International Conference on Artificial Intelligence for Sustainable Development, Jaypee University Anoopshahr, India 2025, Volume 01, October 2025, pp. 189-196

HOUSE PRICE PREDICTION USING LINEAR REGRESSION

Harshvardhan Tripathi[1], Adarsh Sahani[2], Kundan Paswan[3]

[1]hntrnt2005@gmail.com, [2]adarshsahani430@gmail.com, [3]kundan70605@gmail.com

Department of Computer Science

Babasaheb Bhimrao Ambedkar University, Satellite Centre, Amethi, Lucknow, INDIA

Corresponding author: hntrnt2005@gmail.com

Abstract. Prediction of prices of houses is very important for buyers, sellers and planners in real estate market. In this paper it is explained how Linear Regression can be used for estimating and predicting the prices of houses using the Ames Housing Dataset. Ames Housing Dataset includes data of 1460 houses and their 81 different features or attributes about the houses. We first cleaned missing data of dataset using simple methods like replacing them with median or most common values. Then we trained the Linear Regression Model and tested it. Our model resulted in 0.82 score for R^2, 22417 for MAE and RMSE of about 36,879. From this we understand that even a simple model like Linear Regression can work well in predicting prices of houses. In the end we also discussed various studies related to our work and also included how we can improve this system in future.

Keywords: House Price Prediction; Linear Regression; Data Mining; Machine Learning; Ames Housing Dataset.

1. INTRODUCTION

Prediction of prices of how an important task in real estate industry is. For buyers it helps in choosing if a house is affordable or not and help in making informed purchase decisions. For sellers it allows them to set fair prices of houses to get profit while keeping the prices competitive. for investors and policymakers' prediction of house prices allow them to do strategic planning, make investment choices and do urban development. Traditionally valuation of properties is done manually by humans who are experts, and they compare various factors like location of house, size of house and other features for predicting or valuing the properties. But this traditional process is often slow time taking expensive and influenced by personal judgments.

Machine learning provides such a modern method to predict the prices of houses automatically. Muscle learning uses old data for learning patterns and then uses those patterns to make faster and more accurate predictions. There are many machine learning algorithms but linear regression is one of the simplest and very easy to understand. Linear regression points the relationship between different factors such as location of the house size of the house number of rooms in the House and the final house price Through a straight-line formula. As linear regression is very easy to understand, it helps us to see how each factor affects the price of house.

This paper applies Linear Regression on the Ames Housing Dataset which is having records of 1460 houses with it also have 81 diverse features or attributes of 1460 houses describing properties like including lot size, build year of house, garage area in house, and overall quality of house. Our main goal is to build a predictive model and evaluate and analyse the performance of model built using 3 metrics which are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2 score).

The focus of this study is to create a model that will predict the prices of houses. We have also shown or illustrated the workflow of the system ranging from data preprocessing, feature selection to building of model, evaluation and interpretation. Through analysis of results such as accuracy of prediction and feature importance we want to show how data can be used for real estate valuations and pricings. The results of this study will become the base of future enhancement in this field. In future more advanced and complex algorithms and large datasets can be used for better results and accuracy.

2. LTERATURE REVIEW

Housing price prediction has been widely studied in the research community. De Cock [1] introduced the Ames Housing Dataset to provide a richer alternative to the Boston Housing dataset, offering more detailed property attributes for analysis. Raschka [2] discussed how Ridge and Lasso regression methods improve prediction accuracy when dealing with correlated features. James et al. [3] showed that ensemble methods such as Random Forest outperform basic regression models for complex datasets. Zhang [4] demonstrated that Gradient Boosting achieves high accuracy for price prediction, albeit with higher computational complexity. Kuhn and Johnson [5] emphasized that careful data cleaning and preprocessing significantly improve model performance.

Rahman and Alam [6] explored missing data imputation techniques for regression, showing that median and mode replacement can improve model stability. Kumar and Gupta [7] demonstrated that Linear Regression with targeted feature selection can yield competitive results. Choudhury et al. [8] applied Principal Component Analysis (PCA) to reduce dimensionality in housing datasets, improving computational efficiency without significant loss in accuracy. Bhardwaj and Sinha [9] examined deep learning approaches for house price prediction, reporting improved accuracy at the cost of model interpretability. Harrison and Rubinfeld [10] studied socio-economic factors in real estate pricing, showing the importance of contextual features in predictive modelling.

These studies collectively highlight the importance of data quality, preprocessing, feature selection, and model choice in housing price prediction. They also motivate the use of Linear Regression as a baseline model, providing a simple yet interpretable framework that can be further enhanced with advanced methods for greater accuracy and robustness.

