Research & Reviews: Journal of Statistics [E-ISSN: 2278-2273]

Volume 15 Issue 01 2026

Review Article

Received: 07/04/2026

Accepted: 17/04/2026

Published: 30/04/2026

House Price Estimation Using Linear Regression: A Machine Learning Perspective

Varuchi Maurya*

Student, Department of AI-DS , Dr. Akhilesh Das Gupta Institute of Professional Studies (ADGIPS), Delhi, India

Abstract— **House price prediction plays a crucial role in the real estate industry, helping buyers, sellers, and investors make well-informed decisions. Accurate estimation of property values enables stakeholders to assess market trends, plan investments, and minimize financial risks. This study focuses on the application of linear regression, a fundamental and widely used machine learning algorithm, to predict house prices based on multiple influencing factors. These factors include location, property size, number of bedrooms, amenities, and other relevant characteristics that significantly impact pricing.**

Linear regression models the connection between input variables (features) and an output variable (such as house price) using a linear equation. By examining past housing data, it uncovers patterns and relationships among variables, enabling it to predict outcomes for new, unseen cases. Its straightforward nature and ease of interpretation make it a popular option, particularly for initial analysis and baseline modeling in real estate applications.

To assess how well the model performs and how reliable it is, metrics like Mean Squared Error (MSE) and the coefficient of determination (R^2) are used. These measures help evaluate prediction accuracy and indicate how effectively the model explains variations in house prices. The results of this study indicate that linear regression can deliver reasonably accurate predictions when the data is well-structured and relevant features are selected.

Overall, this approach demonstrates that linear regression serves as a simple, efficient, and practical tool for house price estimation, offering valuable support for real estate market analysis and decision-making processes.

Keywords—**House Price Prediction, Linear Regression, Correlation, R^2 Score, Mean Squared Error, Root Mean Squared Error, Absolute Mean Error.**

***Abbreviations*-MSE, RMSE, AME, R^2, SVR, ANN, CNN, HPM, MLR**

I. INTRODUCTION

The real estate market plays a vital role in the global economy, influencing investment decisions, financial planning, and government policies. Accurate prediction of house prices is essential for a wide range of stakeholders—buyers, sellers, investors, and policymakers—so they can make well-informed decisions. In the past, property valuation depended largely on manual evaluations and expert judgment, which often led to subjective and inconsistent outcomes. With the growing emphasis on data-driven approaches and the advancement of machine learning, predictive models have gained popularity as reliable tools for estimating property values [1].

Predicting house prices plays a key role in the real estate sector. Prices are influenced by multiple factors, including location, number of bedrooms, total area, and available amenities. To analyze these variables and generate accurate predictions, machine learning methods—especially regression techniques—are widely applied [2].

Among these methods, linear regression stands out as one of the most basic and commonly used algorithms for predictive analysis. It works by identifying the relationship between independent variables (features) and the dependent variable (house price) through fitting a linear equation to the observed data. The model assumes that the relationship between the input features and the house price can be represented by a straight line modeling [3]. It establishes a relationship between independent variables (features) and the dependent variable (house price) by fitting a linear equation

to the observed data. The model assumes that the relation- ship between the input features and the house price can be represented by a straight line [4].

II. Literature Review

House price prediction has been a widely researched topic in real estate analytics, with numerous studies exploring different models and methodologies to improve accuracy [5]. Traditional statistical approaches, such as Hedonic Pricing Models (HPM) and Multiple Linear Regression (MLR), have been commonly used to estimate housing prices based on factors like size, location, and amenities. Early studies, such as those by Rosen (1974), established the foundation for hedonic pricing, em- phasizing the impact of property characteristics on market value. More recent research has explored the limitations of linear regression, particularly its inability to capture non-linear relationships in complex real estate markets. In response, machine learning models, including Decision Trees, Random Forests, and Support Vector Regression (SVR), have been introduced to improve predictive performance. Studies by Kumar & Ravi (2016) and Zheng et al. (2020) demonstrated that ensemble methods, such as Gradient Boosting Machines (GBM) and XGBoost, significantly enhance accuracy by cap- turing intricate patterns in large datasets [6].

