House Price Prediction Based on Machine Learning Models

Xiaoyan Ouyang *

Shanghai Normal University, Shanghai, China

* Corresponding Author Email:1000502233@smail.shnu.edu.cn

Abstract. This study aims to provide accurate house price predictions using machine learning algorithms. These predictions can assist decision-makers in making informed property investments and planning. Multiple linear regression and random forest were employed to achieve this goal. First, the acquired data underwent thorough analysis, including preprocessing and visualization. Subsequently, the study employed multiple linear regression and random forest models for house price prediction and evaluated their performance. The multiple linear regression model yielded promising results with an R^2 score of 0.73, explaining 73% of the target variable's variance. However, it exhibited prediction errors in specific cases, suggesting potential areas for improvement. In contrast, the Random Forest model achieved a slightly lower R^2 score of 0.69. Nonetheless, it excelled at capturing complex nonlinear relationships. Additionally, it identified the top five key features influencing house prices: house size, number of bathrooms, number of floors, parking spaces, and air conditioning. This study highlights the potential of machine learning models for house price prediction. Future research can further enhance these models and consider other influential factors to explain house price fluctuations comprehensively. The results offer valuable applications for investors, brokers, and government planners in the real estate market.

Keywords: Multiple linear regression, random forest, price prediction.

1. Introduction

Predicting house prices is crucial in real estate market research and investment decision-making. Housing prices are a key economic indicator that affects the financial status of individuals, families, and businesses and has implications for various stakeholders, including tenants, homeowners, real estate analysts, policymakers, and urban planners.

Algorithm-based predictive models play a significant role in assisting stakeholders in making informed decisions within the real estate sector and the broader economy. Regression, inference, neural networks, and deep learning are among the most popular machine learning algorithms, enabling computers to predict outcomes efficiently through data-driven learning [1].

In general, physical features of houses, such as size, construction year, the number of bedrooms, bathrooms, air conditioning, and other factors defining internal characteristics, can impact housing prices [2, 3]. Numerous researchers have employed various algorithms and models for housing price prediction [4].

Adetunji et al. proposed a novel approach, treating house price prediction as a classification problem and utilizing Random Forest machine learning techniques. Although different from traditional regression models, this approach proves practical in predicting price fluctuations under specific circumstances [1]. Gupta and Zhang used Spearman correlation analysis to identify significant factors affecting house prices. They applied a comprehensive analytical algorithm, constructed a multiple linear regression model for price prediction, and tested it on a real estate price dataset in Boston [5]. Liu, Du, and Wen introduced AG-LSTM, a spatiotemporal attention graph convolutional extended short-term memory model that considers location relationships between different communities and the impact of multiple related attributes on house prices. This model demonstrated impressive performance on Lianjia's Beijing housing price data [6]. Liao conducted a modeling analysis of Beijing's housing prices using ARIMA and BP neural network models. ARIMA model showed good fitting performance on the training set but exhibited significant prediction errors on the test set.

In contrast, the BP neural network model, especially after optimization with dynamic weight particle swarm optimization (PSO), performed better, leading to lower prediction errors. Additionally, introducing ensemble models further enhanced prediction accuracy [7]. To build prediction models, Cai employed various regression algorithms, including Random Forest, XGBoost, and multiple linear regression. They utilized grid search and other techniques to fine-tune hyperparameters and achieve robust predictive performance [8]. Ling and Li (2021) utilized a fusion model based on ElasticNet, LightGBM, Support Vector Regression (SVR), and Gradient Boosting Decision Tree (GBDT) as base models. They conducted feature selection and ranking using Random Forest. Inspired by these studies, this ensemble model aims to explore the effectiveness of simple machine learning algorithms in price prediction. We delve deeper into the data through a series of analytical steps, aiming to understand better the correlations between house prices and various factors [9].

This study compares the performance of two models, simplifying existing methods for house price prediction. Through this research, we aspire to offer a concise and reliable solution for price prediction and provide valuable insights for relevant decision-making processes.

