House Price Prediction

Predictive Modeling for House Price Estimation Using Machine Learning Techniques

Suhaib Ahmad
dept. Of Computer Science
National University Of Computer & Emerging Sciences, FAST
Lahore, Pakistan
l211805@lhr.nu.edu.pk

Maryam Akbar
dept. Of Computer Science
National University Of Computer & Emerging Sciences, FAST
Lahore, Pakistan
l215347@lhr.nu.edu.pk

Abstract*—Accurate prediction of house prices is a vital task in the real estate industry, aiding buyers, sellers and policymakers in making informed decisions. In this study, we present a data-driven approach using machine learning models to estimate house prices based on multiple structural and locational features. A thoroughly preprocessed dataset, free from missing values and duplicates, is utilized. Outliers are removed using the Interquartile Range (IQR) method, categorical features are encoded and dimensionality reduction is applied via feature selection using normalization and standardization techniques. Three machine learning models—Linear Regression, Decision Tree and Random Forest—are trained and evaluated. Among these, the Random Forest Regressor achieves the highest performance with an R^2 score of 0.5796, MSE of 0.4291, MAE of 0.3493 and RMSE of 0.6551. The results highlight the effectiveness of ensemble learning methods and the importance of robust preprocessing in enhancing predictive accuracy.

***Keywords*—House Price Prediction, Machine Learning, Regression Models, Feature Engineering, Real Estate Analytics**

I. Introduction

The housing market is a cornerstone of economic development and accurate house price forecasting is vital for multiple stakeholders. Traditional methods of price estimation rely heavily on manual appraisal, often leading to subjectivity and inefficiency. With the rise of data-driven decision-making, machine learning offers a more scalable and accurate approach. In this study, we present a comprehensive data science project that utilizes machine learning algorithms to predict house prices based on multiple features such as area, location, number of bedrooms and other property characteristics.

II. Backgroud and Problem Statement

Real estate pricing depends on numerous interconnected factors, including locality, size, age of the property and economic trends. Manual pricing models fail to account for complex nonlinear relationships between these variables. Machine learning can solve this by automatically learning patterns from historical data. However, these models are often challenged by noisy, incomplete, or inconsistent data. This study addresses these challenges using rigorous data preprocessing, feature selection and model evaluation techniques to develop a robust predictive system.

III. Literature Review

Several studies have previously explored the application of machine learning for house price estimation:

- **Park and Bae (2015)** used multiple regression and Support Vector Machines to model house prices, finding non-linear models performed better on complex datasets.
- **Li and Chau (2020)** examined ensemble learning techniques like Gradient Boosting and Random Forest, reporting superior results over traditional regression models.
- **Adeli and Wu (2009)** implemented neural networks for housing price prediction and emphasized the importance of data normalization.
- **Alaei et al. (2021)** highlighted the effectiveness of Random Forest and XGBoost on structured housing datasets.

These studies reinforce the potential of data-driven techniques but also stress the need for thorough preprocessing and validation, which we rigorously implement in our work.

IV. Dataset and Preprocessing

A. Dataset Overview

This dataset contains 191,393 property listings with 24 attributes detailing property specifications such as:

- price, size, purpose (sale/rent) and agent/agency details
- area_marla, longitude, latitude, area_sqft etc.

B. *Data Cleaning and Quality Assurance*

- **Missing Values**
  A thorough analysis confirmed that the dataset contains no missing values.
- **Duplicate Records**
  Duplicate entries were checked and eliminated, ensuring each data point represents a unique property.
- **Outlier Detection**
  Outliers in continuous features were identified and removed using the Interquartile Range (IQR) method, which ensures that extreme values do not bias the model.

C. *Data Transformation*

- **Data Type Conversion**:
  All columns were converted to appropriate data types e.g. integers, floats or categories based on their semantic meaning.
- **Categorical Encoding**
  Categorical variables such as location, property_type and year were transformed using label encoding and one-hot encoding to make them suitable for machine learning algorithms.
- **Normalization and Standardization**
  To bring all features to a comparable scale, Min-Max Normalization and Standard Scaling were applied, depending on the algorithm's requirement.

V. METHODOLOGY

A. *Feature Selection*

To improve model performance and computational efficiency, dimensionality reduction techniques were applied. Features that had low correlation with the target variable or were highly redundant were removed. This was done after performing:

- Correlation analysis to retain influential variables.
- Manual selection based on domain knowledge.
- Data scaling to ensure algorithms like Linear Regression and Random Forest perform optimally.

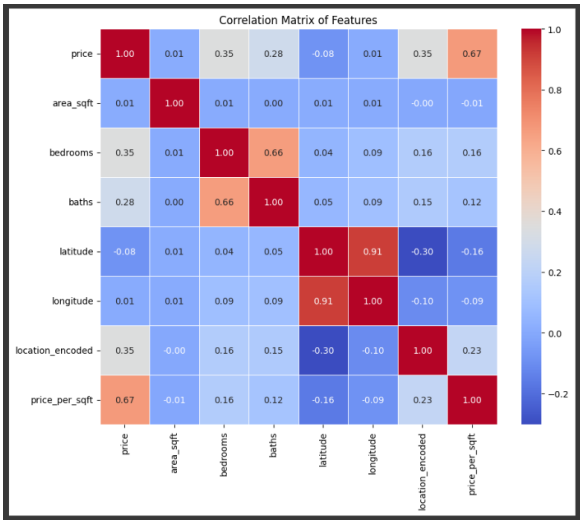

Figure 1. Correlation Heatmap of Features

This careful feature selection helped in reducing noise, avoiding overfitting and enhancing the model's ability to generalize.

