Nitin Jayavarapu

Fremont, CA | nitinjayavarapu12@gmail.com | +1-448-219-7141
https://github.com/nitinjayavarapu12 | https://www.linkedin.com/in/nitinj12/

Professional Summary

Data Engineer / Data Analyst with 2+ years of experience building end-to-end ML pipelines, statistical models, and analytical systems on large-scale operational and time-series data. M.S. in Data Science (GPA: 3.51) with graduate-level coursework in Statistical Modeling, Regression, Machine Learning, and Deep Learning. Proven track record of translating complex data into production-ready tools and business insights across energy, healthcare, and labor market domains.

Technical Skills

Languages & Libraries: Python (Pandas, NumPy, Scikit-learn, XGBoost, SHAP), SQL
Machine Learning: Regression, Classification, Time-series Forecasting, Anomaly Detection, Feature Engineering, Model Evaluation
Deep Learning & NLP: Neural Networks, TF-IDF, Embeddings, Text Processing Pipelines
Data & Analytics: EDA, KPI Design, Statistical Modeling, A/B Testing Concepts, Large-scale Data Pipelines
Visualization & BI: Matplotlib, Seaborn, Streamlit, Tableau
Tools & Infrastructure: Git, Linux, Docker, AWS (S3, EC2), FastAPI

Experience

AI Engineer Intern 2024 – Present
LumisAI *SF Bay Area, CA (Remote)*

- Built and deployed statistical forecasting models for operational time-series data, applying ML pipelines that directly align with M.S. Data Science coursework in regression modeling and advanced analytics.
- Developed and maintained production-grade FastAPI inference endpoints on AWS EC2, maintaining ∼99.9% uptime while handling traffic spikes and optimizing latency-accuracy trade-offs.
- Partnered with cross-functional stakeholders to document data requirements, surface compliance-relevant edge cases, and deliver tools that were auditable and maintainable post-handoff.

Data Analyst January 2022 – November 2023
APSPCL *Vijayawada, India*

- Analyzed high-frequency solar generation and weather data across multiple utility-scale solar parks, processing millions of records per month to track CUF, asset performance, and operational KPIs — delivering consistent reporting to planning and grid operations teams.
- Built and automated SQL- and Python-based data pipelines for inverter-level and meter data, reducing manual reconciliation effort by ∼30–40% and improving reporting turnaround time from days to hours.
- Conducted forecast-versus-actual analysis for day-ahead and intra-day solar generation, identifying recurring deviation and curtailment patterns that directly informed grid scheduling and operational planning decisions.
- Performed trend and anomaly analysis to detect recurring outages and performance degradation, enabling engineering teams to prioritize maintenance actions and reduce repeat operational incidents.

Education

University of West Florida Pensacola, FL
Master of Science in Data Science, GPA: 3.51/4.0 *January 2024 – December 2025*

- Relevant Coursework: Statistical Modeling (A), Modeling in Regression (B+), Machine Learning (A), Deep Learning for Data Science (A), Mathematics for Data Science (A-), Advanced Business Intelligence Applications, Data Science Capstone

Technical Projects

MediBridge – AI-Powered Discharge Summarizer & Follow-Up Tool | *Python, FastAPI, LLaMA 3, Supabase, Twilio*

- Identified a critical healthcare data gap — 40–50% of discharged patients misunderstand clinical notes — and built an end-to-end LLM pipeline over 4,999 real medical transcriptions (MTSamples) to translate, extract, and safety-check discharge instructions.
- Engineered a confidence scoring layer that flags model-inferred vs. explicitly extracted fields, preventing silent hallucinations in safety-critical output — a compliance-first design decision ensuring correctness before delivery.
- Deployed a production system integrating a Twilio SMS check-in bot with secondary LLM red-flag detection, a live caregiver dashboard, and a full FastAPI backend on Render with Supabase (PostgreSQL).

NLP-Driven Job Market Skill & Salary Analysis | *Python, SQL, Pandas, Scikit-learn, TF-IDF*

- Analyzed 3,254 LinkedIn job postings using NLP (TF-IDF, regex taxonomy matching) to quantify what technical skills, skill combinations, and salary signals define the 2024 data job market across 8 role types.
- Built an end-to-end text processing pipeline — filtering 124K mixed-industry postings down to relevant data roles, stripping HTML noise, and converting messy job descriptions into a structured 69-skill binary matrix for downstream analysis.
- Discovered ML/AI skills carry the highest salary premiums while Excel and forecasting skills correlate with lower-paying roles — providing hiring teams and job seekers a concrete, data-backed view of where skill investment pays off most.
- Found SQL and Python co-occur in 67% of postings that mention either, and that cloud + pipeline skills cluster tightly in Data Engineer roles while BI tools stay almost exclusive to Analyst roles — revealing how employers bundle skill expectations by role type.

Urban Air Quality Monitoring & Forecasting | *Python, XGBoost, Random Forest, SHAP, Streamlit, SQL*

- Engineered an automated EPA data pipeline ingesting 175,000+ air quality observations across 5 pollutants and 6 years, enabling daily-refreshable forecasts for city planning teams.
- Developed an XGBoost next-day PM2.5 forecasting model achieving RMSE of 11.7 AQI points with a chronological train/test split, giving public-health officials a 24-hour advance warning system for unhealthy air events.
- Trained a Random Forest severity classifier (CV F1-macro = 0.77) with SHAP explainability across 25 engineered features, reducing ambiguity in air quality triage decisions for non-technical stakeholders.
- Deployed a 4-page Streamlit dashboard with interactive forecasting, anomaly detection (110 flagged pollution events), and an interactive severity predictor — making complex ML outputs accessible to city planners without data science backgrounds.