(+91) 86601-82267
Bangalore, Karnataka
harshpandey472@gmail.com

HARSH PANDEY

Data Scientist

github.com/tortejumpy
linkedin.com/in/harsh472

Profile Summary

Highly motivated learner with hands-on experience in Python, SQL, machine learning, deep learning, and NLP. Skilled in building end-to-end data-driven and AI solutions, extracting insights, and delivering business-ready analytical outcomes.

Skills

Programming Languages	Python, SQL
Data Analytics	Advanced MS Excel, Power BI, Matplotlib, Seaborn, Prompt-based Analysis
Libraries & Tools	Pandas, NumPy, Scikit-learn, TensorFlow, Jupyter Notebook, Git/GitHub, LangChain, Hugging Face
Core Competencies	Exploratory Data Analysis (EDA), Statistical Analysis, Data Visualization, Feature Engineering, RAG
Soft Skills	Analytical Thinking, Problem Solving, Effective Communication, Data Storytelling

Technical Experience

LLM-Powered Document Intelligence & Retrieval System **Jan 2025**

Project Link

- Built multiple LLM-powered document intelligence pipelines using LangChain, OpenAI, Google PaLM, FAISS, and Pinecone, enabling PDF, multi-PDF, CSV chat, and summarization through RAG-based retrieval workflows..
- Optimized semantic search and response accuracy using text chunking, embeddings, conversational memory, and agents, achieving fast, context-aware answers across large documents with scalable vector search.

Text Intelligence: Sentiment Analysis of IMDb Reviews and News Article Classification **Oct 2025**

Project Link

- Built a scalable multi-class NLP pipeline on 50K+ articles using TF-IDF (uni/bi-grams) and Word2Vec with Logistic Regression, Naive Bayes, and SVM; applied cross-validation and GridSearchCV to achieve approx 60% weighted F1-score
- Developed a sentiment analysis system on 50K IMDb reviews with advanced text preprocessing and TF-IDF features; evaluated LR, SVM, NB, and RF via confusion-matrix-driven error analysis, achieving 87.48% accuracy with interpretable TF-IDF+LR

SmartPricing And Retention: Data-Driven Solutions for Airbnb and Telecom Industries **Sept 2025**

Project Link

- Built an end-to-end regression pipeline using Python and scikit-learn (Pipelines, ColumnTransformer, OHE, scaling) with EDA and feature engineering; trained Ridge, Random Forest, and Gradient Boosting models with 5-fold CV and GridSearchCV, achieving 71% of variation and average difference of $50.
- Telecom Customer Churn Prediction & Retention Analytics (MachineLearning): Developed a production-ready churn model using pandas, NumPy, and scikit-learn with SMOTE and GridSearchCV; trained Logistic Regression, Random Forest, and XGBoost models achieving high ROC-AUC and improved churn recall, and translated SHAP insights into targeted retention strategies.

Walmart Retail Insights Optimization Using MySQL **Aug 2025**

Project Link

- Conducted end-to-end sales and profitability analysis on multi-branch Walmart transaction data using advanced MySQL (CTEs, window functions, LAG, RANK, NTILE) to assess growth, margins, customer segments, and product performance across cities.
- Derived actionable business insights through anomaly detection and repeat-customer analysis, identifying 33% high-value customers, 50+ anomalous transactions, top 5 customers driving 36%+ revenue, and peak sales periods to support targeted marketing strategies.

Education

B.E. Information Science and Engineering, *AMC Engineering College*	May 2025
Senior Secondary(PUC), *Ramaiah P.U Composite College*	July 2021

Achievements / Certifications

- MongoDB Sponsor Prize Winner @ NMIT Hackathon
- Complete Data Science Bootcamp 2025 - Udemy
- Data Science With AI – Internshala Training
- Complete Generative AI Course With Langchain and Huggingface - Udemy