# Fahad Shaikh

*Data Scientist / Machine Learning Engineer | 6+ Years Experience (ML, MLOps, Generative AI)*
New York City | fs12516n@pace.edu | +1 (329) 204 9984
Portfolio: fahad-sha.github.io | GitHub: Fahad-sha | LinkedIn: Profile

## Experience

**Data Scientist (ML Engineering / People Analytics)** Nov 2023 – Aug 2024
*Endeavor Group (IMG Media), New York City*

- Architected and productionized a Retrieval-Augmented Generation (RAG) system over **2k+ HR documents**, reducing policy lookup time by **60%** and enabling citation-grounded responses.
- Orchestrated batch inference pipelines in Amazon SageMaker processing **10k+ records per run**, introducing schema and data-quality checks that reduced pipeline failures by **40%**.
- Designed self-serve analytics dashboards in Amazon QuickSight adopted by **30+ HR stakeholders**, cutting reporting turnaround from days to hours.
- Coordinated with **5+ cross-functional teams** (HR, IT, Legal, Analytics) to align ML outputs with governance and compliance requirements.

**Data Scientist — Client Engagement (Microsoft)** Apr 2021 – Jun 2022
*Mu Sigma, Bengaluru*

- Developed segmentation models (K-Means, GMM) on datasets exceeding **1M user records**, identifying **6–8 behavioral cohorts** that informed product prioritization.
- Synthesized analytical findings into executive-ready narratives consumed by PM and leadership teams during quarterly roadmap planning.
- Standardized feature engineering and preprocessing pipelines reused across analyses, reducing setup time by **30%**.

**Data Scientist — Client Engagement (Refinitiv / LSEG)** Apr 2020 – Mar 2021
*Mu Sigma, Bengaluru*

- Produced time-series forecasting solutions (ARIMA, Prophet) across **10+ revenue streams**, achieving backtested error reductions of up to **15–20%**.
- Refined revenue-risk modeling workflows by formalizing evaluation protocols, improving model review cycle time by **25%**.
- Translated business objectives into measurable modeling targets in collaboration with finance and risk stakeholders.

**Data Scientist — Client Engagement (Citigroup)** Apr 2019 – Mar 2020
*Mu Sigma, Bengaluru*

- Engineered anomaly detection models (Isolation Forest, One-Class SVM) screening **millions of transactions**, reducing false positives by **20%**.
- Strengthened feature pipelines spanning **50+ engineered features**, improving interpretability for risk reviewers.
- Established reproducible training and evaluation workflows adopted by multiple analysts, accelerating model handoff cycles.

**Trainee (Automotive Engineering)** Aug 2018 – Mar 2019
*KPIT, Pune*

- Developed Adaptive Cruise Control control logic in **Embedded C** using **PID controllers**, aligning controller behavior with system requirements and safety constraints.
- Verified controller performance across representative scenarios (speed tracking, lead-vehicle changes) and ensured requirements compliance through structured test cases.
- Managed ECU requirements and system specifications in **IBM DOORS**, maintaining traceability across requirements, design artifacts, and validation evidence for engineering compliance.

## Projects

**DashPilot AI Agent — Applied AI & Development** GitHub | Live

- Designed a local-first analytics dashboard supporting CSV ingestion up to **100k+ rows** with sub-second aggregate computation in-browser.
- Developed a Gemini-powered AI analyst constrained to deterministic aggregates, generating grounded insights while preventing hallucinations.
- Implemented a pin-able widget system with optional Supabase persistence, enabling repeatable analysis workflows.
- Optimized bundle size and rendering to ensure fast load times on GitHub Pages with minimal dependencies.

**Enterprise RAG Knowledge Assistant — Generative AI** GitHub

- Assembled a production-grade RAG pipeline indexing **thousands of documents**, enabling long-context retrieval and citation-backed responses.
- Reduced manual document search effort by **50%** through semantic retrieval and structured summaries.

**Audio-Visual JEPA-mini — Advanced Research (Colab Free Tier)** GitHub | Live

- Implemented a JEPA-style cross-modal model trained on a **2k–10k clip VGGSound subset**, learning aligned audio–video embeddings without contrastive negatives.
- Trained lightweight vision and audio encoders under free-tier Colab constraints using BYOL-style EMA targets for stability.
- Demonstrated lift over random baselines on **audio↔video retrieval (Recall@K)** with controlled ablations.

## Skills

**Programming & Data:** Python, SQL, R
**Machine Learning:** supervised and unsupervised learning, feature engineering, model training, hyperparameter tuning, cross-validation, model evaluation
**Deep Learning:** PyTorch, TensorFlow/Keras, Transformers, attention mechanisms
**Generative AI:** Retrieval-Augmented Generation (RAG), LangChain, Gemini, prompt engineering, embeddings, vector databases
**Cloud & Data Engineering:** AWS (S3, SageMaker, Redshift, OpenSearch), Azure (Databricks, Data Factory, Azure SQL), ETL/ELT pipelines
**MLOps:** model deployment, batch inference, reproducible training, CI/CD, monitoring, data quality checks
**Tools:** Git, Docker, Kubernetes, Supabase, React, Vite, IBM DOORS

## Education

**Pace University**, New York City
*MS in Data Science*, GPA: 4.0
Relevant Coursework: Deep Learning, Data Mining, Scalable Databases, Generative AI, Autonomous Systems

## Awards

- **Impact Award (Refinitiv / LSEG):** Revenue-risk model redesign resulting in $400K annual savings.
- **Spot Awards:** Recognized for high-impact ML delivery across Microsoft and Citigroup engagements.