Additionally, the integration of geospatial and socioeconomic factors has been highlighted in research by Brunsdon et al. (1996) and Dube´ et al. (2017), showing that spatial dependen- cies and external economic conditions strongly influence house prices. Recent progress also integrates deep learning methods like Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) to automatically extract features and assess properties from images. Furthermore, studies have emphasized the importance of real-time data from online real estate platforms and economic indicators to improve forecasting accuracy [7]. Despite the effectiveness of these models, challenges remain, including data quality issues, feature selection complexity, and the interpretability of black- box models. This literature review underscores the evolution of house price prediction methods, highlighting the need for hybrid models that balance accuracy, interpretability, and real-time adaptability in dynamic housing markets [8].

III. Methodology

A. Data Collection

- Gathering real estate datasets from online sources such as Zillow or Kaggle (Fig 1).

B. Data Cleaning and Preprocessing

- Handling missing values, removing outliers, and normalizing data.

C. Model Development

- Applying linear regression to establish a relationship between house attributes and prices.

D. Model Evaluation

- Measuring model performance using R^2 score, MSE, and MAE.

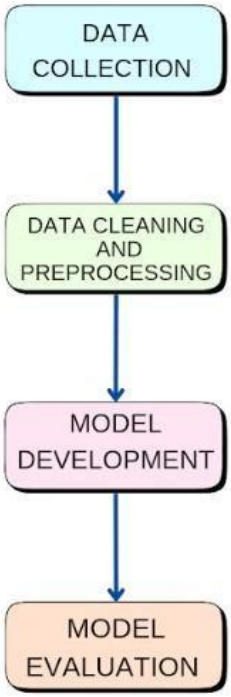

Fig. 1: Flowchart

The above flowchart shows how the process is carried out in prediction of house price using linear regression [9].

IV. Implementation

A. Data collection

Data collection is a crucial step in building a reliable house price prediction model, as the quality and relevance of data directly impact the accuracy of predictions. The data for house price prediction is collected from various sources, including publicly available datasets such as the Ames Housing Dataset and Boston Housing Dataset from Kaggle, real estate platforms like Zillow and Redfin, government property records, and real estate agency reports. Additionally, data can be scraped from online real estate listings to obtain the latest market trends. The dataset consists of multiple features that influence house prices [10]. These features can be categorized into structural attributes (e.g., square footage, number of bedrooms and bathrooms, house type, and age of the property), location- based attributes (e.g., neighborhood, crime rate, proximity to schools, hospitals, parks, and public transportation) (Fig 2).

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basemer
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180.0	
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170.0	4
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770.0	
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050.0	9
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680.0	
5	7237550310	20140512T000000	1230000.0	4	4.50	5420	101930	1.0	0	0	...	11	3890.0	15
6	1321400060	20140627T000000	257500.0	3	2.25	1715	6819	2.0	0	0	...	7	1715.0	
7	2008000270	20150115T000000	291850.0	3	1.50	1060	9711	1.0	0	0	...	7	1060.0	
8	2414600126	20150415T000000	229500.0	3	1.00	1780	7470	1.0	0	0	...	7	1050.0	7
9	3793500160	20150312T000000	323000.0	3	2.50	1890	6560	2.0	0	0	...	7	1890.0	

10 rows × 21 columns

Fig. 2: Unstructured Dataset

B. Data Cleaning and Preprocessing

To improve the accuracy and dependability of the house price prediction model, the collected dataset is subjected to a data cleaning and preprocessing stage [11]. This step is crucial in handling miss- ing values, removing inconsistencies, transforming variables, and preparing the data for linear regression analysis (Fig 3).