2. Methodology

2.1. Data Collection and Explore

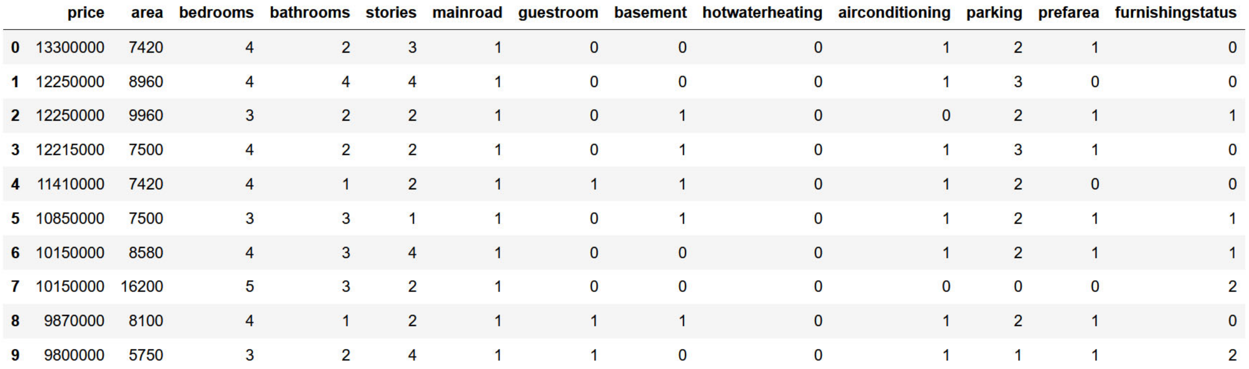
This paper aims to provide a comprehensive and in-depth descriptive analysis of a dataset containing multidimensional attributes of house prices. The real estate market has been in the spotlight for a long time, and this dataset from Kaggle, curated from multiple resources and web scraping, provides multidimensional information about the price of a home, ranging from the basic information about the home to its location and accessibility, parking and geography, to the state of its furnishings. The features present in the dataset are price, which is the price of the house, area which is the total Area of the house in square feet; Bedroom, which is the number of bedrooms; Bathrooms, which is the number of bathrooms; Stories, which is the number of stories, Mainroad which is that whether the house is connected to the main road, Guestroom which is whether the house has a guest room, Basement which is whether the house has a basement, Hot water heating which is whether the house has a hot water heating system, Airconditioning which is whether the house has an air conditioning system, Parking which is the number of parking spaces available, Prefarea which is whether the house is located in a preferred area, Furnishing status which is the furnishing status of the house. The dataset contains a total of 546 training and testing samples. For the training dataset, 80% is used to train the model, and the remaining 20% is used for testing.

2.2. Data Pre-Processing

Prior to model building, data validation and analysis must be performed to ensure the quality of the data set and to understand the data required for the task. The data exploration phase is a critical step in understanding the data. In this phase, this paper takes a number of steps to gain insight into the characteristics of the data.

First, the paper performed a data quality validation to ensure that there are no missing values in the dataset. This is very important because missing values may negatively affect the accuracy and stability of the model.

For the sub-typed values, a case-by-case coding approach was adopted by using the LabelEncoder from the sklearn library to code the dataset guestroom, mainroad, basement, hotwaterheating, airconditioning, furnishingstatus, and prefarea columns by encoding these categorical variables as integers, converting them to a numerical form understandable by the model, and the unique values of the encoded variables were verified by calling the .unique() function to ensure that the encoding did not introduce errors or exceptions. The processed data is shown in Fig.1.



	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	1	0	0	0	1	2	1	0
1	12250000	8960	4	4	4	1	0	0	0	1	3	0	0
2	12250000	9960	3	2	2	1	0	1	0	0	2	1	1
3	12215000	7500	4	2	2	1	0	1	0	1	3	1	0
4	11410000	7420	4	1	2	1	1	1	0	1	2	0	0
5	10850000	7500	3	3	1	1	0	1	0	1	2	1	1
6	10150000	8580	4	3	4	1	0	0	0	1	2	1	1
7	10150000	16200	5	3	2	1	0	0	0	0	0	0	2
8	9870000	8100	4	1	2	1	1	1	0	1	2	1	0
9	9800000	5750	3	2	4	1	1	0	0	1	1	1	2

