

STEPHEN BARAIK

Data Analyst | Data Scientist | Data Engineer

+91 91375 54921 | stephenbaraik@gmail.com | linkedin.com/in/stephenbaraik | github.com/stephenbaraik

Summary

Data professional with expertise in ETL pipelines, machine learning, and AI application development. Strong in Python, SQL, and production data infrastructure, with a focus on building reliable, scalable systems.

Experience

Zoc Learnings

Jan 2024 – Present

Data Management Analyst

Remote

- Centralized data from 10 sales persons by building per-person CRM sheets pipelined into one compliant master Zoc CRM.
- Automated weekly payment-pending & monthly sales reports via Apps Script, auto-generating HTML emails for stakeholders.
- Prototyped an experimental RAG chatbot (FastAPI, LangChain, ChromaDB, Next.js) for automated lead qualification.
- Audited company websites using data science techniques, fixing 115 broken pages and adding sitewide schema coverage.

Projects

Extreme Weather Prediction System ML Model [link](#)

- Trained ML models forecasting cyclones 98%, heatwaves 99%, & rainfall 88% across 15 Indian cities on 90k+ real records.
- Fixed label leakage across 3 models, correcting inflated 99.7% rainfall accuracy to a genuine 88.3%.
- Deployed 5-endpoint FastAPI service with MLflow registry and Docker Compose, processing 90k + records via PySpark ETL.
- Automated CI/CD with GitHub Actions, validated by 80 passing pytest tests and KS-test drift monitoring.

PaisaPro: AI Investment Advisory Platform [link](#)

- Deployed a forecasting and risk engine for 502 NSE/BSE stocks using ARIMA, GARCH(1,1), & Fama-French 4-factor models.
- Trained a walk-forward-validated machine learning ensemble on 20+ OHLCV features for 30-day price prediction.
- Built a FastAPI backend with 35+ endpoints across 19 services, powering an agentic Llama 3.3 with live tool calls over SSE.
- Shipped an 18-page React SPA on Netlify, backed by a Supabase database and a Hugging Face Spaces Docker deployment.

ChartCraft: Open-Source Python Dashboard Framework [link](#)

- Designed an open-source Python dashboard library that rivals Power BI and Tableau, shipping 18+ chart types and 11 themes.
- Engineered the core on Python's standard library and Apache ECharts 5.5, requiring zero external dependencies.
- Synced a drag-and-drop canvas with the underlying Python in real time using a bidirectional, AST-based codegen.
- Shipped SSE-based live streaming with independent per-chart refresh, plus export to HTML, Jupyter, Docker, and PDF.

Bitcoin Data Pipeline [link](#)

- Automated a Python module that pulls historical and real-time OHLC candles from the Binance API.
- Built two Airflow DAGs: one for historical backfill, one for real-time ingestion running every minute.
- Persisted candle data in PostgreSQL with idempotent inserts to prevent duplicate records.
- Containerized the full stack (Airflow, PostgreSQL, pgAdmin) with Docker Compose for reproducible local setup.

Skills

Programming: Python, SQL, R, Bash

Machine Learning & AI: Scikit-learn, TensorFlow, PyTorch, MLflow, LangChain (RAG), Stable-Baselines3, LLM Integration

Data Engineering: Apache Spark, Hadoop (HDFS, MapReduce), Airflow, ETL Pipelines, Docker

Backend & Web: FastAPI, React, Next.js, Supabase, Tailwind CSS

Databases & Cloud Data: PostgreSQL, MongoDB, Supabase

Analytics & Visualization: Pandas, NumPy, Matplotlib, Seaborn, Tableau, Power BI

Developer Tools: Git, Jupyter, WSL, LaTeX, Windows, Linux

Education

NMIMS University – B.Sc. in Data Science

2023 – 2026

Relevant coursework: Machine Learning, Data Warehousing, Statistical Methods, Big Data Analytics, Python Programming.

Certifications

- DataCamp Associate Data Engineer (2025) [link](#)