7004928439
Patna, Bihar
Souravraj664@gmail.com

Sourav Raj

Data Scientist

github.com/sourav664
linkedin.com/in/sourav664/
portfolio/sourav664
medium.com/@souravraj664

Technical Skills

- Programming & Analysis: Python (Pandas, NumPy, SciPy), SQL (MySQL)
- Visualization & BI: Matplotlib, Seaborn, Power BI, Excel
- Machine Learning & Deep Learning: Scikit-learn, XGBoost, LightGBM, PyTorch, TensorFlow, Keras
- MLOps & Deployment: MLflow, DVC, Docker, FastAPI
- Cloud & DevOps: AWS (EC2, S3, ECR, ECS, CodeDeploy)
- Version Control: Git, GitHub, GitHub Actions
- GenAI & NLP: LangChain, LangGraph, Retrieval-Augmented Generation (RAG).

Personal Projects

REAL ESTATE STREAMLIT APP — December 2025

- Trained a **Real Estate Price Prediction model** using **LightGBM**, trained on **40K+ property listings**, achieving **R² = 0.90** and **MAE ≈ ₹0.6 Cr on test data** through Optuna-based hyperparameter optimization.
- Performed EDA and feature engineering on 40K+ property records, selecting and refining 13 key features including location (region, locality), floor attributes, property configuration, and area, contributing to an R² improvement to 0.90.
- Developed an **end-to-end ML workflow** with **MLflow (experiment tracking & model registry), DVC, Docker, and AWS (S3, EC2, ECR)**, and deployed a **Streamlit application** for analytics and price prediction.

TWITTER SENTIMENT ANALYSIS — August 2025

- Built a sentiment classification model on **50,000+ Twitter posts** using **PyTorch LSTM**, achieving **90% test accuracy**.
- Implemented NLP preprocessing (spaCy, NLTK) to reduce noise and stabilize model training across 50K+ tweets.
- Performed Optuna-based hyperparameter tuning over multiple trials, improving training efficiency and final model accuracy.

HYBRID SPOTIFY RECOMMENDER SYSTEM — May 2025

- Formulated a **hybrid recommender system** combining **collaborative filtering and content-based filtering** on **1M+ user listening records** and **50K+ music metadata entries**.
- Applied EDA on user listening behavior and audio features (genre, tempo, danceability) to define similarity features and user-item representations.
- Built a production-ready recommendation pipeline using cosine similarity, DVC, Docker, GitHub Actions, and AWS (S3, EC2, ECR, CodeDeploy) to enable scalable similarity computation and deployment.

SWIGGY DELIVERY TIME PREDICTION — March 2025

- Built a Delivery Time Prediction Model for Swiggy using Stacking Regression (Random Forest + LightGBM + Linear Regression) on 45k+ records, achieving R² = 0.83 and MAE = 3.13 minutes on test data.
- Performed data cleaning, feature engineering, and EDA on variables like distance, traffic, weather, and order time to uncover key factors influencing delivery duration.
- Deployed the model as a REST API using FastAPI and implemented a production-grade ML pipeline with MLflow, DVC, Docker, GitHub Actions, and AWS (S3, EC2, ECR, CodeDeploy) for automated training, versioning, and scalable deployment.

Education

BACHELOR OF SCIENCE IN MATHEMATICS: SCORED (76%), *Patna Science College, Patna* — 2020-2023

Certifications

DATA SCEINCE MENTORSHIP PROGRAM - CAMPUSX — Feb 2025
MACHINE LEARNING A-Z: AI, PYTHON & R - UDEMY — July 2025