Vinay Ram Gazula

Newark, NJ | 908-552-1879 | gazulavinayram@gmail.com | vinaygazula.dev | linkedin.com/in/vinayramg

Professional Summary

MS **Data Science** graduate with strong research background and multiple publications focusing on machine learning and explainable AI. Holds over 3 years of experience as Data Engineer designing, building, and maintaining scalable data pipelines to enhance data quality and drive business growth. Proficient in data modeling, data analysis, data warehousing, and cloud technologies. Effective at collaborating with cross-functional teams.

Education

New Jersey Institute of Technology | Newark, NJ — 2023 — 2025

*Master of Science in **Data Science*** — GPA: 3.89/4

Coursework: Applied Statistics, Machine Learning, Deep Learning, Data Analytics with R Program, Data Visualization

Experience

New Jersey Institute of Technology — Newark, NJ

Research Assistant | NJIT Engineering Education Research — July 2024 — May 2025

- Analyzed institutional undergraduate student data using machine learning and developed regression models (Multiple Linear Regression, Random Forest, XGBoost, LightGBM) in R to predict GPA
- Performed ablation studies to determine the influence of academic, demographic and socioeconomic factors on students performance
- Leveraged SHAP values to generate variable importance as a model agnostic metric to compare various machine learning algorithms

Research Assistant | Data and Knowledge Engineering Lab — Sept 2023 — June 2024

- Developed "SolarFlareNet"—a transformer framework used to forecast the occurance of solar flares with 90.7% accuracy
- Integrated xAI tools like LIME, SHAP, Anchors, PDP, and ALE into SolarFlareNet to analyze the black-box model predictions
- Published and presented research findings at FLAIRS 2024 and IEEE ICTAI 2024

Impetus — Bengaluru, India

Data Engineer — July 2020 — Aug 2023

- Designed and implemented scalable ETL pipelines integrated with data quality checks to ingest and process 10 TB of raw data using PySpark, reducing processing times by 40% and accelerating access to business insights
- Leveraged Apache Parquet snappy compression and Amazon S3 life cycle policies reducing storage and I/O costs by 30%
- Proposed a data transformation plan utilizing DBT and Aiflow to achieve a 15% increase in data transformation efficiency
- Collaborated with stakeholders to design and implement robust data models for generating 10+ key KPIs, ensuring data accuracy and alignment with business reporting needs
- Accelerated the migration of data from Snowflake warehouse to S3 for a data lake solution, leveraging Athena for ad hoc analysis and Redshift Spectrum with materialized views for BI dashboards, resulting in 50% faster analytics report generation

Technical Skills

Languages	:	Python (PySpark, Pandas, Tensorflow, PyTorch, scikit-learn, Streamlit), SQL, R, Bash
Databases	:	PostgreSQL, MySQL, Oracle (PL/SQL), MongoDB, DuckDB
Cloud	:	AWS (S3, Glue, Lambda, Athena, Redshift, Aurora, RDS, DynamoDB, Firehose, SageMaker) GCP (Cloud Storage, BigQuery, Dataflow, Dataproc, Bigtable)
Big Data	:	Trino, Databricks, Snowflake, Apache Spark, DBT
Visualization	:	Tableau, Looker, Power BI, Dash (Python), Apache Superset, Excel
Data Modeling	:	Normalization (3NF), OBT, Star Schema, Snowflake Schema, Data Vault
CI/CD	:	Git, GitHub, GitLab, GitHub Actions, Docker, Kubernetes, Terraform, Jenkins

Research Publications

1. Interpretable Deep Learning for Solar Flare Prediction — IEEE ICTAI 2024 — 2024
2. An Interpretable Transformer Model for Operational Flare Forecasting — FLAIRS 2024 — 2024

Projects

TradeForecast | *Python, PyTorch, PyTorch Lightning, yfinance, Polars, scikit-learn* — Report

- Developed three deep-learning architectures (LSTM, CNN+LSTM, Transformer) for multi-horizon timeseries forecasting of stock price
- Orchestrated training using "ReduceLROnPlateau" learning rate scheduler in PyTorch for faster convergence, and implemented hyperparameter tuning using grid search
- Implemented feature engineering by ingesting OHLCV data via yfinance and adding temporal variables (Day of week, Fiscal Quarter) and technical indicators (MA, MACD, RSI, ATR)

Data Engineer Playground | *Docker, Airflow, Trino, Spark, MinIO, PostgreSQL, Project Nessie, Unity Catalog* — Github

- Built a fully containerized multi-service environment to prototype end-to-end ETL workflows, from data ingestion in MinIO to batch or stream processing with Spark and workflow orchestration via Airflow. Enabled interactive SQL analytics through Trino with connectors for Postgres DB, Nessie Catalog and Unity Catalog

AlgoTrade API | *Python, yfinance, Pandas, Tensorflow, ks-api-client* — Github

- Developed a fully automated NSE stock trading bot in Python by integrating real-time and historical data with yFinance, training ML models (including LSTM) for stock price prediction, and executing live trades via the Kotak Securities API

Certifications

- Google Data Analytics Professional Certificate — Coursera — 2023