Sourav Raj

Data Scientist

souravraj664@gmail.com 7004928439 Patna, Bihar linkedin.com/in/sourav664/

github.com/sourav664 portfolio/sourav664

PROJECTS

REAL ESTATE PRICE PREDICTION SYSTEM (ML & DASHBOARD) Oct 2025 - Dec 2025

- Collected and analyzed **40K+ real estate property listings**, engineered 13 high-impact features spanning location, floor attributes, and property configuration to improve prediction accuracy and model stability.
- Trained a LightGBM-based regression model, achieved **R2 = 0.90 and MAE = ₹0.6 Cr** on test data, with **Optuna**-driven hyperparameter optimization and MLflow-based logging of parameters and evaluation metrics.
- Built an end-to-end ML pipeline using **MLflow (experiment tracking & Model Registry), DVC, Docker, and AWS (S3, EC2, ECR)**; promoted registered models to production and deployed a **Streamlit application** for real-time analytics and price prediction.

TWITTER SENTIMENT ANALYSIS (NLP & DEEP LEARNING) Aug 2025 - current

- Conducted Exploratory Data Analysis (EDA) on **69K+ Twitter posts** to examine sentiment distribution, class imbalance, text length patterns, and noise characteristics, guiding preprocessing and model design decisions.
- Tracked experiments and performance metrics using **MLflow**, integrated with **Optuna**-based hyperparameter tuning, attained **88% accuracy** and **88% macro F1 score** on the **test** dataset with a **PyTorch LSTM model.**
- Productionized the model by deploying a **Streamlit**-based inference application on **Streamlit Cloud**, enabling real-time predictions via a public web interface (Live App).

HYBRID SPOTIFY RECOMMENDER SYSTEM Apr 2025 - May 2025

- Formulated a hybrid recommender system combining **collaborative filtering and content-based filtering on 1M+ user listening records and 50K+ music** metadata entries.
- Investigated listening behavior and audio features (danceability, energy, loudness, valence, tempo, etc.) to construct similarity-based user and item representations.
- Implemented a production-ready recommendation pipeline leveraging **cosine similarity, DVC, Docker, GitHub Actions (CI/CD), and AWS (S3, EC2, ECR, CodeDeploy)** to enable scalable recommendation generation and deployment.

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 & CI/CD:** Git, GitHub, GitHub Actions
- **Generative AI & NLP:** LangChain, LangGraph, Retrieval-Augmented Generation (RAG)

EDUCATION

Patna Science College - Bachelor of Science, Mathematics

2020 - 2023 Patna

- Scored (76%)

CERTIFICATIONS

- **Data Science Mentorship Program 2.0 - CampusX**
- **Machine Learning A-Z: AI, Python & R- udemy**