HAFSA MOTIWALA

Junior / Entry-Level / Associate Data Scientist
Karachi, Pakistan | Open to Remote Roles
Email: motiwala603@gmail.com | LinkedIn: linkedin.com/in/hafsa-motiwala-a9608835b | GitHub: github.com/hafsamotiwala/hafsa-portfolio

SUMMARY

Junior / Entry-Level Data Scientist with experience in data analysis, baseline modeling, and advanced machine learning pipelines. Skilled at structuring workflows, preparing datasets, and building models with measurable performance. Enthusiastic about learning in a **mentor-led environment** and growing into more advanced data science responsibilities.

TECHNICAL SKILLS

Programming & Data Analysis: Python (pandas, NumPy), SQL
Machine Learning & Modeling: scikit-learn; regression & classification models (baseline and ensemble methods), decision trees, tree-based ensembles
Advanced Techniques: Feature engineering, cross-validation, hyperparameter tuning, model stacking, bias–variance trade-off
Evaluation & Statistics: ROC-AUC, F1-score, RMSE, train/validation split, statistical analysis
Data Exploration & Visualization: Exploratory Data Analysis (EDA), data cleaning, matplotlib, seaborn, Jupyter Notebook
Tools & Workflow: Git (local version control), GitHub (repository management, project documentation)

PROJECT EXPERIENCE

Junior Data Scientist (Project-Based) — Personal & Academic Projects
2025 – Present | Remote

Exploratory Data Analysis (EDA) Project | Python, pandas

- Conducted structured exploratory data analysis to understand dataset structure, distributions, missing values, and feature relationships.
- Cleaned and organized raw data using pandas, focusing on data quality and analytical readiness.
- Produced clear visual and written insights to guide downstream modeling decisions.

House Price Prediction (Baseline Regression) | Python, scikit-learn

- Built a **strong baseline regression model** to predict house prices, emphasizing high-quality data cleaning and feature preparation.
- Demonstrated how a well-prepared dataset and simple modeling approach can yield competitive performance.
- Achieved **RMSE of 0.13** on validation data without heavy model complexity.
- Clearly documented assumptions, preprocessing steps, and limitations in a reproducible Jupyter notebook.

Bank Term Deposit Subscription Prediction (Advanced Machine Learning) | Python, Random Forest, XGBoost, LightGBM

- Built an end-to-end machine learning pipeline to predict customer subscription to term deposits using structured banking data.
- Performed **hyperparameter tuning** across Random Forest, XGBoost, and LightGBM models using cross-validation.
- Applied **model stacking** to combine tuned base learners and improve generalization performance.
- Achieved **ROC-AUC of 0.80** on validation data with the final stacked model.
- Analyzed feature importance and model behavior to explain key drivers of customer response.

SQL Business Analysis Project | SQL

- Solved business-oriented analytical questions using structured SQL queries.
- Applied joins, aggregations, subqueries, and window functions to analyze sales and customer data.
- Focused on clarity, correctness, and reproducibility of queries for real-world analytical workflows.

EDUCATION

Higher Secondary School (Intermediate) — Computer Science & Statistics
Board of Intermediate Education, Karachi | *2022 – 2024*
Relevant Coursework: Probability & Statistics, Computer Science, Programming Fundamentals

CONTINUOUS LEARNING

- Self-taught data science through structured practice, technical documentation, online tutorials, and applied projects
- Consistently use generative AI tools (e.g., ChatGPT) as a **learning and productivity aid**, while independently validating logic, results, and implementations
- Ongoing focus on strengthening fundamentals in statistics, machine learning, and analytical problem-solving