Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Predicting House Prices Using Machine Learning Models

Tianying Zhao

Department of Computer Science, XYZ University, Dongguan, China

zty19574988692@outlook.com

Abstract. Predicting house prices is crucial for stakeholders such as buyers, developers, financial institutions, and policymakers. This paper explores how machine learning models can help with house price prediction. Using the Ames Housing Dataset, we compare three models: linear regression, decision tree, and random forest. The goal is to understand which model gives better results in terms of accuracy and how well it works with new data. Data preprocessing, including imputation for missing values, standardization, and feature selection, was performed to ensure compatibility with the models. The models were tested using three common evaluation tools: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 score. The results show that the random forest model performs the best. It has the lowest error and the highest R^2 score among the three. This is because random forest uses many decision trees and combines their results, which helps reduce overfitting and improves prediction. The study highlights the potential of machine learning in improving house price prediction accuracy and discusses future avenues for model optimization. These include incorporating real-time data, enhancing model adaptability, and tailoring predictions for specific stakeholder needs. The findings suggest that machine learning models, particularly random forest, can provide valuable insights into housing market trends and support more informed decision-making.

Keywords: House Price Prediction; Machine Learning; Linear Regression; Decision Tree; Random Forest; MSE; R^2 Score.

1. Introduction

House price prediction plays a vital role for various stakeholders. For homebuyers, accurate forecasts enable better budgeting and timing of purchases. Developers rely on price predictions to formulate development plans and pricing strategies that ensure profitability and competitiveness. Financial institutions assess mortgage collateral value and lending risks based on predicted prices to safeguard capital. Government agencies, on the other hand, utilize price forecasts to devise informed policies that regulate market supply and demand, thus promoting the stable development of the real estate sector.

Classical approaches, linear regression and decision tree models, for example, have long been employed due to their simplicity and interpretability [1]. According to Montgomery et al. [2] in *Introduction to Linear Regression Analysis*, linear regression predicts house prices by modeling the linear relationship between price and basic attributes such as floor area, geographic location, and nearby amenities. However, the relationship between house prices and influencing factors in practice is often highly complex, exhibiting non-linear patterns and interactions. This complexity limits the predictive accuracy of linear regression. Decision trees, while capable of handling non-linearity, are kind of overfitting. They may perform well on training data but show poor generalization to unseen data.

Recently, the application of next-generation information technologies to large datasets has proven effective in overcoming prediction challenges. Chen et al. [3] believe that machine learning techniques can extract valuable insights from vast datasets, thereby enhancing prediction accuracy—a view echoed by Nguyen et al. [4]. Zhou et al. [5] highlight the strength of deep learning methods in improving a model's ability to capture data features. The Random Forest algorithm, which integrates machine learning and decision tree methodologies, represents a significant advancement. Breiman [6]

points out that Random Forest reduces the correlation among individual trees by introducing randomness (e.g., random feature and sample selection), thereby effectively mitigating the risk of overfitting.

This study draws on the methodology of Kumar and Toshniwal [7], applying linear regression, decision trees, and random forest models to housing price prediction. Using the Ames Housing Dataset, the study compares different models for predicting house prices. The comparison is based on three main evaluation tools: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). By looking at how well the models predict prices and how well they work with new data, the research aims to understand the strengths of both traditional and machine learning methods in house price prediction.

Moreover, the dynamic nature of the housing market—impacted by interest rates, demographic shifts, and policy changes—demands models that are not only accurate but also adaptable to evolving conditions.

The structure of this paper is as follows: Section 2 describes the dataset and the steps taken to prepare it; Section 3 introduces the prediction models used in this study; Section 4 shows the results and explains how the models were tested; and Section 5 ends with a discussion of the findings and suggestions for future research.

2. Data Description

2.1. Dataset

The Ames Housing Dataset employed in this study comprises 2,930 samples and 79 housing-related features, offering a rich source of information for training and validating prediction models. These features include property location, size, number of bedrooms and bathrooms, and overall quality, among others. The sale price serves as the target variable, representing the market value of the property.

As noted by Li et al. [8], proper feature selection significantly influences model performances. In this study, 1,000 rows of data are selected from the Ames Housing Dataset, focusing on the most relevant features (as presented in Table 1):

Table 1. Selected Features

Numerical data:	
Size	Ranges from 800 to 4,500 square feet
Number of Bedrooms	Ranges from 1 to 6
Number of Bathrooms	Ranges from 1 to 5
Age of the House	Ranges from 0 to 100+ years
Lot Size	Ranges from 1,000 to 10,000 square feet
Year Built	Range from 1900 to 2020
Categorical data:	
Location	Includes 5 neighborhoods (e.g., Downtown, Suburb, Rural, etc.)
House Type	Categories like detached, semi-detached, townhouse, apartment
Condition	Categories include new, renovated, and old

2.2. Preprocessing

Raw data often has problems such as missing values that are not suitable for the prediction model. Therefore, three procedures for data preprocessing were taken in this study before the model application.

Firstly, to address missing values in the dataset, imputation techniques were applied. For categorical features (e.g., location, house type), One-Hot Encoding was used to convert them to numerical form,

with the aim of ensuring compatibility with machine learning models. For numerical features (e.g., house size, bedroom count), standardization was employed to scale the data, transforming the data to a standard normal distribution with a mean of 0 and a standard deviation of 1.

