Data Scientist

SOURAV RAJ

70049 28439| souravraj664@gmail.com
https://www.linkedin.com/in/sourav664
https://sourav664.github.io/portfolio/
Patna, Bihar

Skills

- Python (Pandas, NumPy, SciPy, MatPlotLib, Seaborn, Plotly)
- SQL (MySQL)
- Scikit-learn, XGBoost, LightGBM
- Pytorch

- AWS (EC2, S3 Bucket, ECR, ECS, CodeDeploy)
- FastAPI
- DVC, MLflow
- Docker
- Git, GitHub, GitHub Action

Projects

TWITTER SENTIMENT ANALYSIS – Personal Project   August 2025

- Developed a **Twitter Sentiment Analysis model** using **PyTorch with LSTM**, processing 50,000 tweets with **SpaCy and NLTK** for text preprocessing.
- Optimized model hyperparameters with **Optuna**, achieving **97% accuracy on training data** and **90% accuracy on test data**.
- Implemented an end-to-end pipeline for **data cleaning, tokenization, model training, and evaluation**, ensuring reproducibility and scalability**.**

HYBRID SPOTIFY RECOMMENDER SYSTEM – Personal Project   May 2025

- Built a **Hybrid Spotify Recommender System** using **Collaborative Filtering** and **Content-Based Filtering** on **1M+ listening history records** and **50k+ music metadata entries**. Transformed and processed client's data by using DAX and Excel to ensure data completeness and validity.
- Conducted **Exploratory Data Analysis (EDA)** to understand user listening patterns and music attributes such as genre, tempo, and danceability.
- Developed a **production-ready ML pipeline** with **DVC, Docker, GitHub Actions, and AWS (S3, EC2, ECR, CodeDeploy)** for automated training, versioning, and deployment at scale.

SWIGGY DELIVERY TIME PREDICTION – Personal Project   March 2025

- Built a **Delivery Time Prediction Model** for Swiggy using **Stacking Regression (Random Forest + LightGBM + Linear Regression)** on 45k+ records, achieving **R^2 = 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.
- Designed a **production-grade ML pipeline** with **MLflow (experimentation & model staging), DVC, Docker, GitHub Actions, and AWS (S3, EC2, ECR, CodeDeploy)** for automated training, versioning, and deployment.

Education

BACHELOR OF SCIENCE IN MATHEMATICS (SUBSIDIARY: STATISTCS)   2020 - 2023
Patna Science College, Patna

Certifications

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