Figure 1. The processed data

And for numerical values, normalization was performed. In this task, the standard scaler provided in the Python Scikit-learn module was used. For the mean and standard deviation of each feature, the mean and standard deviation of that feature in the entire dataset are calculated and the scaled data is obtained by normalizing the original data x through the transform function. The scaling process is based on the following equation:

$$\frac{x_i - x_{mean}}{x_{stdev}}, \tag{1}$$

Here, x_i is the raw data, x_{mean} is the average of the data, and x_{stdev} represents the standard deviation of the data. The normalized data will have a mean of 0 and a standard deviation of 1, which helps to ensure that different features have similar scales to improve model performance and stability.

2.3. Linear Regression

Multiple linear regression is a regression analysis method commonly used in the field of statistics and machine learning to establish and analyze the linear relationship between the dependent variable and multiple independent variables. The core idea of the method is to establish a link between a linear combination of independent variables and the dependent variable by finding the line of best fit (or hyperplane), which has good interpretability, and is widely used in various fields, including economics, finance, social sciences, natural sciences, etc., to establish and analyze the relationship between the effects of multiple factors on a phenomenon or variable [10].

Assuming that there is a linear relationship between house prices and each attribute, and there is no particularly high correlation between the attributes to be studied, the following mathematical model is established:

$$Price = \beta 0 + \beta 1 * Area + \beta 2 * Bedrooms + \beta 3 * Bathrooms + \cdots + \beta n * FurnishingStatus + \epsilon, \tag{2}$$

Where Price is the house price as the dependent variable, Area, Bedrooms, Bathrooms, etc. are the attributes used as independent variables, and β is the regression coefficient. The goal of the model is to estimate the optimal regression coefficients using the available dataset to minimize the sum of squares of the residuals between the actual observations and the model predictions.

Linear regression typically uses the least squares method to estimate the regression coefficients with the goal of minimizing the sum of squared residuals between the actual observations and the model predictions to find the line of best fit. The method seeks a set of regression coefficients that minimize the sum of squares of the differences between the actual observations and the model predictions. This can be accomplished by solving the least squares regular equation.

2.4. Random Forest

Random Forest is a powerful and widely used integrated learning algorithm for machine learning, often used to solve regression and classification problems. The main idea of the method is to make predictions by constructing multiple decision trees, each based on a random selection of different subsets and attributes of the data, and combining the results of these trees into a complex nonlinear model, which has the advantages of high accuracy, resistance to overfitting, handling large-scale data, and feature importance assessment. The following is a summary of the principles of the Random Forest algorithm:

i. Decision Tree Integration: random sampling with put-back is performed from the training dataset to generate several different training subsets, each of which contains a portion of the original data. These subsets are used to train each decision tree. Each decision tree is an independent classifier or regressor for making predictions on the data. The results of these trees are voted or averaged to form the final prediction.

ii. Randomness: an element of randomness is introduced to ensure that each decision tree is unique. Specifically, it employs the following randomness elements:

(a). Self-sampling: each tree is constructed based on Bootstrap Sampling's training data, and the training data for each tree is randomly sampled from the original data in a relaxed manner, which allows each tree to have a different view of the data.

(b). Randomized Feature Selection: In the process of constructing each tree, a subset of features is randomly chosen from the entire pool of available features instead of utilizing all of them.

iii. Integration Advantage: The overfitting problem of a single decision tree is overcome by combining the results of multiple decision trees. The integrated results are usually more stable and have better generalization ability.

In the Random Forest Classifier, 500 decision trees are chosen in this paper to construct the random forest. Increasing the number of decision trees usually improves the performance of the model and also increases the computational cost.500 decision trees are considered a reasonable choice to provide accurate predictions while maintaining computational efficiency.

2.5. Evaluation metrics

In statistical analysis, there are many ways to calculate the prediction error. To evaluate the effectiveness of the house price prediction model in this research, two key performance indicators are used in this paper: mean square error and coefficient of determination. These metrics help quantify the predictive accuracy and explanatory power of the model.

