Harsha Vardhan Reddy Tamatam

• harshavardhan.io • linkedin.com/in/harsharshavardhantamatam/ • harsha.vardhan@rutgers.edu • +1 617-642-4707

EDUCATION

RUTGERS UNIVERSITY, NEW BRUNSWICK, NJ, USA　　　　　　　　　　　　　09/2023 - 05/2025
Master of Science in Data Science, Dept of Statistics

MAHINDRA ECOLE CENTRALE, HYDERABAD, INDIA　　　　　　　　　　　　08/2016 - 05/2020
Bachelor of Technology in Computer Science

WORK EXPERIENCE

DATA ANALYST INTERN - LS DIRECT, NY, USA　　　　　　　　　　　　　　06/2024 - 12/2024
- Redesigned an existing imbalanced classification model by incorporating text field embeddings and new features, then deployed and maintained it in production, improving recall and boosting ad targeting efficiency by 4%.
- Created vector data stores for 15M+ users and implemented Azure AI Search to power high-accuracy semantic retrievals.
- Built scalable ETL pipelines using Airflow, dbt, and GCS to automate SQL-to-BigQuery migration, integrating version control, data validation, and scheduling to ensure seamless, reliable data flow.

PRODUCT ASSOCIATE - VEDANTU LEARNING, INDIA　　　　　　　　　　　　12/2021 - 08/2023
- Fine-tuned and deployed a DistilBERT model with PyTorch and FastAPI to intelligently classify and map student queries(or tickets) to tutors in real-time, reducing response times by 65% and boosting chat engagement by 1.5X.
- Developed XGBoost models on student behavioral data to identify high-risk churn users, and developed dashboards to help sales and leadership teams to design targeted retention strategies.
- Built dashboards to optimize the ad funnel across recall, rough-sort, and fine-sort stages, tracking ranking quality, personalization impact, and drop-off metrics to guide delivery strategy.

DATA SCIENTIST - KHABRI AUDIO(YC STARTUP), INDIA　　　　　　　　　　11/2020 - 12/2021
- Built and deployed a deep learning-based recommendation and ranking model, combining FastText embeddings and behavioral features, which served personalized content to 150K+ daily active users, improving CTR by 12%.
- Designed A/B tests and offline evaluation pipelines with PMs and Data Scientists, optimizing search and recommendation models using metrics like dwell time and CTR.

MACHINE LEARNING ENGINEER (Co-Founder) - AUTODUB, INDIA　　　　　04/2018 - 11/2020
- Founded and led Autodub.in, developing AI-powered speech translation software to enable multilingual accessibility.
- Scaled AI infrastructure from scratch, enabling rapid video dubbing at 90% lower cost and near-human accuracy.
- Autodub's technology was acquired by Khabri Audio after national media recognition for dubbing COVID-19 awareness content in Indian languages.

SKILLS

- **Programming:** Python, Pytorch, HF-Transformers, LangChain, FastAPI, Django, Streamlit
- **AI & ML:** Gen-AI, Machine Learning, Deep learning, NLP, LLMs, RAG, ML System Design, Data Modeling
- **Databases & Cloud:** SQL, MongoDB, Pinecone, BigQuery, Redis, Airflow, GCP, AWS, Kubernetes, CI/CD

RESEARCH EXPERIENCE & PROJECTS

FOUNDATION MODELS EVALUATION & USE-CASE MAPPING　　　　　　　01/2025 - Present
AI ETHICS LAB, RUTGERS UNIVERSITY　　　　　　　　　　　　　　Advisor - Dr. Mhamed Bettaieb
- Co-Authoring a research study on existing foundation models, comparing architectures and applications across domains.
- Developing a structured model selection framework, guiding optimal LLM choices for different real-world scenarios.

CSHORTS - AI NEWS SUMMARIZATION PLATFORM - Website
- Fine-tuned and quantized a T5 model for abstractive summarization, reducing model size by 45% and optimizing for low-cost, scalable inference; deployed the model on AWS SageMaker.
- Built a real-time FastAPI backend and Streamlit dashboard to ingest unstructured news feeds; implemented text extraction workflows via PDF parsers and transcription tools, and store the processed text in MongoDB.

MOOLAH - PERSONAL FINANCE AI AGENT - Website
- Fine-tuned Llama 3.1-70B with LoRA adapters and served it on an Amazon SageMaker Large-Model-Inference (Inf2) endpoint to power an AI personal-finance agent that tracks discretionary spending and delivers real-time budget insights.