Karan Badlani

Seeking AI Data Science and Machine Learning Opportunities

Boston, MA | 617-992-1174 | badlani.k@northeastern.edu | GitHub | LinkedIn

SUMMARY

AI Data Scientist with 2+ years of experience building **predictive models**, **RAG platforms**, and **production ML pipelines** that directly impact business outcomes. Reduced document retrieval from 40 minutes to 30 seconds with a RAG platform at MFS; deployed **XGBoost** churn pipelines at 88% recall and cut forecast error from 7.5% to 1.65%.

TECHNICAL SKILLS

Languages: Python, SQL, R, Java
Data and ML Libraries: Pandas, NumPy, Matplotlib, Scikit-learn, TensorFlow, PyTorch, LangChain, LangGraph, HuggingFace
Machine Learning: XGBoost, Random Forest, Gradient Boosting, Decision Trees, Logistic Regression, SVM, Regression Analysis
Cloud & Infra: Azure (ML Studio, Data Factory, Delta Lake Storage), AWS (S3, ECS), GCP (BigQuery), Snowflake, Databricks (Spark), Apache Airflow, Docker, MLflow
Tools & Visualization: Tableau, Power BI, Streamlit, MS Excel, Git/GitHub, AI-Assisted Development (GitHub Copilot, Claude Code)
Concepts: Predictive Modeling, EDA, Feature Engineering, A/B Testing, Clustering, Time-Series Analysis, Generative AI, RAG, LLM Evaluation, Agentic AI, Responsible AI, Model Monitoring, Drift Detection, MLOps, CI/CD, NLP (LLMs, Embeddings, Transformers)

WORK EXPERIENCE

ML Engineer (Part-time) | **Squark AI** | *Remote, US* Mar 2026 – Present

- Refactored the **XGBoost** classification pipeline within Squark AI's **AutoML** platform, improving hyperparameter tuning efficiency and reducing average model training time by **20%** across benchmark client datasets.
- Extended preprocessor **data validation** logic to handle edge-case null distributions, reducing pipeline failure rate by **15%**; contributed to **Docker** containerization and **AWS** deployment documentation to streamline CI/CD workflows.

Data Scientist | **MFS Investment Management** | *Boston, MA* Jan 2025 – Jul 2025

- Orchestrated an end-to-end **sales forecasting pipeline** benchmarking **SARIMA** against **Prophet**, applying **forward chaining** for bottom-up monthly forecasts and cutting advisor sales error from **7.5% to 1.65%** across 500 financial advisors.
- Architected **Snowflake SQL** models across four sources (Sales, AUM, Interaction, Advisor metadata), resolving a **200GB+** compute bottleneck to enable reliable downstream modeling and reporting at scale.
- Engineered **behavioral features** from 200GB+ of advisor activity data to power a **Random Forest** classifier segmenting advisors into Rising Stars and Falling Angels at **80% recall**, enabling targeted retention interventions.
- Designed **A/B testing** experiments with **Welch's t-test** across 1,000 advisors, translating churn model outputs into data-driven recommendations that drove a **17% lift** in sales conversion for senior leadership.
- Led development of **FiBi**, an internal Research Intelligence Platform centralizing **1,500+** investment documents with AI-powered Q&A for portfolio managers, reducing research discovery time from **40 minutes to under 30 seconds**.
- Designed an **agentic retrieval layer** using **LangGraph** that graded document relevance and rewrote weak queries automatically, improving retrieval **F1 from 0.64 to 0.85** and reducing false positives by **55%**.

Data Analyst | **InfoCepts** | *India* Dec 2022 – Jun 2023

- Designed an **ADF ETL/ELT pipeline** ingesting CRM and Sales data from on-prem SQL Server through SHIR, applying **SQL stored procedure** transformations and surfacing KPIs in **Power BI**, cutting CFO-office reporting errors by 17%.
- Integrated 45GB of partner API data through **Java Spring Boot** by validating JSON payloads with Postman and loading records into **ADLS Gen2** through version-controlled pipelines, eliminating manual extraction delays.

Data Analyst Intern | **Superfine Minerals** | *India* Mar 2022 – Oct 2022

- Produced a **SARIMA** demand forecasting model and **Tableau** dashboard by automating Oracle SQL data ingestion, cutting analyst reporting time from 3 days to under 30 minutes.

PROJECTS

Document Q&A RAG System GPT-4 LLM, FastAPI, Qdrant, LlamaIndex, OpenAI, Streamlit GitHub

Built an end-to-end **Generative AI** RAG pipeline using **LlamaIndex** for section-aware PDF chunking and **Qdrant** vector search with idempotent upserts, leveraging **LLMs** for scalable information retrieval from unstructured documents. Evaluated with **RAGAS** achieving 0.92 faithfulness and 0.78 context precision across 50 test queries through **LLM-as-a-Judge**; served through a **FastAPI** backend with Streamlit frontend.

Churn Risk: MLOps Python, XGBoost, MLflow, FastAPI, Docker, AWS ECS GitHub

Developed a production-grade classification workflow using **XGBoost** (88% recall) for customer attrition risk, with systematic **hyper-parameter tuning**, **feature importance** analysis on behavioral signals, and experiment tracking via **MLflow** for model monitoring and drift detection. Containerized as a cloud-native **FastAPI** microservice on **Docker**, deployed to **AWS ECS** with sub-10ms inference latency, and automated build-test-deploy through **GitHub Actions CI/CD** for enterprise-scale serving.

EDUCATION

Northeastern University Boston, MA
Master of Science in Data Science – GPA: 3.6/4.0 Sep 2023 – Dec 2025
Coursework: Machine Learning, Statistical Analysis, Data Mining, Database Management, Algorithms

Ramdeobaba University Nagpur, India
B.E. Computer Engineering – GPA: 3.55/4.0 Jul 2019 – Apr 2023