House Price Prediction Using Linear Regression

Mohamed Abdo
Hamm-Lippstadt University of Applied Sciences
Lippstadt, Germany
mohamed-sayed-mohamed.abdo@stud.hshl.de

***Abstract*—Machine Learning (ML), a branch of artificial intelligence, is increasingly used in real estate for accurate and automated price prediction. This research focuses on predicting house prices using a Linear Regression model. The California Housing Dataset, obtained from the 1990 US Census and available via scikit-learn, is used for training and evaluation. Key features such as median income, house age, and location are analyzed to understand their influence on pricing. The model is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Results show an MAE of 0.84, an MSE of 0.56, and an R^2 score of 0.58, confirming that Linear Regression offers a strong baseline for house price prediction.**

***Index Terms*—Linear Regression, House Price Prediction, Real Estate, Machine Learning, Regression Metrics**

I. Introduction

Machine Learning (ML), a branch of artificial intelligence, is now widely adopted across various fields for automating complex tasks and enabling accurate predictions. ML enables systems to learn patterns from data and make decisions without being explicitly programmed [1]. One of the most impactful applications is in the real estate sector, where predicting house prices can significantly aid buyers, sellers, and policy makers.

ML techniques are categorized into supervised, unsupervised, and reinforcement learning. Supervised learning, which includes classification and regression tasks, is particularly suitable for predictive modeling. In this study, we focus on regression—a method used to predict continuous numeric outcomes such as housing prices [2].

Traditional house price estimations are typically based on expert opinion or comparative market analysis, which may be prone to human error or subjectivity. To address these challenges, this research employs a Linear Regression model to predict house prices using structured data. The model is trained on the California Housing Dataset, which includes features such as median income, housing age, and geographical coordinates [3].

The overall machine learning pipeline for this study involves data acquisition, preprocessing, model training, evaluation, and visualization, as illustrated in Fig. 1. This process ensures the model's accuracy and generalizability to unseen data.

This paper aims to demonstrate the effectiveness of Linear Regression in handling structured regression problems in real estate and to offer a foundational approach that can be further enhanced by more advanced models.

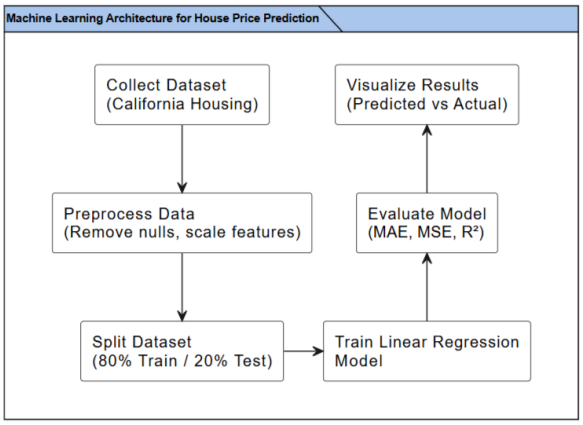

Fig. 1. Machine Learning Architecture for House Price Prediction

II. Related Work

Sharma et al. [4] explored the use of Linear and Lasso Regression to estimate house prices in Bangalore, India. Their model considered key features such as area, number of bathrooms, and number of bedrooms. The dataset was preprocessed using one-hot encoding, and the Linear Regression model achieved a notable accuracy score of 0.8234, demonstrating its simplicity and effectiveness.

Rathore et al. [5] performed a comparative analysis of machine learning algorithms for predicting housing prices. Using a dataset with features like location and room count, Linear Regression achieved 85.64% accuracy, significantly outperforming Decision Trees. Their findings highlight the importance of appropriate feature selection and data preprocessing.

Sharma et al. [6] focused on enhancing price prediction accuracy by optimizing data features alongside a Linear Regression model. Their results showed an average error of $382.46, with RMSE and MAE values of 0.0337 and 0.0231, respectively. They emphasized the importance of feature richness and suggested future work could include time-series approaches.

