based on Scikit-learn's outlier detection model with 96% recall improving security systems efficiency by 40%
- Developed a Technical Assistant Chatbot using RASA to help the operation support team efficiently resolve customer issues, resulting in a 30% reduction in the engineering team's ticket and workload.
- Led a 4-member team that developed two industry-scale Cloud-managed VPN and SD-WAN projects using open-source Linux libraries and MySQL. Instrumental in securing a major tenure worth $6 million
- Enhanced and modernized frontend of Firewall's UI, utilizing ExtJS, Pygal, and Chartjs
- Served as technical representative at stakeholder meetings, collaborating with clients to architect the deployment of over 3000 Firewalls and cloud-based management

TECHNICAL PROJECTS

Multilingual Translation System – Workshop on Machine Translation 2024 **July 2024**

- Representing Stevens at an international conference with novel multilingual machine translation research
- Developed an iterative back-translation finetuning pipeline for NLLB and MT-5 models, raising BLEU score to 59.8
- Preparing a comprehensive research paper for WMT 2024, detailing innovative methodology and significant results

Forex Rate Prediction – Deep Learning *-GitHub* **Mar 2023**

- Architected a baseline RNN & LSTM neural network with In-Domain world Forex dataset using TF and Keras
- Hyper parameter tuned those models and used transfer learning fine-tuning on target data to achieve 0.45 R2 score

CitiBike Usage Forecast – Big Data ML *-GitHub* **Apr 2023**

- Leveraged PySpark on Google Cloud Platform's Dataproc to create ETL pipeline on 40GB time-series dataset
- Utilized Spark's MLlib to train a distributed multivariate regression model achieving a 0.85 R2 score

Telemarketing Campaign Prediction *-GitHub* **Sept 2021**

- Performed Exploratory Data Analysis to gain data insights, selected and tuned best Machine Learning algorithm
- Awarded excellence certificate for achieving an accuracy of 95% employing KNN Classifier with silhouette analysis

TECHNICAL SKILLS & ACCOMPLISHMENTS

Certifications:	**AWS: Solutions Architect Associate - *Cert***	IIT Roorkee's Data Science - *Cert*
	Fundamentals of Data Mining - *Cert*	Algorithmic Toolbox, Coursera - *Cert*
Tech Stacks:	Python, C++, MLOps, Shell, JavaScript, MySQL, SQLite, MSSQL, JSON	
Frameworks:	TensorFlow, PyTorch, Keras, Fairseq, HuggingFace, LangChain, OpenCV, Spark, Django, Flask	
Other Libraries:	Scikit-Learn, Transformer, Amazon Web Services, SageMaker, S3, EC2, Lambda, GCP, Docker, Kafka, Linux (CentOS), Apache Web Servers, Full-Stack, SDLC, Nvidia-Triton, Kanban, Agile	
Publication:	Sign Language Interpreter Glove | IEEE ICATE-2K18 Event | ISBN: 978-93-5267-422-0	