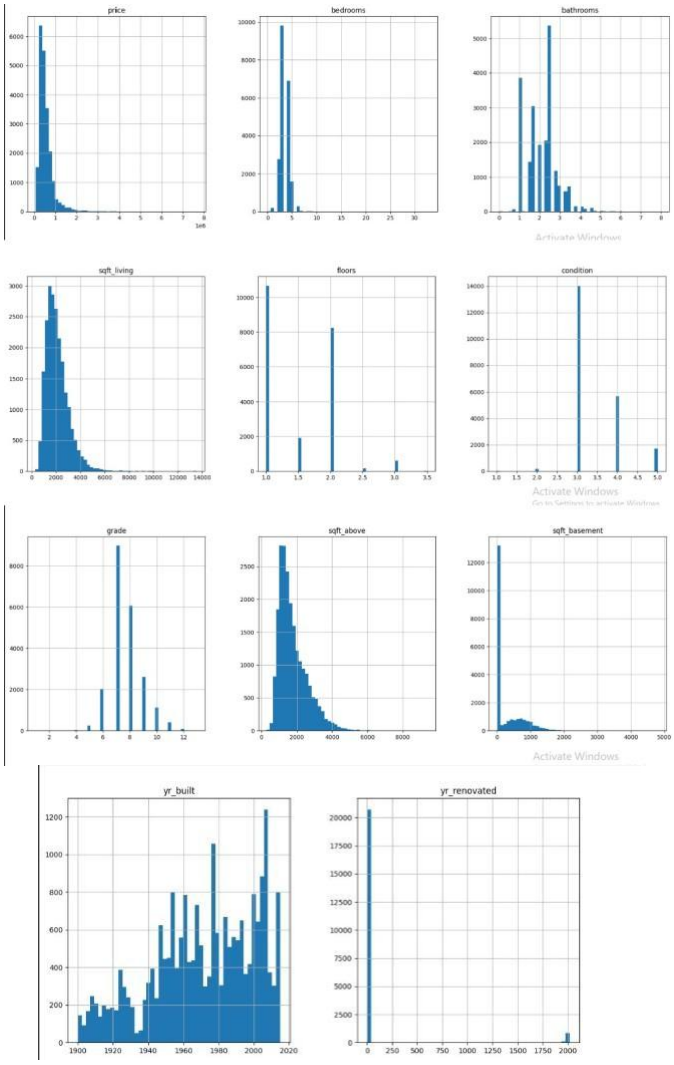

Fig. 3: Distribution of Attributes

C. Model Development

1) Algorithm Selection: Selecting the right model is crucial for accurate house price prediction. The selection of an algorithm is influenced by factors like the complexity of the data, the need for interpretability, and the efficiency of computation [12]. In this study, Linear Regression is selected as the primary algorithm due to its simplicity, interpretability, and effectiveness in modeling relationships between numerical features and house prices [13].

In linear regression, the dependent variable's value is determined based on the independent variables. The value being predicted depends on how strongly it is related to those independent variables, and this relationship is measured using correlation (Fig 4).

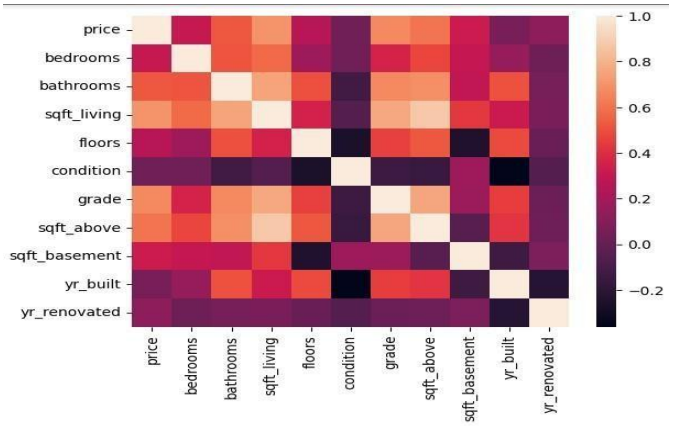

Fig. 4: Heatmap for showing Correlation

The independent variables are plotted on the x-axis while the dependent variables are plotted on the y-axis. The formula

for multiple regression is given by :

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n \quad (1)$$

here, a is the y-intercept, y is the dependent variable, $x_1, x_2, ..., x_n$ are the independent variables and $b_1, b_2, ..., b_n$ are the coefficients for the independent variables respectively.

2) Training the model: Model training is carried out by supplying the cleaned and preprocessed dataset to a linear regression algorithm so it can identify relationships between the independent variables (features) and the dependent variable (house price).