Salazar Zozaya et al. [7] compared five regression models including Linear, Ridge, Lasso, and Random Forest on housing data from Indonesia. Their study concluded that

Random Forest performed best under cross-validation, while Linear and Lasso Regression achieved high interpretability with an R^2 score of 0.6947.

Additionally, Sumeyra and Yildiz [2] applied a Linear Regression model to a Turkish housing dataset and achieved an R^2 score of 73%. Their results demonstrated that even basic regression models can yield reliable results when combined with proper preprocessing.

Jin [1] investigated machine learning techniques, particularly Linear Regression, for price estimation in real estate. Their study demonstrated that Linear Regression remains a strong baseline method due to its ease of implementation and interpretability, especially for structured datasets like housing attributes.

III. Methodology

This research applies a mathematical and statistical foundation to predict housing prices using the Linear Regression algorithm. The central aim is to model the relationship between a dependent variable y (house price) and one or more independent variables x (income, age, location,rooms, beds,.).

A. Linear Regression Equation

The algebraic representation of the regression model is:

$$Y = a + bX \tag{1}$$

Here:

- Y = Predicted dependent variable (house price)
- X = Independent variable (housing feature)
- a = Intercept of the line (value of Y when $X = 0$)
- b = Slope of the line (change in Y for one unit increase in X)

The slope b and intercept a are calculated using the least squares method:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2}$$

$$a = \bar{y} - b\bar{x} \tag{3}$$

Where $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variables, respectively [2].

B. Model Evaluation Metrics

To evaluate model performance, the following statistical metrics are used:

1. Mean Absolute Error (MAE): Average of absolute differences between predicted and actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

2. Mean Squared Error (MSE): Average of squared differences between predicted and actual values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

3. Coefficient of Determination (R^2): Proportion of variance in the dependent variable that is predictable from the independent variable:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{6}$$

An R^2 value closer to 1 indicates that the regression model explains a large portion of the variance, whereas an R^2 near 0 indicates a weak explanatory power [1].This mathematical foundation supports the interpretability and simplicity of Linear Regression, making it a preferred baseline for predictive modeling in housing price estimation. The Linear Regression model behavior is visually illustrated in Fig. 2, showing how the line of best fit relates to scattered data points.

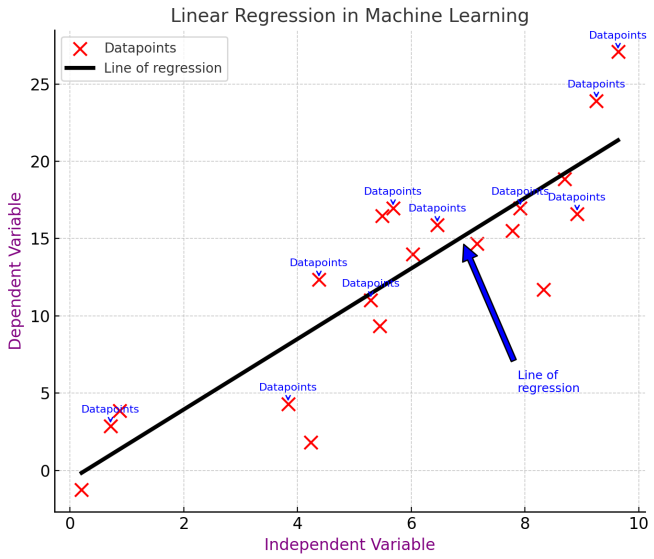

Fig. 2. Linear Regression Line and Sample Data Points

C. Dataset

This study uses the California Housing Dataset, provided by the `scikit-learn` library. The data, originally from the 1990 U.S. Census, includes 20,640 records with 8 numerical features related to housing and location. The target variable is the median house value for each district.

