Lokesh Kanna Rajaram

rajaramlokeshkanna@gmail.com | Portfolio | LinkedIn | +1(716) 520 7064| GitHub

Summary

Data Scientist with hands-on experience building scalable analytics and ML pipelines across computer vision, NLP, and time-series data. Proven ability to translate large, noisy datasets into actionable insights through statistical modeling, distributed data processing, and production-style experimentation. Experienced working with Python, SQL, Spark, and cloud-based data systems to support data-driven decision-making.

TECHNICAL SKILLS

Programming Languages & Databases: Python, R studio, R Shiny, MySQL, Pytorch, Hadoop, Apache Spark, Kafka, Pandas, MongoDB, PostgreSQL, NoSQL, Informatica, AWS.
Tools & Platforms: PowerBI (DAX), Tableau, Docker, GitHub/Git, Microsoft Office, Generative AI, LLM, Amazon Web Service (AWS), Redshift, ETL Pipelines, CI/CD Pipelines, Statistical Analysis, Cloud Infrastructure, Data Visualization, Matlab.
Certifications: Informatica Cloud (IICS), Cloud Practitioner – CLF-CO2.

EDUCATION

University at Buffalo, The State University of New York, NY, USA **Dec 2025**
Master of Science, Data Science
Coursework: Statistical Data Mining, Data Intensive Computing, Introduction to Machine Learning.

WORK EXPERIENCE

Research Assistant – University at Buffalo (SMILE lab), Buffalo, NY **Dec 2025 - Present**
- **Built self-supervised preprocessing pipelines to denoise and enhance ultrasound image data, improving data usability for downstream modelling tasks.**
- **Applied generative modelling techniques to learn mappings between low- and high-quality sensor data distributions, increasing robustness under limited-label conditions.**
- Developed data-centric validation workflows to support modelling on **high-variance, sensor-generated datasets**.

Data Scientist Intern (Capstone) - Nissha Medical Technologies, Buffalo, New York **Aug 2025 - Dec 2025**
- **Designed and deployed a real-time computer vision pipeline using YOLOv8, OpenCV, and PyTorch to detect manufacturing defects under strict latency constraints, reducing unplanned machine downtime by 12%.**
- **Engineered a hybrid ML + Statistical Process Control monitoring system, combining deep learning-based object detection with rolling-window SPC analytics, early failure detection and reducing defect rates from 4% to 3%.**
- **Built** a production-style analytics workflow using **Pandas and SciPy** to track **temporal quality drift (dimensional and colour metrics, ΔE)**, **transforming raw model outputs into actionable predictive maintenance signals**.

PROJECTS

Generative Models Benchmarking GenAI Models: Gans, Vaes, And Diffusion Models **Aug 2025**
- **Designed and analyzed large-scale ML experiment datasets** across multiple image benchmarks (**MNIST, CIFAR-10, CelebA**), **defining and tracking performance KPIs (FID, Inception Score)** to compare model quality and stability.
- **Led experimental analysis and benchmarking**, **evaluating convergence trends, variance, and output diversity** across generative models, and **translating technical metrics into clear comparative insights**.
- **Built automated analysis and visualization workflows**, **standardizing metric computation, logging, and reporting**, and **reducing experiment analysis time by ~30%** through reusable notebooks and scripts.

Amazon Book Review Using Big Data Pipeline **May 2025**
- **Built an end-to-end big data analytics pipeline** using **Hadoop and PySpark** to **ingest, clean, and transform 1M+ Amazon book reviews**, enabling large-scale feature engineering and model-ready datasets.
- **Conducted scalable experimentation and evaluation**, **validating model performance across distributed datasets** and translating results into actionable insights on sentiment trends and review quality.
- **Engineered NLP features at scale** using **Spark ML (Tokenizer, StopWordsRemover, HashingTF, IDF)**, **improving model training efficiency by 40%** while achieving **90.4% classification accuracy**.

Optimized Bulk Stock Selling Strategies With Machine Learning **Dec 2024**
- **Built and evaluated predictive models** (Random Forest, Gradient Boosting, LSTM) on **4 years of NVIDIA market data (~1,000+ trading days)**, **improving forecast accuracy by 15%** and **increasing simulated strategy returns by ~8–10%** over baseline approaches.
- **Designed execution and risk-aware analytics** using **RSI, Bollinger Bands, VWAP, and TWAP**, **reducing adverse price impact during bulk trades by ~6%** in backtesting through regression, clustering, and time-series visualization.