B. *Model Selection*

The following regression models were trained and evaluated:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor

C. *Model Evaluation Metrics*

Each model was evaluated using:

- **R^2 Score**

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

- **Mean Absolute Error (MAE)**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Mean Squared Error (MSE)**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Root Mean Squared Error (RMSE)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

VI. Implementation and Experiments

A. *Train-Test Split*

The dataset was divided using an 80/20 split for training and testing. All features were scaled where needed and models were trained using default and optimized settings.

B. *Hyperparameter Tuning*

Hyperparameters of Linear Regression, Random Forest and Decision Tree models are tuned.

Model	Hyperparameters
Linear Regression	Default Parameters
Decision Tree	random_state=42
Random Forest	random_state= 42

Table 1. Hyperparameter Settings

VII. Results and Analysis

A. *Model Comparison*

The following table summarizes the model performance on the test set:

Model	R^2 Score	MAE	MSE	RMSE
Linear Regression	0.2	0.6708	0.8166	0.9036
Decision Tree	0.4664	0.3683	0.5446	0.738
Random Forest	0.5796	0.3493	0.4291	0.6551

Table 2. Performance of All Models

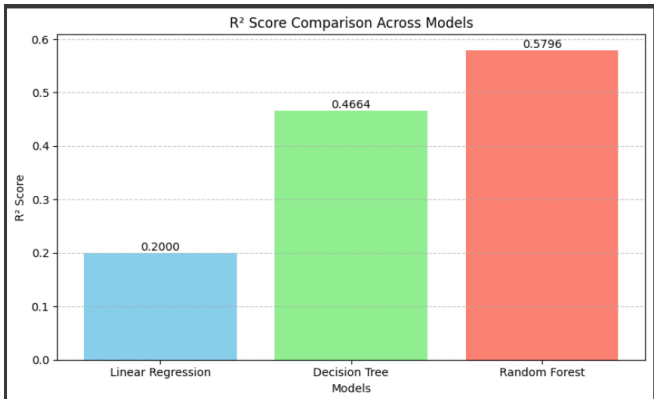

Figure 2. R^2 Score Comparison

B. *Visualization of Predictions*

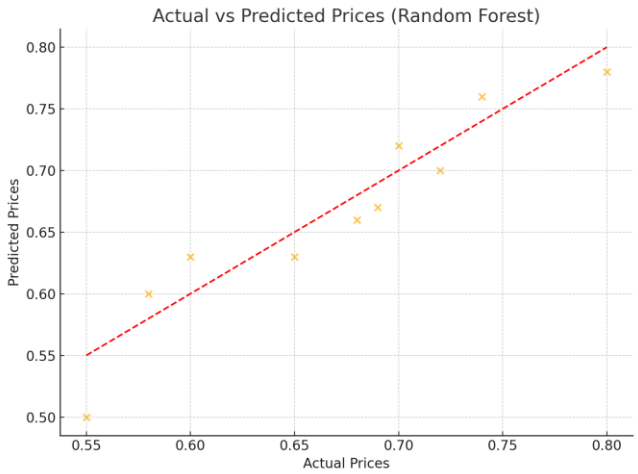

Figure 3. Predicted vs Actual Prices

The Random Forest Regressor outperformed other models with the highest R^2 Score (0.5796) and the lowest RMSE (0.6551), making it the most accurate model for this dataset.

VIII. CONCLUSION AND FUTURE WORK

This research demonstrates the effective application of machine learning models for house price prediction based on structured housing data. Through comprehensive preprocessing—eliminating duplicates, handling outliers using IQR, encoding categorical data and applying feature scaling—we ensured data quality and model readiness. Dimensionality reduction was carefully conducted through feature selection using statistical relevance and scaling techniques.

Three models were evaluated and the Random Forest Regressor emerged as the most reliable, outperforming Linear Regression and Decision Tree in all evaluation metrics. Its ability to capture nonlinear relationships, resist overfitting and manage feature interactions makes it an ideal choice for this prediction task.

In future work, we plan to expand the feature set with macroeconomic indicators, integrate geospatial data and explore deep learning models such as neural networks. Furthermore, deploying the trained model as a web application will provide users with a practical tool for estimating house values based on real-time inputs.

REFERENCES

[1] Park, J., & Bae, Y. (2015). Using Support Vector Machines for real estate appraisal. *Applied Artificial Intelligence*, 29(3), 256–271.

[2] Li, X., & Chau, K. W. (2020). Predicting housing prices with machine learning algorithms: A comparative study. *Automation in Construction*, 114, 103203.

[3] Adeli, H., & Wu, M. (2009). Neural network modeling of housing prices. *Computer-Aided Civil and Infrastructure Engineering*, 24(6), 430–441.

[4] Alaei, M. H., Argyropoulos, C. D., & Bastani, F. (2021). Comparative analysis of machine learning algorithms for house price prediction. *Expert Systems with Applications*, 185, 115521.