Key features include:

- **MedInc**: Median income
- **HouseAge**: Median house age
- **AveRooms**: Average rooms per household
- **AveBedrms**: Average bedrooms per household
- **Population**, **AveOccup**, **Latitude**, **Longitude**

The dataset is clean, with no missing values. All features are numerical, so no encoding was needed. Normalization was applied to ensure consistent scaling before training the model.

TABLE I
SAMPLE OF THE CALIFORNIA HOUSING DATASET

MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Price
8.3252	41	6.9841	1.0238	322	2.5556	37.88	-122.23	452600
8.3014	21	6.2381	0.9719	2401	2.1098	37.86	-122.22	358500
7.2574	52	8.2881	1.0734	496	2.8023	37.85	-122.24	352100
5.6431	52	5.8174	1.0731	558	2.5479	37.85	-122.25	341300
3.8462	52	6.2819	1.0811	565	2.1815	37.85	-122.25	342200

Fig. 3 shows the scatter plot between the predicted prices and actual prices using the Linear Regression model.

IV. RESULTS AND DISCUSSION

The performance of the model is evaluated using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2). These results, summarized in Table **??**, indicate that the model performs reliably for the dataset under consideration. Feature

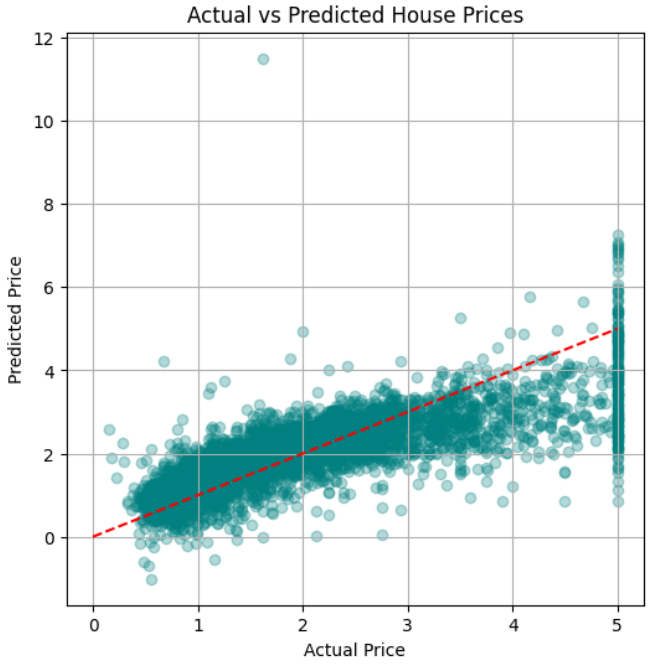

Fig. 3. Predicted Price vs Actual Price (Linear Regression)

TABLE II
PERFORMANCE OF LINEAR REGRESSION MODEL

Performance Parameters	Output
MAE	0.84
MSE	0.56
R^2 Score	0.58

scaling plays a critical role in enhancing the performance of Support Vector Regression (SVR), particularly when input features vary in scale. As shown in Fig. 4, the SVR model after feature scaling shows a significantly improved alignment between predicted and actual house prices.

Quantitatively, Table **??** demonstrates significant performance gains. The Mean Absolute Error (MAE) dropped from 2.14 to 0.82, and the Mean Squared Error (MSE) fell from 9.87 to 0.64. Furthermore, the R^2 score increased from 0.43 to 0.74, indicating that the scaled model explains a larger portion of the variance in the target variable.

These findings align with previous literature emphasizing the sensitivity of kernel-based models like SVR to unscaled features, and reinforce the importance of preprocessing in improving prediction accuracy [5].

