Nagharjun Mathi Mariappan

Brooklyn, New York, USA
(917)-688-8843 nm4074@nyu.edu linkedin.com/nagharjun17 nagharjun17.github.io

Skills

Languages: Python (Pandas, Numpy, Scikit-learn), C++, SQL
Tech Stack: PyTorch, TensorFlow, AWS, Azure, Docker, Git, Apache Toolkit, Tableau, Linux
Machine Learning: Classification, Regression, Clustering, LLMs, CNN, LSTM, Transformers, PCA

Experience

Data Science Analyst | Python, PyTorch, AWS, Pandas **July 2024 - Present**
Mount Sinai Health System New York, NY

- Achieved 84% recall and 75% precision on test set by forecasting every second of missing respiratory flow signals as flow or no flow in polysomnography sleep studies. Trained a Dilated Residual CNN with Additive Attention using plethysmography signal as the dynamic covariate feature of sleep apnea patients.
- Optimized clinical note extraction through LLMs by reducing prompt size by 70% and costs by 40%. Achieved this using a Cosine Similarity-based Nearest Neighbor Search to group and consider semantically similar text chunks.
- Developed a tool to assist doctors in detecting fixed airway obstruction, achieving an F1 Score of 89.3%. Built an LSTM-based multiclass classifier that analyzed spirometry respiratory flow-volume data from 187 patients.
- Secured patient details by reducing data leaks by 98%. Implemented a pretrained zero shot GLiNER to extract and censor personal details before forwarding clinical notes to Llama3.1 405B on AWS Bedrock for symptoms extraction.
- Eliminated 5 hours of weekly overhead by automating downloads of previous night sleep studies. Implemented a Python script with SQL commands, scheduled via Task Scheduler, to download polysomnography and EEG studies.

Data Science Intern | Python, OpenAI, Apache Airflow, AWS **January 2024 - May 2024**
LOCOMeX New York, NY

- Improved contract engagement by 40% by engineering a two-tower deep learning model for contract recommendation, incorporating BERT-based text embeddings, company metadata, and binary cross entropy loss to optimize contract relevance prediction.
- Automated data integration and contract document scraping, saving 8 hours of weekly effort. Developed an ETL pipeline with Python and Apache Airflow, storing processed data in multiple formats on AWS Redshift and S3.
- Extracted structured data from unstructured sources with 95% accuracy, reducing processing time by 90%. Designed an API endpoint on AWS Lambda using GPT-3.5 to extract data from PDFs, DOCXs, and PPTs.

Projects

CitiBikeFlow: NYC's Cycling Visuals | Python, Apache Airflow, Snowflake **December 2023 - February 2024**

- Engineered a **real-time data pipeline** aggregating Citi Bike data from multiple APIs to AWS S3. Automated data ingestion into Snowflake with Snowpipe, facilitating dynamic Tableau visualizations orchestrated by **Apache Airflow** on AWS EC2.

NYC StationSense: Predicting Crowds | Python, Apache Spark, MongoDB **February 2023 - May 2023**

- Parallelized the data transformation phase of over 11 million subway crowd records using the distributed **Apache Spark** framework. Further constructed crowd forecasting ML models using **SparkML** with 0.62 Mean Squared Error.

Education

New York University **September 2022 - May 2024**
Master of Science in Computer Engineering **(GPA: 3.96/4)** New York, NY
Courses: Big Data, Machine Learning, Deep Learning, Algorithms and Data Structures, Image Processing, Databases

Leadership

Teaching Assistant - Python for Bioinformatics **January 2023 - May 2023**
New York University

- Led weekly doubt-clearing sessions for 25+ students, addressing challenges in Python programming.
- Enhanced learning outcomes by updating and refining course assignments and materials for improved quality.