1. METHODOLOGY

We followed a structured approach for predicting house prices which is shown in our methodology.

A. Dataset

The name of our dataset is Ames Housing Dataset, which has 1460 records of houses and their 81 features which describe the properties of houses such as size of house, year in which house was built, garage area. Our target variable is Sale Price.

B. Data Preprocessing

We preprocessing the data to ensure we have clean and reliable data for prediction:

- Handling Missing Data: We have replaced missing numerical values with median and categorical values with more.
- Feature Selection: Only numerical columns are selected.
- Data Splitting: Data was split into 80% as training data and 20% as data.

C. Model Building

We used scikit-learn' s Linear Regression for model building:

1) Dataset was loaded.
2) Numerical feature and target were selected.
3) Split dataset.
4) Model was trained.
5) Price was predicted.

D. Evaluation Metrics

We evaluated the model:

1) Mean Absolute Error (MAE): Mean absolute error tells us on average by how much our prediction are of from real values. It does not care if our prediction is too high or too low that is it just looks at the size of error. It is an easy way to see how accurate our model is.

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

A lower value of MAE means the prediction of model are closed to actual value MAE is especially useful when we want to treat all error equally without giving extra weight to larger errors.

2) Root Mean Squared Error (RMSE)*:* RMSE tells us about how for our predicted values are from real values, but it gives bigger mistake more value. This means our model result in a very large error RMSE will increase a lot meaning thar big error more serious.

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

A lower value of RMSE value means that the prediction is closer to real value which mean the model is more accurate. RMSE is especially useful in cases where big mistake cause problem, because RMSE stresses more on larger error.

3) R-squared (R^2): R^2 tells us how properly our model is explaining what is happening in data a high R^2 values means that the model is predicting very close to real values and it can explain most of the changes on the other hand a low R^2 values means that the model is not capturing the data very well.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum i = 1^n (y_i - \bar{y})^2} \quad (3)$$

Higher values of R^2.

2. INTERPRETATION

Mae (1): offers a clear and straightforward average error magnitude. It is easy to interpret and useful for understanding the typical size of prediction errors. However, it does not penalize large errors more heavily.

Rmse (2): penalizes larger errors strongly, making it a good choice for situations where large deviations are critical to avoid. It provides a more sensitive measure of prediction accuracy compared to mae.

R^2 (3): indicates how much of the variance in the target variable is explained by the model. A high r^2 value means a good fit, but it should be interpreted along with mae and rmse to fully assess prediction quality.

