Sourabh Sonker

Data Scientist · ML · NLP · Time Series · LLMs · MLOps

Delhi, India · +91-8447727696 · sourabhsonker17@gmail.com
linkedin.com/in/sourabh-sonker · github.com/Sourabh1710
Portfolio

PROFILE

Data Scientist transitioning from Mechanical Engineering (DTU 2024, CGPA 7.86), with **5 production-deployed projects** across ML, NLP, Time Series, LLM/RAG, and MLOps — each with a live deployment and measurable results. Strong interest in Finance and FinTech; skills fully transferable across any data-intensive domain. **Available immediately.**

TECHNICAL SKILLS

Languages	Python · SQL (CTEs, Window Functions, Aggregations) · Excel (Power Query, Pivot Tables, DAX)
ML / DS	XGBoost · LightGBM · scikit-learn · SMOTE · SHAP · Optuna · Prophet · SARIMA · PyTorch · Pandas · NumPy
LLM / NLP	LangChain · FAISS · HuggingFace Transformers · FinBERT · Sentence-Transformers · Google Gemini · Groq
MLOps	FastAPI · Docker · GitHub Actions (CI/CD) · MLflow · Render · Streamlit · Plotly · Power BI

PROJECTS

RAG-Based Document Q&A System (LLM)

Live App ↗ GitHub ↗

`Python · LangChain · FAISS · HuggingFace Sentence-Transformers · Google Gemini 2.5 Flash · Groq Llama 3 · Streamlit · PyPDF2`

- End-to-end RAG pipeline: PyPDF2 → RecursiveCharacterTextSplitter (800-char/100-overlap) → all-MiniLM-L6-v2 FAISS embeddings (**384-dim**) → dual LLM backends (Gemini 2.5 Flash + Groq Llama 3) with anti-hallucination system prompt
- ConversationBufferMemory for multi-turn Q&A; **zero hallucination** by design — answers grounded strictly in retrieved context; @st.cache_resource caching deployed on Streamlit Cloud

House Price Prediction — End-to-End MLOps Pipeline

Live API ↗ GitHub ↗

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

- Leak-proof sklearn pipeline (ColumnTransformer fit on train folds only); MLflow tracking across 4 models — XGBoost selected at **CV RMSLE 0.1209**; FastAPI + Pydantic validation + auto-generated Swagger docs
- Multi-stage Docker + GitHub Actions CI/CD: auto-tests → retrain → build → deploy to Render on every push; failing tests block deployment; resolved 5 production issues independently

Financial News Sentiment Analysis + Stock Correlation

Live App ↗ GitHub ↗

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

- FinBERT zero-shot on 5,842 headlines: **75.1% accuracy**, **4.3× improvement** in bearish signal detection (F1: 0.14→0.60) over TF-IDF+SVM; caught label-misalignment bug inflating VADER baseline by 25.3 pts
- 4-tab Streamlit dashboard; ~1hr cold-start resolved via precomputed artifacts (tab load <1 sec); Pearson/Spearman sentiment-price pipeline across 5 tickers, 2 years

Retail + NSE Stock Time Series Forecasting

Live App ↗ GitHub ↗

`Python · Statsmodels · pmdarima · Prophet · yfinance · Plotly · Streamlit`

- SARIMA: **8.8% MAPE** on 90-day retail horizon (3M+ rows); macro signal: WTI oil price r=−0.47 with grocery sales; April 2016 earthquake modelled as named Prophet event (2× peak demand)
- Extended to live NSE stocks (TCS.NS via yfinance); ARIMA(0,1,1) documented against weak-form EMH; Streamlit app: 10 tickers, dual ARIMA+Prophet, 30–90 day configurable horizons

Bank Customer Churn Prediction & Explainability

Live App ↗ GitHub ↗

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

- XGBoost: **86.9% AUC-ROC** on 10K records (+3.6 pts via Optuna, 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); SHAP TreeExplainer identified top 5 retention levers including age 41–60, inactive membership, mono-product high-balance holders

EDUCATION & DEVELOPMENT

B.Tech — Mechanical Engineering
Delhi Technological University (DTU) · 2020–2024
CGPA: **7.86 / 10.0**

Self-Directed Data Science Transition
Independent Study & Portfolio Development · 2024–Present
5 production-deployed projects across ML, NLP, Time Series, LLMs, MLOps

CERTIFICATIONS

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