I. The mean square error represents the squared expected value of the error and is used as a measure of the difference between the model's predicted value and the actual observed value. It is calculated using the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{3}$$

Where n denotes the number of samples, the actual observed value of the ith sample, the predicted value of the model for the ith sample, and the smaller the MSE, the smaller the model.

II. The coefficient of determination measures the extent to which this model explains the variance of the target variable. Its calculation formula is as follows:

$$R^2 = 1 - \frac{SSR}{SST}. \tag{4}$$

Here, SSR signifies the summation of squared deviations between the model's predicted values and the mean of the target variable, whereas SST represents the summation of squared deviations between the observed actual values and the mean of the target variable. The $R^\wedge 2$ Score, or the coefficient of determination, has values that range between 0 and 1. A value closer to 1 indicates that the model has a stronger ability to explain the variance in the target variable, resulting in a better fit.

3. Result and Discussion

3.1. Result of the Data Exploration Process

In the data exploration phase, this study created Kernel Density Estimate (KDE) plots, using ‘Price’ as an example. As shown in Fig. 1, the horizontal axis represents the ‘Price’, while the vertical axis represents the probability density, which indicates the probability of each price value occurring. This graph illustrates the distribution of price data, and it can be observed that prices are predominantly concentrated in the range close to 4 million dollars. This concentrated distribution trend reflects a common pattern in housing prices.

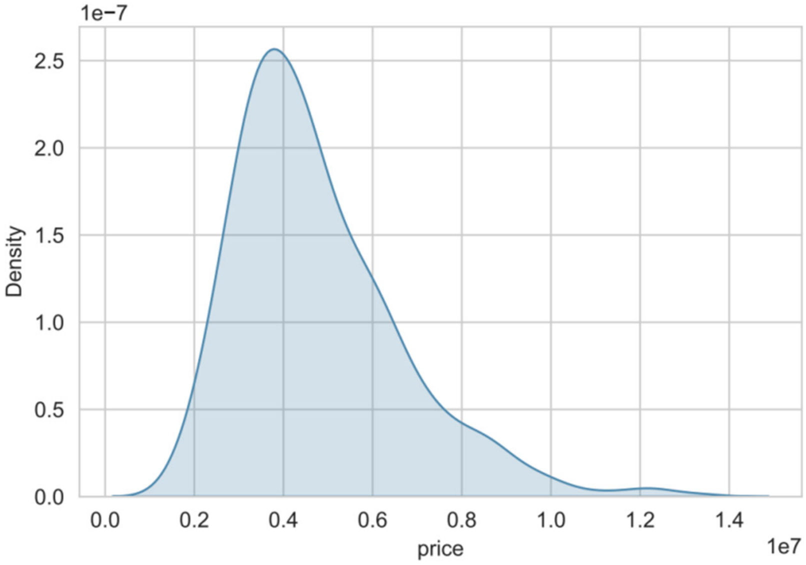


Figure 2. Kernel density estimation plot (Price)

Based on custom-defined categorical visualizations of the relationships and distributions among different categories, as observed from Fig. 3, it can be noted that Bathrooms have an impact on Price. When the number of bathrooms is 1, house prices are notably lower, primarily clustered around approximately 0.4 million dollars, displaying a distinct concentration trend. This may indicate that single-bathroom houses are relatively more affordable in terms of price. However, when the number of bathrooms increases to 2, there is a significant uptick in the distribution of house prices, with an average exceeding the price level of houses with one bathroom. Homes with two bathrooms tend to be more expensive. When the number of bathrooms reaches 3, the distribution of house prices becomes more scattered, with no apparent concentration trend. This situation may suggest that houses with three bathrooms exhibit a wider range of price variations, influenced by multiple factors, without a clear consistency trend.