To do this, the dataset is divided into two portions: 80% is used for training the model, while the remaining 20% is reserved for testing. This approach allows the model to learn patterns from one subset and then be evaluated on previously unseen data [14]. The model is based on the multiple linear regression framework, where house price is represented as a linear combination of factors such as square footage, number of bedrooms, and number of floors. Using Scikit-Learn's Linear Regression function, the model is fitted to the training data (Table 1). After the training process, the coefficients for each feature are obtained as follows:

Attributes	Coefficients
Bedrooms	-4.62019078e+04
Bathrooms	4.23588267e+04
sqft livng	1.29352678e+02
Floors	3.35883649e+04
Condition	2.22162529e+04
Grade	1.30501032e+05
sqft above	5.14309869e+01
sqft basement	7.79216914e+01
Year Built	-3.79589152e+03
Year Renovated	1.98399425e+01

TABLE I: Coefficients of Attributes

D. Model Evaluation

Once the linear regression model has been trained, it is important to assess its performance to understand how accurately it predicts house prices. This evaluation is carried out using several statistical metrics that capture aspects such as accuracy, error, and how well the model fits the data. Commonly used metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R^2 score (coefficient of determination) [15].

For this evaluation, the R^2 score has been selected as the primary metric.

The R^2 score, also known as the coefficient of determination, measures how effectively the independent variables account for the variation in the dependent variable. In other words, it indicates how well the model fits the observed data.
The Formula for calculating R^2-score is as follows:

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

here,

y_i = Actual house price
$\hat{y}_i$ = Predicted house prices
$\bar{y}$ = Mean of actual house prices
$\sum(y_i - \hat{y}_i)^2$ = Sum of squared errors (SSE)
$\sum(y_i - \bar{y})^2$ = Total sum of squares

if, $R^2 = 1$: The model perfectly predicts the target variable (best fit).
$R^2 = 0$: The model does not explain any variance in the target variable.
$R^2 < 0$: The model performs worse than a simple mean-based prediction, indicating poor fit.
Higher R^2 values (closer to 1) indicate that the model explains most of the variability in the data, making it a better predictor [16].

A higher R^2 is generally preferred, it does not always guarantee a good model. An extremely high R^2 may signal overfitting, meaning the model has learned the training data too closely instead of capturing patterns that apply well to new, unseen data. Conversely, a low R^2 does not always mean the model is bad, especially in domains like economics or real estate, where many external factors influence prices [17].

V. Results

Scatter plots are commonly used to visualize errors. In this study, we used a residual plot, which shows the difference between actual and predicted values. Ideally, these residuals should be randomly scattered around zero. Residuals around zero indicate that the error generated are small. This means that the predicted values are close to actual values and the model fits the data well (Fig 5).

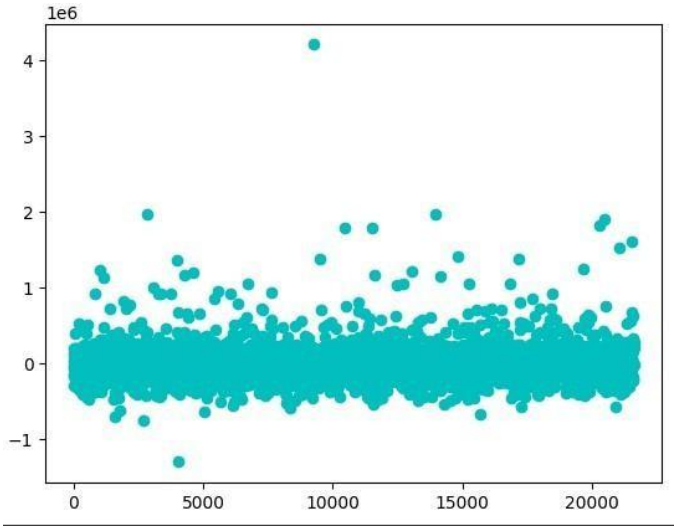

Fig. 5: Scatter Plot for visualising error

The R^2-score calculated for this model is 0.6320477597526137. This score indicates that 63% of the variance in housing prices is explained by our model while the remaining 37% is due to some missing factors like noise,location.

R^2 Score = 0.63 (63%)

VI. Acknowledgement

We express our sincere gratitude to all those who contributed to the successful completion of this research on house price prediction using linear regression. First and foremost, we extend our heartfelt thanks to our mentors and faculty members for their invaluable guidance, constructive feedback, and continuous support throughout this study. Their knowledge and support have played a crucial role in guiding our research.

We also acknowledge the support of various data sources and real estate platforms that provided access to crucial datasets, enabling the development and validation of our predictive model. Additionally, we appreciate the contributions of researchers whose prior work in the field of machine learning and real estate analytics has served as a foundation for our study.