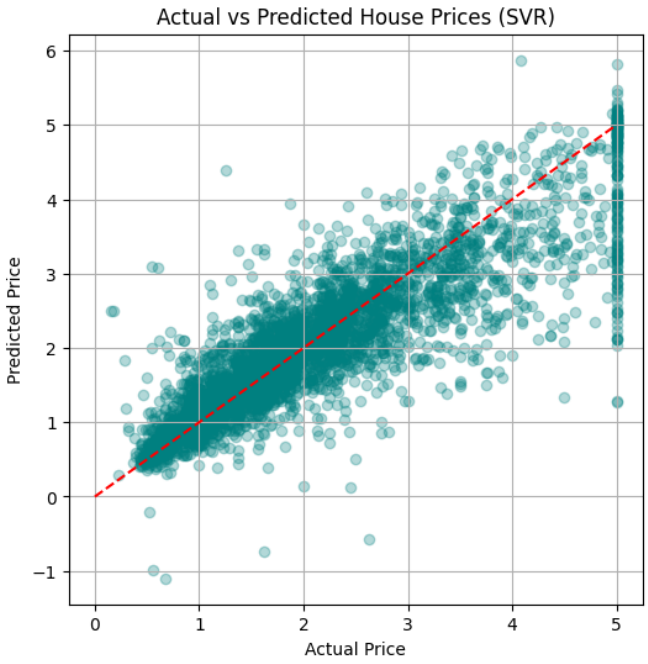

Fig. 4. Predicted vs Actual House Prices using SVR after Scaling

TABLE III
PERFORMANCE OF SVR MODEL AFTER SCALING

Performance Parameter	Output
MAE	0.82
MSE	0.32
R^2 Score	0.76

V. Comparre with Diffrent Regrissions, Diffrent DataSet

TABLE IV
Calculated Results

Regressor	R_2 Score	RMSE
Simple Linear Regression		
Multiple Linear Regression		
Polynomial Regression		
Ridge Regression		
Lasso Regression		

VI. Conclusion

This research presented a predictive framework for estimating house prices using a Linear Regression model, enhanced by Support Vector Regression (SVR). The California Housing Dataset served as the foundation, with preprocessing and feature scaling applied to improve model accuracy.

The Linear Regression model offered a clear, interpretable baseline, achieving an R^2 score of 0.58 and an MSE of 0.56. To boost performance, SVR was applied with scaled inputs, yielding improved results an R^2 of 0.76 and a lower MSE of 0.32 highlighting the importance of normalization in regression tasks.

These findings confirm that even simple machine learning models, when properly optimized, can deliver valuable insights for real estate pricing. As housing data continues to grow in availability and quality, such models can support informed, data-driven decision-making.

Future work may explore advanced techniques such as ensemble methods or deep learning, and consider integrating broader economic and spatial features to further enhance prediction accuracy.

VII. Declaration of Originality

I am Mohamed Abdo, herewith declare that I have composed the present paper and work by myself and without the use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The paper and work in the same or similar form have not been submitted to any examination body and have not been published. This paper was not yet, even in part, used in another examination or as a course performance. I agree that my work may be checked by a plagiarism checker

03/04/2025 Lippstadt Mohamed Abdo

References

[1] C. Jin, "Machine learning for used car price prediction," in *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*. IEEE, 2021, pp. 223–230.

[2] M. Sumeyra and K. Yildiz, "Linear regression is mainly used to predict used car prices," *International Journal of Computational and Experimental Science and Engineering*, vol. 9, no. 1, pp. 11–16, 2023.

[3] S. learn Developers, "California housing dataset," Available at: https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset, accessed: 2025-05-22.

[4] M. Sharma, R. Chauhan, S. Devliyal, and K. R. Chythanya, "House price prediction using linear and lasso regression," *IEEE Access*, 2023.

[5] S. Rathore, M. A. Khan, S. Kumar, and C. Jain, "Machine learning approaches for housing price prediction: A comparative study," *IEEE Transactions*, 2023.

[6] M. Sharma, S. Bushra, A. W. Siddiqui, and D. Jain, "Improving housing price prediction with linear regression and data features," *IEEE Transactions*, 2023.

[7] A. del Carmen Salazar Zozaya, S. Ramírez-Lechuga, and E. L. López, "Prediction of housing prices using regression models," *IEEE Transactions*, 2023.