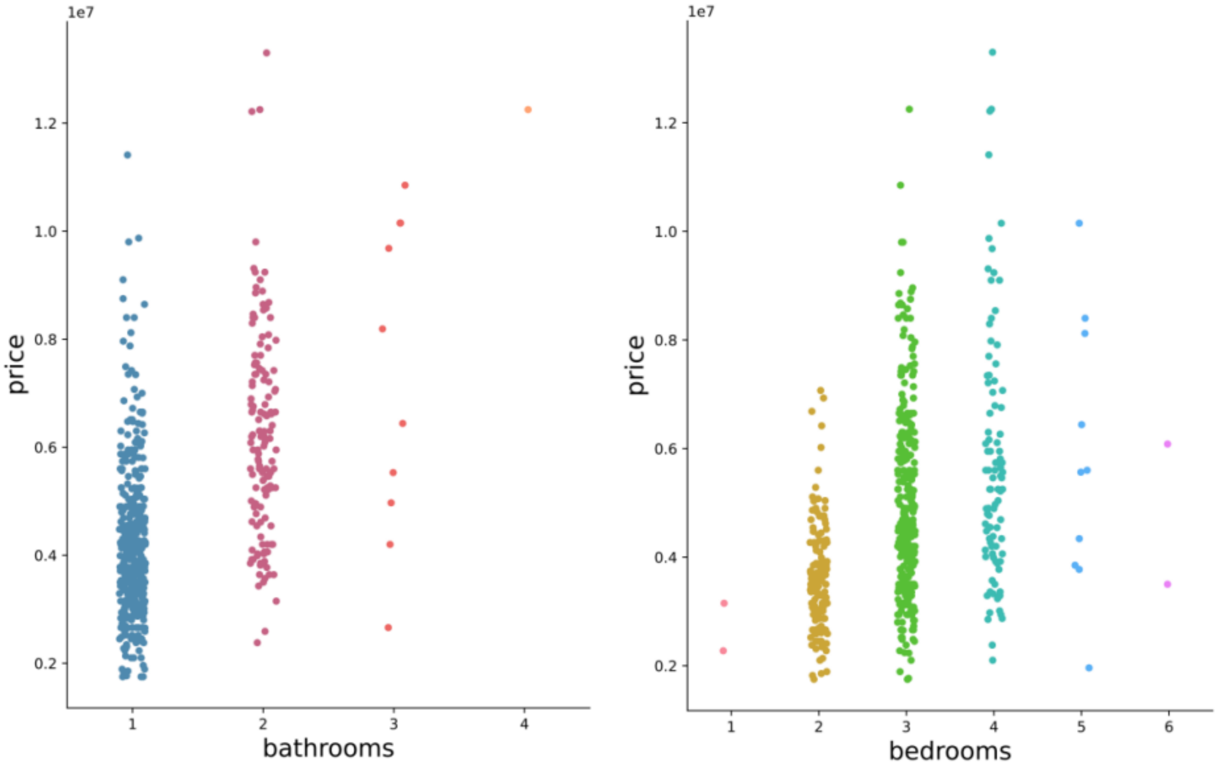


Figure 3. Categorical plot (bathroom, bedroom)

By utilizing customized scatter plots in Fig. 4, the relationship between Price and Area is illustrated. Despite some data points displaying dispersion, it broadly reflects the trend of house prices increasing with the growth of house size. However, this conclusion is not absolute and requires further model analysis and validation. This will aid in gaining a deeper understanding of the data's correlations, feature importance, and how to make precise and effective predictions of house prices while considering multiple factors simultaneously.