Finally, we extend our gratitude to our peers, friends, and family for their unwavering support and encouragement. Their unwavering support has been essential in completing this research successfully.

VII. Conclusion

This study explored the use of Linear Regression for predicting house prices based on various factors such as size, location, and amenities. The research followed a structured approach, including data collection, preprocessing, model development, evaluation, and optimization. The findings demonstrate that Multiple Linear Regression effectively captures relationships between house features and price, making it a valuable tool for real estate valuation.

The results indicate that factors like square footage, number of bedrooms, and neighborhood quality significantly impact house prices. Although Linear Regression offers a straightforward and easy-to-interpret model, its accuracy can be impacted by issues like multicollinearity, outliers, and non-linear relationships. To address these limitations, regularization techniques (Ridge and Lasso Regression) and advanced machine learning models (e.g., Decision Trees, Neural Networks) can be explored in future research.

In conclusion, Linear Regression serves as a strong baseline model for house price prediction, offering insights for buyers, sellers, and real estate professionals. However, incorporating more complex models and larger datasets could further enhance predictive accuracy and market analysis.

VIII. Future Scope

The study on house price prediction using linear regression provides a strong foundation, but there are several areas for future improvement and expansion. One key area is the incorporation of advanced machine learning models, such as Random Forest, Support Vector Regression (SVR), and Neural Networks, which can capture complex, non-linear relationships in housing data and potentially improve accuracy. Additionally, real-time data integration from real estate platforms and economic indicators can enhance prediction reliability by accounting for market fluctuations. Another promising direction is the use of geospatial analysis, incorporating factors such as proximity to essential services, crime rates, and neighborhood development trends.

References

[1] Bhagat, N., Mohokar, A., Mane, S. (2016). “House Price Forecasting using Data Mining”. International Journal of Computer Applications, 152(2),23–26.

[2] M. Bhuiyan and M. A. Hasan, “Waiting to Be Sold: Prediction of Time-Dependent House Selling Probability”, IEEE 2016 International Conferenceon Data Science and Advanced Analytics (DSAA), Montreal, QC,2016, pp. 468-477.

[3] N. N. Ghosalkar and S. N. Dhage, ”Real Estate Value Prediction Using Linear Regression”, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5.

[4] Vishal Venkat Raman, S. V. (2014). “Identifying Customer Interest in Real Estate Using Data Mining Techniques” (Vol. 5 (3)). Vellore, Tamil Nadu, India: International Journal of Computer Science and Information Technologies

[5] D. Sangani, K. Erickson and M. A. Hasan, “Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting”, 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Orlando, FL, 2017, pp. 530-534.

[6] Azadeh A. et al. “A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations”. Expert Systems with Applications, 2012, 39(1): 298–315.

[7] Cock D. D. Ames, Iowa: “Alternative to the Boston housing data as an end of semester regression project Journal of Statistics Education”. 2011, 19(3): 11-13.

[8] Truong Q., et al. “Housing Price Prediction via Improved Machine Learning Techniques”. Procedia Computer Science, 2020, 174: 433-442.

[9] Zauhar R., et al. “As in Real Estate, Location Matters: Cellular Expression of Complement Varies Between Macular and Peripheral Regions of the Retina and Supporting Tissue”. Front Immunol, 2020, 13: 519.

[10] Arshiya Shaikh, R. Vinayaki, G. Siddhanth, Y. Phanindra Varma - ” House price prediction using multivariate analysis” 2020, IJCRT.

[11] Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, Dr. Vinayk A. Bharadi - “House Price Prediction Using Machine Learning” 2021, IRE Journals.

[12] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei - “Housing Price Prediction via Improved Machine Learning Techniques” 2020

[13] Ms. A. Vidhyavani, O. Bhargav Sathwik, Hemanth.T, Vishnu Vardhan Yadav.M - ”House Price Prediction Using Machine Learning” 2021, Ijcrt.

[14] Thuraiya Mohd, Suraya Masrom, Noraini Johari - “Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia” 2019, IJRTE.

[15] M Thamarai, S P Malarvizhi - “House Price Prediction Modeling Using Machine Learning” 2020, DJIEEB.

[16] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh - ”A hybrid regression technique for house prices prediction” 2017, IEEE.

[17] Sayan Putatunda – “PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market