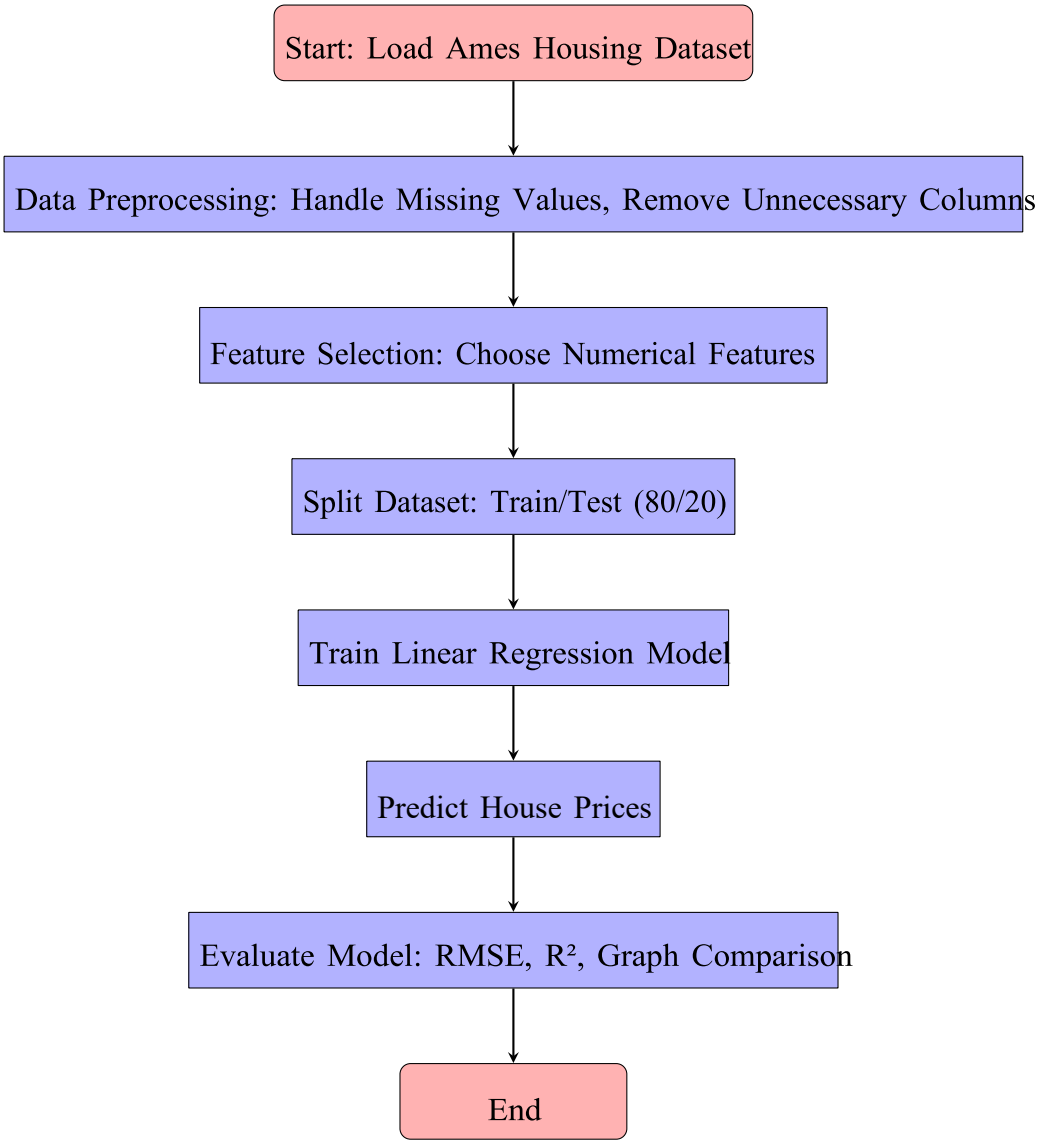

Fig. 1: flowchart of methodology for house price prediction

3. RESULTS

TABLE I: Model Performance Metrics

Metric	Value
Mean Absolute Error (MAE)	22,417.65
Root Mean Squared Error (RMSE)	36,879.82
R^2 Score	0.82

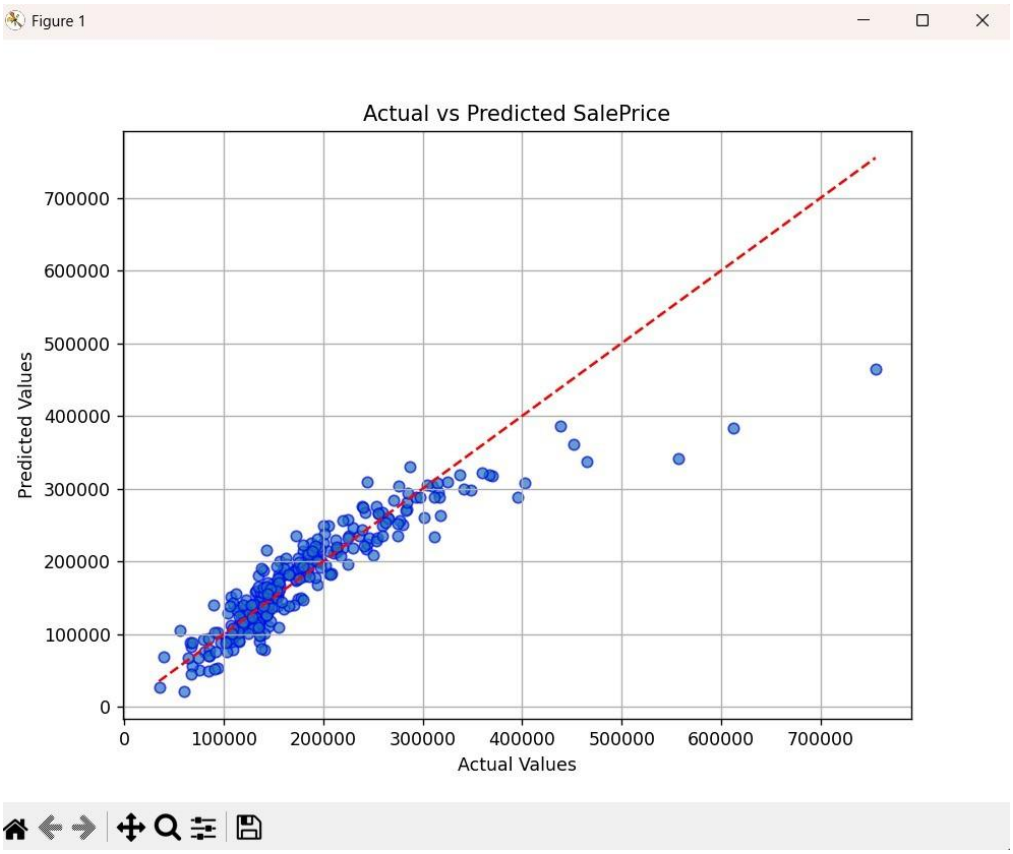

Fig. 2: Actual vs Predicted House Prices. Points close to the dashed line show accurate predictions.