Additionally, outlier detection was conducted using interquartile range (IQR) analysis to remove extreme values that could skew the training process. This ensures cleaner input data, which improves model stability and predictive performance.

3. Methodology

3.1. Linear Regression

The linear regression model predicts house prices using a linear equation that represents the relationship between the dependent variable (price) and multiple independent variables (features) as a weighted sum. The general form of the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots\ldots + \beta_n x_n + \epsilon$$

In this study, linear regression is used to capture linear associations between features and house prices, with model parameters estimated via the least squares method.

3.2. Decision Tree

A decision tree is a non-linear model based on a hierarchical tree structure. It recursively partitions the dataset into subsets based on feature values that best minimize prediction error. Each internal node represents a feature, and each branch is a decision rule, culminating in a leaf node with a predicted value.

To prevent overfitting, pruning techniques were applied. These include rule-based methods (e.g., limiting maximum tree depth or minimum sample size per leaf) and cost-complexity pruning, which balances model complexity and prediction accuracy.

3.3. Random Forest

Random forest is an ensemble learning algorithm that builds upon the decision tree model by introducing bagging (bootstrap aggregation) and feature randomness to improve prediction accuracy and control overfitting. In this approach, multiple decision trees are constructed using different random subsets of the training data and a random selection of features at each split. The final prediction is made by aggregating the outputs of all trees, typically through averaging in regression tasks.

In this study, the random forest model is implemented with 100 decision trees and was trained on 80% of the data and tested on the remaining 20%. To ensure robustness and mitigate overfitting, a 5-fold cross-validation strategy during the training process was applied. This enables the model to be evaluated on multiple data splits and supports the tuning of key hyperparameters such as the number of trees, maximum tree depth, and minimum samples per leaf. Cross-validation helps optimize model performance and ensures generalization to unseen data.

4. Model Evaluation

The performance of the three models, linear regression, decision tree, and Random Forest, was evaluated using three metrics:

(1) Mean Squared Error (MSE), quantifies the average squared difference between predicted and actual values; lower values indicate higher accuracy.

(2) Root Mean Squared Error (RMSE), the square root of MSE, offers an error estimate in the same unit as the data, making it more interpretable.

(3) Coefficient of Determination (R^2), which reflects the proportion of variance in the dependent variable that is predictable from the independent variables; values closer to 1 indicate better model fit.

These results (as shown in Table 2) show that the Random Forest model outperforms the other two across all evaluation metrics.

The superior performance of the Random Forest model can be attributed to its ensemble nature, which mitigates overfitting and enhances generalization. Moreover, its ability to handle both numerical and categorical features makes it particularly suitable for heterogeneous datasets like the Ames Housing data.

Table 2. Model Performance

Model	MSE	RMSE	R^2
Linear Regression	25,600,000	5,060	0.65
Decision Tree	15,200,000	3,898	0.78
Random Forest	12,500,000	3,536	0.85

5. Conclusion

This study explains that the application of machine learning algorithms significantly improves the accuracy of house price prediction models. Looking ahead, two key trends are anticipated to further optimize predictive modeling in this field.

First, the continued development of robust toolkits across programming languages will play a crucial role. Kumari Sandhya et al. [9] have already utilized mature predictive models available in Python libraries. Second, tailoring models to meet the specific needs of different stakeholder groups will become increasingly important. Singh et al. [10], for example, constructed a model based on homebuyers' preferences for urban location, regional resources, and financial conditions.

Besides, integrating real-time data sources such as online listings and user behavior from real estate platforms could further enhance model responsiveness. These data streams may allow for near-instantaneous updates, paving the way for real-time prediction systems.

There are still some limits in this study. For example, the data is from only one city, so the results might not work well in other places. Also, we only used simple data like number of rooms or size, but not things like location or market trends. In the future, we can add more types of data to make the predictions better.

While highlighting the importance of both technical tool advancement and user-centered model customization, the trends offer clear directions for future research and practical application.

References

[1] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[2] Montgomery D. C., Peck E. A., Vining G. G. Introduction to Linear Regression Analysis. Wiley, 2012.

[3] Chen K., Li H., Wu Y. Predicting housing prices using machine learning: A comparison of tree-based models. Journal of Real Estate Research, 2022, 44 (2): 233-256.

[4] Nguyen H., Pham H., Do T. Comparative analysis of machine learning models for house price prediction. International Journal of Forecasting, 2020, 36 (3): 1010-1025.

[5] Zhou Y., Zhang H., Liu J. An empirical study on housing price prediction using deep learning. Neural Computing and Applications, 2018, 30 (8): 2335-2347.

[6] Breiman L. Random forests. Machine Learning, 2010, 45 (1): 5-32.

[7] Kumar A., Toshniwal D. A novel framework for real estate price prediction using ensemble learning. Expert Systems with Applications, 2020, 159: 113623.

[8] Li L., Wang X., Zhao D. Real estate price prediction based on feature selection and machine learning. IEEE Access, 2019, 7: 132066-132080.

[9] Kumari Sandhya, Sarwar Siddiqui. House Price Prediction Using Machine Learning. Journal For Research in Applied Science and Engineering Technology, 2022, 11 (6): 1970-1873.

[10] A. P. Singh, K. Rastogi and S. Rajpoot. House Price Prediction Using Machine Learning. 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, 203-206.