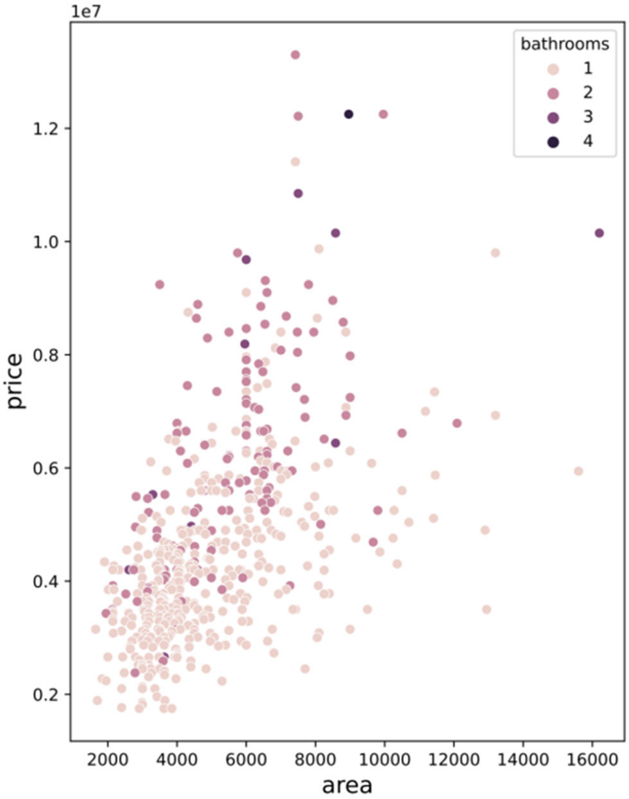


Figure 4. Scatter plot (bathroom, bedroom)

3.2. Testing the Proposed Model

The performance of these two models was determined through a detailed evaluation and comparison of the test data. Multiple Linear Regression excelled on the test data with an MSE of 750,408.72 and a coefficient of determination of 0.73. This indicates that the model can explain 73% of the variance in the target variable, and the errors between the predicted and actual values are relatively small.

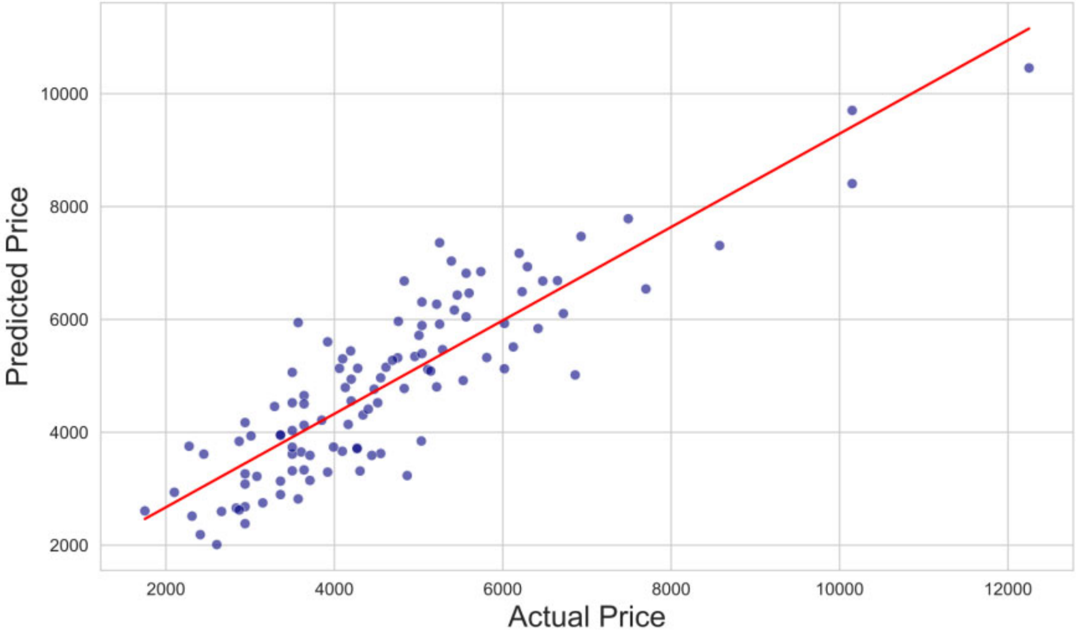


Figure 5. Actual vs predicted price with the line of best fit

Furthermore, using the pre-trained Random Forest model, the 'feature_importances' attribute was employed to obtain importance scores for each feature. These scores reflect the contribution of each feature to the model's performance. The top five features with the highest scores are Area, Bathrooms, Stories, Parking, and Furnishing status. These attributes have a significant influence on the model's prediction outcomes.

