

HIMANSHI RATHORE

Data Scientist • ML • Agentic AI • NLP • Data Analyst

Phone: 9718431726 • Email: himanshirathore25102005@gmail.com

LinkedIn: [linkedin.com/in/himanshi-rathore-hr2520](https://www.linkedin.com/in/himanshi-rathore-hr2520) • GitHub: github.com/Himanshi252005

SUMMARY

Data Science graduate and early-career Data Scientist with hands-on experience taking machine learning and NLP solutions from messy, real-world data all the way to deployed, user-facing products. Strong in exactly the parts this role demands: gathering and cleaning data through web scraping and APIs, engineering features with domain reasoning, and validating models with statistical hypothesis testing, with a clear grasp of correlation versus causation and when a simple linear model outperforms a complex one (OLS versus GBM). Delivered an NLP fake-news detection model at 94%+ accuracy, built and deployed a multi-agent research system that autonomously gathers and analyzes web data, and produced statistically defensible predictive models. Fluent in Python and SQL, working knowledge of R, across the full machine learning workflow, resourceful, business-minded, and comfortable owning a problem end to end.

EXPERIENCE

Data Analyst Intern | Early-stage B2B Startup

Oct 2025 – Dec 2025

Sales & Engagement Analytics

- Cleaned and performed EDA on 3 real-world business datasets (sales, engagement, performance), uncovering 4 key user-behavior patterns flagged by the product team for roadmap decisions.
- Built Power BI and Tableau dashboards tracking 10+ KPIs, reducing manual reporting effort by an estimated 35% through automated data pipelines.
- Documented full analytical workflows and presented findings to non-technical stakeholders, resulting in 2 process changes adopted by the operations team.

PROJECTS

Agentic Research Assistant | Python, LangChain, LangGraph, Groq LLM, Tavily, Streamlit

- Built and deployed a multi-agent research system that autonomously searches the web, scrapes the most relevant sources, writes a structured report, and self-critiques it, orchestrating four LangChain/LangGraph agents (search, reader, writer, critic) in a single pipeline.
- Engineered autonomous data-gathering tools using the Tavily search API and BeautifulSoup web scraping, letting the agents collect and clean real-world web data on any topic before analysis, directly tackling the reality that usable data must first be found and gathered.
- Deployed the full application on Streamlit Community Cloud with live output streaming, powered by a Groq-hosted LLM, taking it from a notebook prototype to a public, user-facing product.

AI Emotion-Aware Music Generator | Python, DeepFace, OpenCV, CNN, Streamlit, YouTube Data API v3

- Architected a fully deployed AI web application that identifies a user's facial emotion from an uploaded photo using DeepFace CNN-based recognition, and generates a personalized YouTube playlist in under 5 seconds.
- Assembled the complete pipeline using OpenCV for face detection and image preprocessing, Streamlit for the web interface, and YouTube Data API v3 for real-time music retrieval, making the app publicly accessible from any device.
- Took the project from a Jupyter notebook prototype to a live, user-facing product, demonstrating practical experience deploying production-level machine learning applications.

Insurance Charges Prediction | Python, Pandas, Scikit-learn, Feature Engineering, Statistical Testing

- Implemented an end-to-end predictive modeling pipeline covering data preprocessing (label encoding, one-hot encoding) and custom feature creation, with BMI and age categories crafted from scratch using domain reasoning.
- Applied Pearson Correlation and Chi-Square hypothesis testing to statistically validate feature selection, pinpointing smoking status as the strongest predictor of insurance charges and distinguishing genuine predictive signal from mere correlation.
- Produced a clean, scaled, model-ready dataset with chi-square p-values as the quantitative basis for all keep/drop feature decisions, ensuring statistically defensible inputs.

Superstore Sales Analysis | Excel, SQL, Python, Power BI

- Performed end-to-end data analysis on a 9,994-row retail dataset using Excel, SQLite, Python, and Power BI, covering data cleaning, feature engineering, and business insight generation across all four tools.
- Wrote advanced SQL queries using GROUP BY, CASE WHEN, and window functions to identify that 18.7% of orders were loss-making and that discounts above 40% consistently generated negative profit.
- Built an interactive Power BI dashboard with DAX measures, cross-filtering slicers, and multi-page drill-down reports enabling dynamic analysis of sales, profit, and customer performance by region, segment, and category.

SKILLS

Programming: Python, R, SQL

Machine Learning: Predictive Modeling, Supervised Learning, Classification, Regression, Model Evaluation, Scikit-learn

Modeling Techniques: Linear Regression (OLS), Logistic Regression, Decision Trees, Gradient Boosting (GBM)

NLP: Natural Language Processing (NLP), Text Classification, TF-IDF

Statistics: Hypothesis Testing, Chi-Square Test, Pearson Correlation, Correlation vs. Causation, Statistical Validation, EDA

Data Wrangling: Data Collection, Data Gathering, Web Scraping (BeautifulSoup, Tavily API), Data Cleaning, Feature Engineering

AI / Agents: LangChain, LangGraph, Groq LLM, Streamlit

BI & Visualization: Power BI, Tableau, Excel, Matplotlib, Seaborn

Data & Tools: Pandas, NumPy, MySQL, MongoDB, SQLite, Git, GitHub

Strengths: Data Storytelling, Stakeholder Communication, Business Acumen, Problem Solving

EDUCATION

Bachelor of Computer Applications (BCA) — Data Science

SRM University • CGPA: 9.5 / 10.0 • Relevant Coursework: Machine Learning, Data Structures, DBMS, Statistics, Data Visualization