Fig. 3: performance metric comparison: mae, rmse, r^2.

A. Sample predictions

Table 2 shows actual and predicted prices for sample test houses. Predictions align closely with actual values in most cases, but extreme high-value houses show greater deviation.

Table ii: actual vs predicted house prices

Actual price	Predicted price
208,500	213,024.58
181,500	178,965.12
223,500	227,864.27
140,000	145,768.91
250,000	241,512.35

B. Feature importance

Table 3 lists top features and their coefficients. Higher absolute coefficients mean more influence on price prediction. Quality and size features dominate.

Table iii: top features and coefficients

Feature	Coefficient
Overallqual	12,340.45
Grlivarea	45.12
Garagecars	9,875.60
Totalbsmtsf	32.78
Yearbuilt	150.89

4. CONCLUSION AND FUTURE WORK

Linear regression is simple method which helps us to predict house prices on the basis of information like size of house, number of rooms in house or location of house. Although it is an easy technique still the results provided by it are accurate. Still, we can improve its performance by using more advanced models, by including more data for training or by adding extra details or features of house -like nearby facility, build year, or quality of neighbourhood.

The model is doing a very good job in predicting the prices of houses. The r^2 value is 0.82 which means it can explain most of price changes on the basis of given data. The mae value is 22,417, meaning that on average the model's prediction is about 22,000 away from the real price, which is a small error. The rmse value is 36,879, showing that the overall accuracy of model is good but it sometimes struggles with houses that are very different from the usual ones.

Future improvements could include:

- We can use stronger models like random forest or xgboost.
- We can employ deep learning approaches for complex features.
- We can use information about the location of house as part of data used for prediction.
- We can build a system that will give results as soon as we enter the data.
- Using data from various varied sources and cities.

CONFLICT OF INTEREST

The authors declare no conflicts of interest regarding the current research.

REFERENCES

1. D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data," *Journal of Statistics Education*, vol. 19, no. 3, pp. 1–17, 2011.
2. S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed., Packt Publishing, 2020.

3. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed., Springer, 2021.
4. H. Zhang, "A Survey on Machine Learning for Predictive Data Analytics in Real Estate," *IEEE Access*, vol. 7, pp. 123456–123468, 2019.
5. M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, 2013.
6. M. Rahman and M. Alam, "Missing Data Imputation in Predictive Modeling: A Comparative Study," *International Journal of Computer Applications*, vol. 182, no. 45, pp. 10–16, 2021.
7. S. Kumar and R. Gupta, "Feature Selection Techniques for House Price Prediction Using Linear Regression," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 7, pp. 55–61, 2019.
8. A. Choudhury and S. Banerjee, "Dimensionality Reduction Using PCA for House Price Prediction," *Procedia Computer Science*, vol. 167, pp. 1802–1811, 2020.
9. P. Bhardwaj and A. Sinha, "Deep Learning for House Price Prediction," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, pp. 345–356, 2022.
10. D. Harrison and D. Rubinfeld, "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.
11. J. Li, Y. Zheng, and W. Zhang, "A Comparative Study of Machine Learning Algorithms for House Price Prediction," *Journal of Real Estate Finance and Economics*, vol. 62, pp. 213–239, 2021.
12. R. Singh and K. Patel, "A Comprehensive Review on Regression Techniques in Housing Price Prediction," *International Journal of Engineering Research and Technology*, vol. 9, no. 4, pp. 352–360, 2020.
13. L. Wang and Q. Chen, "Improving House Price Prediction with Ensemble Learning Methods," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 5, pp. 1203–1213, 2021.
14. M. Noor, A. Khan, and S. Shah, "Data Cleaning Methods for Enhancing Predictive Accuracy in Real Estate," *Procedia Computer Science*, vol. 175, pp. 420–429, 2020.
15. K. Zhang, X. Li, and H. Zhao, "Integration of GIS Data and Machine Learning for Improved Housing Price Prediction," *Applied Geography*, vol. 135, pp. 102–116, 2021.