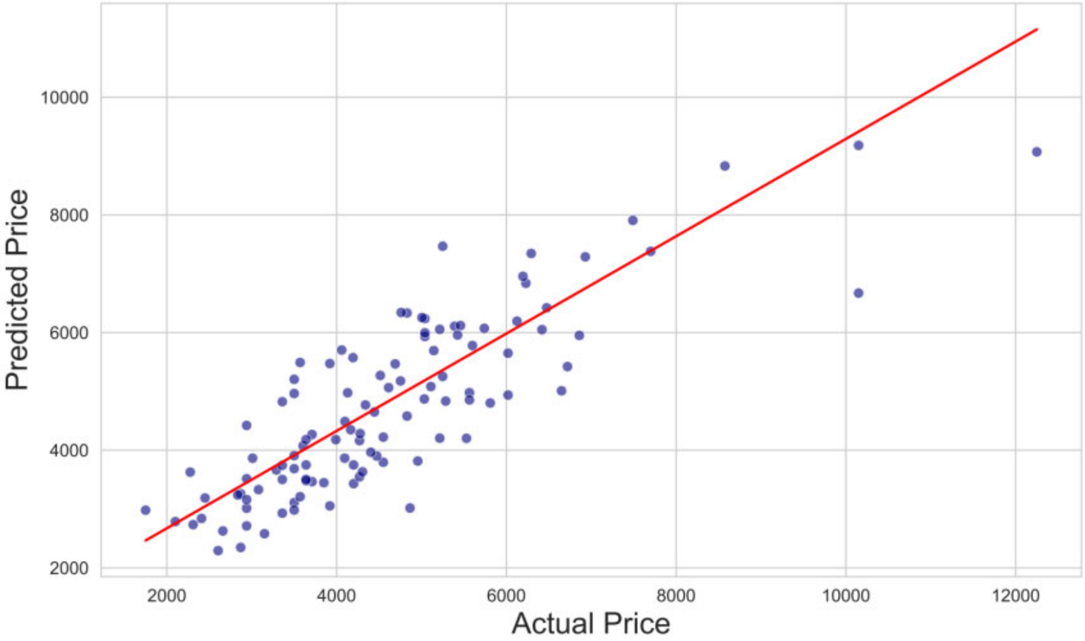


Figure 6. Actual vs predicted price with line of best fit (Random Forest)

Additionally, by utilizing the pre-trained Random Forest model and the 'feature_importances' attribute, importance scores for each feature were obtained. These scores indicate the extent to which each feature contributes to the model's overall performance. The top five features with the highest scores are Area, Bathrooms, Stories, Parking, and Furnishing status. These attributes exert a significant influence on the model's predictive outcomes.

