

Sourabh Sonker

Data Scientist & AI Engineer · LLMs · RAG · NLP · ML · MLOps · Python

Delhi, India · +91-8447727696 · sourabhsonker17@gmail.com

[linkedin.com/in/sourabh-sonker](https://www.linkedin.com/in/sourabh-sonker) · github.com/Sourabh1710

sourabh1710.github.io

PROFILE

Data Scientist and AI Engineer with **hands-on production experience** across LLM pipelines, RAG systems, NLP, classical ML, and MLOps — each project deployed live with measurable outcomes. Comfortable building end-to-end: from data ingestion and model evaluation to REST APIs, Docker containers, and CI/CD pipelines. Strong Python and HuggingFace ecosystem depth; experienced with multi-model LLM orchestration, vector databases, and automated evaluation frameworks. **Available immediately, open to relocate anywhere in India.**

TECHNICAL SKILLS

LLM / AI	LangChain · FAISS · Qdrant · HuggingFace Transformers · Sentence-Transformers · Gemini 2.5 Flash · Sarvam AI (sarvam-30b) · Groq (Llama 3.3 70B) · Ragas · Prompt Engineering
ML / NLP	PyTorch · scikit-learn · XGBoost · LightGBM · SMOTE · SHAP · Optuna · Prophet · SARIMA · FinBERT · Pandas · NumPy
Engineering	FastAPI · Docker · GitHub Actions (CI/CD) · MLflow · asyncio · httpx · pytest · Ruff · Black · mypy · Pydantic · Render · Streamlit
Data / BI	SQL (CTEs, Window Functions, Aggregations) · Python (Plotly, Seaborn) · Power BI · Excel (Power Query, DAX) · yfinance

PROJECTS

Multilingual Indic RAG Pipeline — Cross-Lingual Hinglish Q&A System

[Live App ↗](#) [GitHub ↗](#)

Python · asyncio · httpx · Qdrant Cloud · paraphrase-multilingual-MiniLM-L12-v2 · Gemini 2.5 Flash · Sarvam AI (sarvam-30b) · Groq Llama 3.3 70B · Ragas · pytest

- ▶ Built **cross-lingual retrieval without translation overhead** — Hinglish queries map directly to English document chunks via shared 384-dim multilingual vector space (paraphrase-multilingual-MiniLM-L12-v2), bypassing semantic loss from translation loops; vendor-agnostic generation layer dynamically routes to Gemini 2.5 Flash, Sarvam AI sarvam-30b, or Groq Llama 3.3 70B via custom async httpx client
- ▶ Production-grade engineering: **circuit-breaker graceful degradation** (429/400/403 failures caught, pre-cached fallback served — 100% uptime); deterministic UUID v5 idempotency for safe upserts; Ragas automated eval harness scored **Faithfulness 0.9250, Context Recall 1.0000** on human Golden Dataset; **8/8 tests passing**, Ruff + Black + mypy clean

Financial News Sentiment — FinBERT Evaluation vs. Classical Baselines

[Live App ↗](#) [GitHub ↗](#)

Python · FinBERT (HuggingFace) · PyTorch · scikit-learn · VADER · Scipy · yfinance · Plotly · Streamlit

- ▶ Structured a fuzzy NLP question into a **rigorous 3-model evaluation** on 5,842 financial headlines with predefined metrics — FinBERT: **75.1% accuracy, 4.3× bearish F1 gain** (0.14→0.60) vs TF-IDF+SVM; domain-aware preprocessing validated via ablation study; caught label-misalignment bug inflating VADER baseline by **25.3 percentage points** through systematic EDA
- ▶ Built Pearson/Spearman sentiment-price correlation pipeline across 5 tickers, 2 years with p-value reporting; resolved ~1hr cold-start to **<1 sec** via precomputed artifact loading; 4-tab Streamlit dashboard deployed on Streamlit Cloud

House Price Prediction — End-to-End MLOps Pipeline

[Live API ↗](#) [GitHub ↗](#)

Python · scikit-learn · XGBoost · LightGBM · MLflow · FastAPI · Pydantic · Docker · GitHub Actions · Render

- ▶ Leak-proof sklearn pipeline (ColumnTransformer fit exclusively on train folds); MLflow experiment tracking across 4 models — XGBoost selected at **CV RMSLE 0.1209** vs LightGBM (0.1284) and Ridge (0.1392); FastAPI REST service with Pydantic validation and auto-generated Swagger docs
- ▶ Multi-stage Docker build + GitHub Actions CI/CD: auto-tests → retrain → build → deploy to Render on every push; **failing tests block deployment**; diagnosed and resolved 5 independent production issues including CI mock scoping and Docker image bloat

Bank Customer Churn Prediction & Explainability Dashboard

[Live App ↗](#) [GitHub ↗](#)

Python · XGBoost · scikit-learn · SMOTE · SHAP · Optuna · Plotly · Streamlit

- ▶ XGBoost classifier: **86.9% AUC-ROC** on 10K bank records; +3.6 pts via Optuna Bayesian tuning (50 trials); benchmarked LR (77.2%) and RF (86.2%); SMOTE for 4:1 class imbalance; engineered 4 domain features — products_per_tenure ranked #2 in SHAP importance, revealing non-linear 3–4 product churn cliff
- ▶ SHAP TreeExplainer identified top 5 retention levers (age 41–60, inactive membership, mono-product high-balance holders); live Streamlit app delivers per-customer churn probability with SHAP waterfall — business framing throughout: false negative cost vs false positive spend

EDUCATION

B.Tech — Mechanical Engineering

Delhi Technological University (DTU) · 2020–2024

CGPA: **7.86 / 10.0**

Self-Directed AI & Data Science Portfolio

Independent Study · 2024–Present

4 production-deployed projects · LLMs, RAG, NLP, ML, MLOps

CERTIFICATIONS

✓ Agentic AI — HuggingFace · 2026	View ↗
✓ Basic Statistics — University of Amsterdam · Coursera	View ↗
✓ Excel Skills for Business (Essentials + Advanced) — Macquarie University · Coursera	View ↗
✓ Kaggle Micro-Courses (10 certificates) — Python, Pandas, SQL, ML, Deep Learning, Time Series, Feature Engineering, Explainability	View ↗