Fahad Shaikh

New York City | fs12516n@pace.edu | +1 (329) 204 9984 | Portfolio | LinkedIn | GitHub

Experience

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

- Built a RAG pipeline for HR knowledge retrieval using Amazon S3 (document store) + Amazon Bedrock (embeddings + LLM) + Amazon OpenSearch Service (vector index), enabling semantic search and grounded QA over policies, SOPs, and HR documentation with citations.
- Implemented automated batch scoring using Amazon SageMaker with scheduled runs (e.g., EventBridge triggers), adding reliability checks (schema/quality validations) to reduce pipeline failures.
- Delivered BI dashboards using Amazon QuickSight backed by Amazon Redshift/Athena, translating complex workforce KPIs into self-serve views used by HR stakeholders.

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

- Built **segmentation** models (K-means, GMM) to identify behavioral cohorts and drive product strategy; improved interpretability via clear cluster profiling.
- Produced stakeholder-ready analysis packs (cohort definitions, drivers, recommendations) to support decision-making and roadmap prioritization.
- Implemented scalable data prep and feature computation patterns for large datasets; emphasized reliability and repeatable runs.

Data Scientist (Contract / Client Engagement) *Refinitiv (an LSEG business)* — *Blr, Apr 2020 – Mar 2021*

- Delivered **time series forecasting** (ARIMA, Prophet) for planning, including backtesting and error analysis to support model selection and deployment readiness.
- Redesigned/validated revenue-risk modeling workflows; improved rigor through evaluation protocols and clear assumptions documentation.
- Partnered cross-functionally to translate business objectives into measurable modeling targets and deliverables.

Data Scientist (Contract / Client Engagement) *Citigroup* — *Blr, Apr 2019 – Mar 2020*

- Developed **anomaly detection** solutions (Isolation Forest, One-Class SVM) for risk/fraud use cases; partnered with stakeholders to define thresholds and validation strategy.
- Built robust feature engineering and validation steps to reduce false positives and improve interpretability for risk reviewers.
- Delivered repeatable model training/evaluation workflows with clear documentation for handoff.

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

- Developed autonomous control logic for Adaptive Cruise Control using **embedded C** and PID controllers; validated behavior against system requirements.
- Managed ECU requirements and system specifications using IBM DOORS, ensuring traceability and compliance with engineering standards.

Skills

Programming: Python, SQL, R **Software:** Git, Docker, Kubernetes, CI/CD, Testing (unit/integration)

ML & Modeling: scikit-learn, XGBoost, PyTorch, TensorFlow/Keras; feature engineering, hyperparameter tuning, cross-validation, model evaluation

Data & Cloud: Azure Databricks, Azure Data Factory, Azure SQL; AWS EMR, S3, Redshift; ETL/ELT, orchestration, scalable batch processing

MLOps: reproducible training, batch scoring pipelines, monitoring-ready hooks (data quality + drift), governance basics

GenAI : LangChain, RAG pipelines, prompt engineering, transformers (BERT/GPT/Llama), vector DBs (ChromaDB)

Education

Pace University, *MS in Data Science,* GPA 4.0 — *NYC, Sept 2024 – Dec 2025*

- **Relevant Coursework:** Deep Learning, Data Mining, Scalable Databases, Generative AI, Autonomous Systems.

Accomplishments

- **Impact Award, Refinitiv (LSEG engagement):** Redesigned revenue-risk model, resulting in $400K annual savings.
- **Spot Awards (consulting/vendor engagement):** Recognized for delivering critical ML solutions supporting Microsoft and Citigroup stakeholders.

Projects

Multi-Modal Video Summarizer Agent GitHub Link

- Built a generative AI system using **RAG** concepts to ingest transcripts (and contextual signals) and generate structured summaries for long-form content.
- Engineered NLP processing for long context and evaluation-oriented outputs (concise notes, key concepts, and sectioned summaries).

Transformer Architecture Re-implementation (PyTorch) GitHub Link

- Re-implemented core Transformer components from scratch in **PyTorch** (multi-head attention, positional encoding, FFN) to deepen understanding of performant model code.
- Validated correctness via controlled experiments and reproducibility practices (fixed seeds, clear configs, repeatable runs).

Flight Delay Prediction System (Distributed Data + ML) GitHub Link

- Built a classification pipeline to predict flight delays using **XGBoost**, including feature engineering and evaluation with decision-focused metrics.
- Processed large-scale records using **distributed data tooling** (Hive/Spark-style workflows) and packaged results for downstream consumption.