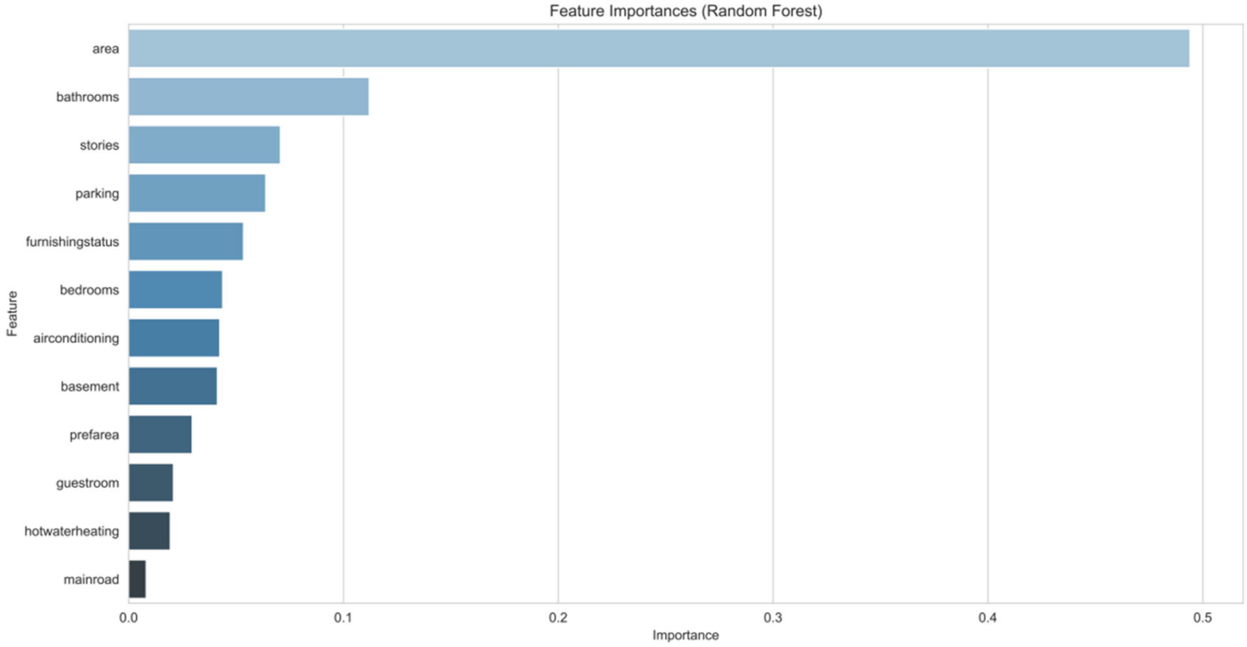


Figure 7. Feature importance (Random Forest)

In summary, the Multiple Linear Regression model demonstrated superior performance in this study, exhibiting lower mean squared error and a higher coefficient of determination, indicating its greater accuracy in predicting house prices. However, the Random Forest model remains a valuable alternative, especially when nonlinear relationships and feature importance need to be considered.

3.3. Discussion

Although the value of Random Forest's R2 in the test dataset shows mediocre results, the results are valuable for accurate, extensive house price data. Its feature importance analysis shows the extent to which each feature contributes to the model, which decision-makers can use to gain a finer understanding of home pricing in the context of the many factors that influence it.

In the future, consideration could be given to reducing the feature dimensions and removing or merging fewer essential features to simplify the model and improve computational efficiency. On the other hand, for features of high importance, attempts could be made to optimize the data collection further so that they better capture the critical information in the data. Performance can be improved by tuning the model's hyperparameters or choosing a different model type.

4. Conclusion

In The house price dataset was selected in this study, and a series of preprocessing tasks were performed, including data cleaning, feature selection, and normalization. Subsequently, this study introduces two models for house price prediction tasks: linear regression and random forest. Detailed characterization and visualization of the dataset are performed better to understand the characteristics and distribution of the data. In the model validation phase, this paper uses R2 as a performance evaluation metric to measure the model fitting effectiveness. The results show that the linear regression model achieves an R2 score of 0.73 in the experiments of this study, indicating that linear regression has relatively better performance in house price prediction tasks.

Overall, this study provides a preliminary exploration and analysis of the house price prediction problem, proposes applying different machine learning algorithms, and achieves particular success.

However, there is still much potential room for improvement and directions for in-depth research, including more complex feature engineering, model tuning, etc. The results of this study can provide valuable references and inspiration for future related research and practical applications.

References

[1] Adetunji, O. N. Akande A. B., Ajala F. A., et al. House price prediction using random forest machine learning technique. Procedia Computer Science, 2022, 199: 806 - 813.

[2] Kang Y., Zhang F., Peng W., et al. Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, July, 2020, 104919.

[3] Truong Q., Nguyen M., Dang H., et al. Housing price prediction via improved machine learning techniques. Procedia Computer Science, 2020, 174: 433 - 442.

[4] Zulkifley N. H., Rahman S. A., Ubaidullah N. H., et al. House price prediction using a machine learning model: a survey of literature. I.J. Modern Education and Computer Science, 2020, 6: 46 - 54.

[5] Gupta P., Zhang Q. Housing price prediction based on multiple linear regression. Scientific Programming, 2021.

[6] Liu X., Du H., Wen D. Research on house price prediction model based on graph neural network and long short-term memory model. Computer Application Research, 2023.

[7] Liao A. Research on house price prediction in Beijing based on ARIMA and PSO-BP combination model. M.S. thesis, Northern Minzu University, 2023.

[8] Cai T. Application of data mining techniques in house price prediction and analysis. Statistics Science and Practice, 2022, 10: 61 - 64.

[9] Ling F., Li Y. House price prediction model based on ensemble learning algorithm. Information and Computer (Theory Edition), 2022, 22: 96 - 100.

[10] Maxey J. The Effect of Pricing Factors on Real Estate Transactions in Prince George’s County. Maryland, 2013.