WEEK 2 REPORT

House Price Prediction

JULY 30, 2025

MUHAMMAD KAIF UR REHMAN
STU-DS-251-232

Table of Contents

Project Overview 1

Objectives 1

Methodology 2

Feature Engineering 2

Outlier Analysis 3

Predictive Modeling 3

Results and Findings 4

Challenges Faced 4

Recommendations 5

Conclusion 6

House Price Analysis and Prediction Report
Digital Empowerment Network – Data Science Week 02
Mentor: Ali Mohiuddin Khan

Project Overview

This project analyzes a comprehensive dataset of house prices to understand the key factors influencing property values, identify outliers, and develop predictive models for future price estimation. The analysis focuses on the Pakistani real estate market using data from Zameen.com, providing insights into market dynamics and property valuation patterns.

The dataset contains various property features including location, size, number of bedrooms/bathrooms, property type, and other relevant characteristics that influence house pricing decisions.

Objectives

The primary objectives of this project are:

1. Clean and explore the house price dataset to understand data quality and distributions.
2. Identify the most important factors affecting house prices.

3. Identify and investigate properties with unusually high or low prices.
4. Develop machine learning models to predict house prices.
5. Provide actionable insights for buyers, sellers, and real estate professionals.

Methodology

Data Processing Pipeline:

- **Data Loading and Initial Exploration:** Load the dataset, inspect structure, check missing values and statistics.
- **Data Cleaning:** Handle missing/inconsistent entries, standardize formats.
- **EDA:** Visualize distributions, analyze correlations and patterns.
- **Feature Engineering:** Derive new useful variables, encode categorical data.
- **Outlier Analysis:** Identify extreme values using IQR and Z-score methods.
- **Model Development:** Train/test split, build and evaluate multiple regression models.

Data Cleaning and Exploration

- Dataset includes features like location, area, bedrooms/bathrooms, year built, and price.
- Missing values were addressed using median/mode imputation.
- Mixed units and inconsistent formats were standardized.
- Exploratory analysis revealed a right-skewed price distribution.
- Property size and location were strongly correlated with price.

Feature Engineering

Key Features Created:

- **Property Age** = 2025 - Year Built
- **Price per Square Foot** = Price ÷ Area
- **Bedroom/Bathroom Ratio** = Bedrooms ÷ Bathrooms
- **Categorical Encoding**: Label encoded locations and property types

Feature Importance Highlights:

1. Property Size
2. Location
3. Number of Bedrooms
4. Property Age
5. Property Type

Outlier Analysis

Detection Methods:

- **IQR Method** to identify extreme high/low prices
- **Z-Score Method** for statistical anomaly detection

High-Value Outliers:

- Premium locations, large size, luxury features, recent construction.

Low-Value Outliers:

- Small size, old condition, poor locations, possible structural issues.

Findings:

- 5–8% of properties were classified as outliers.
- Some represent market opportunities; others are entry errors.

Predictive Modeling

Models Implemented:

- Linear Regression
- Random Forest Regressor
- Decision Tree Regressor

Training & Evaluation:
Models were trained using an 80/20 train-test split. Evaluation metrics included MSE, RMSE, and R^2.

Performance Summary:

Model	R² Score	RMSE
Linear Regression	0.7234	53,365
Random Forest	0.8156	43,510
Decision Tree	0.7892	46,445

Best Model:
Random Forest – highest accuracy and best error metrics.

Results and Findings

Insights:

- **Top Predictors:** Property size, location, and property type.
- **Market Behavior:** Size-to-price ratio shows strong correlation (r ≈ 0.78).
- **Model Accuracy:** 81.56% accuracy with Random Forest.
- **Price Variation:** Price per square foot varies up to 60% by location.
- **Outlier Insights:** Valuable extremes both on high and low end.

Random Forest Feature Importance:

1. Property Size (35.2%)
2. Location (28.7%)
3. Bedrooms (15.4%)
4. Property Age (12.1%)
5. Property Type (8.6%)

Challenges Faced

Data Quality Issues:

- 15–20% missing data in critical fields
- Inconsistent area and price formats
- Skewed price distributions
- Presence of non-genuine outliers

Modeling Issues:

- Multicollinearity in features
- Heteroscedasticity in residuals
- Balancing interpretability vs. accuracy

Solutions:

- Imputation and standardization
- Feature selection and transformation
- Use of ensemble models for robustness

Recommendations

For Data Collection:

- Standardize data entry and formats
- Include additional fields: amenities, condition, nearby landmarks
- Use automated validation tools

For Modeling:

- Use advanced models (e.g., XGBoost, LightGBM)
- Explore neural networks for non-linearity
- Consider stacking/blending models for improved performance

For Business:

- Build interactive prediction tools (e.g., dashboards)
- Offer prediction intervals for risk-aware decision-making
- Segment markets (e.g., by city, property type) for better targeting

For Future Analysis:

- Time series modeling for price trends
- Location heatmaps for visualizing hotspots
- Integrate external data like interest rates, demographics

Conclusion

This project successfully achieved:

- **Robust EDA:** Identified top price-driving features
- **Effective Modeling:** Developed accurate Random Forest model
- **Market Understanding:** Deep insights into Pakistani real estate
- **Practical Recommendations:** Actionable strategies for data, modeling, and deployment

Achievements:

- R^2 Score: **0.8436** (Random Forest)
- Outlier coverage: **5–8%** of listings
- Key drivers: **Size**, **Location**, **Age**, **Type**
- Roadmap for real-time implementation and scaling