NAGHARJUN MATHI MARIAPPAN

917-688-8843 • nm4074@nyu.edu • Portfolio • LinkedIn • Github • New York City, NY

TECHNICAL SKILLS

Languages: Python (Pandas, Numpy, Scikit-learn), C++, SQL
Tech Stack: PyTorch, TensorFlow, MLflow, Apache Toolkit (Airflow, Spark), Docker, Kubernetes, Git, Tableau, Linux
Cloud Platforms: AWS - EC2, S3, Lambda, Bedrock, Redshift, ECS; Azure - VMs, Functions, Cognitive Services
Machine Learning: Classification, Regression, Clustering, Hypothesis Testing, LLMs, CNN, LSTM, Attention, PCA

PROFESSIONAL EXPERIENCE

Data Scientist July 2024 - Present
Mount Sinai Health System, New York, NY

- **Reconstructed** missing respiratory signals in polysomnography studies at **84% recall** and **75% precision** by building a Dilated Residual CNN with Attention, leveraging plethysmography data for sleep apnea patients.
- **Restored** MLflow platform within 24 hours with **100% data integrity** by creating custom **backup/restore** pods and automating nightly MinIO backups via Kubernetes CronJobs after a service failure.
- **Improved** team model tracking, logging, experimentation, and versioning by **30%** by deploying **MLflow** on Ubuntu Server, accelerating workflow efficiency and reproducibility.
- **Reduced** clinical data extraction **costs 40%** by optimizing token sizes. Employed Retrieval Augmented Generation(**RAG**) and Cosine Similarity based nearest neighbor grouping to minimize redundant text chunks.
- **Prevented 98%** of patient data leaks via a zero-shot GLiNER Named Entity Recognition(**NER**) model, censoring personal details before forwarding clinical notes to LLMs on AWS Bedrock.

Data Science Intern January 2024 - May 2024
LOCOMeX, New York, NY

- **Boosted** contract **engagement 40%** with a two-tower deep learning **recommender system**, incorporating BERT text embeddings and company metadata for relevance scoring.
- **Reduced** processing **time 90%** while extracting structured data from PDFs, DOCXs, and PPTs. Built an **AWS Lambda** API leveraging GPT-3.5.
- **Saved** 8 weekly **hours** by automating data pipelines with **Apache Airflow** and Python, storing processed documents on AWS Redshift and S3.

ACADEMIC PROJECTS

AdEvade: Outsmarting ML Ad Blockers - Link December 2023 - March 2024

- Trained a **YOLOv5**-based ad blocker at **90% mean IoU** for detecting ads. Demonstrated a **15% drop in detection** by performing adversarial image manipulations and successfully retrained the model to restore performance.

CitiBikeFlow: NYC's Cycling Visuals - Link December 2023 - February 2024

- Built a real-time **data pipeline** for Citi Bike feeds using **AWS S3 + Snowpipe** into Snowflake. Orchestrated ingestion and transformations with Apache Airflow on **AWS EC2**, powering Tableau dashboards.

NYC StationSense: Predicting Crowds - Link February 2023 - May 2023

- **Parallelized** transformations on 11M+ subway crowd records with Apache **Spark**. Developed forecasting ML models in **SparkML (R2 = 0.62)**, enabling accurate crowd-level predictions.

OPEN SOURCE CONTRIBUTIONS

Exporting MLflow Experiments from Restricted HPC Systems - Link April 2025

- Published a Towards Data Science article on secure MLflow experiment migration from restricted HPC clusters.

EDUCATION

New York University, New York, NY September 2022 - May 2024
Master of Science in Computer Engineering GPA: 3.96/4
Relevant coursework: Big Data, Machine Learning, Deep Learning, DS and Algorithms, Image Processing, Databases