Model For Predicting London House Prices

Mingzhe Yin*

University of Minnesota Twin City, Minneapolis, United States

* Corresponding Author Email: Yin00378@umn.edu

Abstract. With the growth of population, housing prices have become one of the important indicators to reflect the economic performance. Forecasting prices around the world will be part of social and economic development. Through the analysis of various factors affecting the market housing price, it is helpful to make a more accurate assessment of the future housing price trend. In a broad sense, accurate housing price forecast is helpful to the country's macro-control of the market housing price trend, and the prediction of future housing price is part of the strategic planning of enterprises. For consumers, the housing price forecast has a positive effect on the rational planning of individual economy. Since house prices are related to many factors, and there is a linear relationship between house prices and some factors that affect house prices, there are many linear graphs to study this problem. The research direction of this paper is to obtain data from Kaggle and conduct exploratory analysis on these data sets. The main method used in this article is linear regression, and no other forecasting methods will be added. What this work needs to do is use the right machine learning methods to predict home values to create a complex property model under generally stable conditions, which is this work aims to do in this study. Through many data sets show that the multiple linear regression model does achieve a relatively excellent housing price prediction effect.

Keywords: House price; forecasting; linear regression; London.

1. Introduction

With the rapid development of artificial intelligence technology, machine learning is widely used in all walks of life and has achieved good results in many fields. Housing price is a hot issue with complex influencing factors, so it is difficult to make a comprehensive and accurate forecast. For our personal lives, when AI technology can use the right algorithms to guess which method predicts prices with the lowest error rate, it tends to help people make the most favorable choices. It is inevitable that people will be faced with buying or selling a house sooner or later in life. Therefore, it is of great significance to build an effective housing price forecasting model for the financial market and people's livelihood. The property market in the UK has long been a focus of public discussion, with news websites dedicated sections to covering important stories that may have an impact on recent developments and home prices. In addition to the tendencies. They represent the current housing market, which both buyers and owners find worrying. The nation's social climate and economic position. A lot of people consider purchasing a home to be one of their most crucial decisions. In addition to a home's affordability, other elements such as well as the long-term investment potential and the location's appeal influence how decisions are made.

This report will use housing price data from the City of London and its surrounding areas to predict house prices. These data are obtained from Kaggle. Specifically, this research conducts exploratory analysis on the data set, data transformation, fitting multiple linear regression model, model testing, and finally predicts the test data set.

The conclusions are as follows:

(1) Area_in_sq_ft_log has a positive correlation with house prices, City_County and House_Type has different influences according to different types.

(2) The multiple linear regression model has certain advantages for housing price prediction. The data has a total of 3480 observations and 10 variables, which provide information such as the location of the house and the room configuration. Some variables in this information are nominal variables, and they are meaningless from a statistical point of view, such as Postal_ Code; some variables are categorical variables, such as House_Type, so it is necessary for us to perform the

variables selection, finally this paper chooses Price, House_Type, Area_in_sq_ft, Bedrooms, Bathrooms, Receptions, City_County Total 7 variables enter the next step of analysis.

2. Methods

In addition to the multiple linear regression model used in this article to predict house prices, there are other ways to predict house prices. For example: OLS linear regression, random forest, K-NN algorithm, etc. The K-Nearest Neighbors (KNN) algorithm is a supervised learning method that can be used for both classification and regression, also known as the K-Nearest Neighbors algorithm. As a non-parametric supervised learning classifier, the KNN algorithm has the advantages of being very easy to understand and easy to implement. The KNN algorithm only needs to calculate the distance between different points of different feature data, so the distance can be easily calculated. The shortcomings of the KNN algorithm are also obvious. Its estimation and classification are slow. When the data set gradually increases, the efficiency of the KNN algorithm will drop a lot. Second, the KNN algorithm does not have the ability to handle missing.

OLS linear regression algorithm is the second one. It is a relatively simple algorithm, and OLS can be easily trained even on low-configuration systems. Of course, teaching linear regression is also very easy to understand and master, so linear regression is a very popular method. Second, the shortcomings of OLS linear regression. Like KNN algorithm, linear regression also has the problem of sensitivity to outliers. Outliers are outliers or extreme values where other data points deviate from the distribution, so it is best to analyze and remove outliers before they are applied to the dataset.

3. Exploratory data analysis

The typical approach to statistical analysis assumes that the sample follows a predetermined distribution and models the data before investigation. However, the findings of conventional statistical analysis are frequently disappointing since the majority of data cannot fit the hypothetical distribution. The exploratory data analysis method concentrates on the true distribution of the data and emphasizes data visualization by cleaning the data, describing, comparing the relationship between the data, cultivating the intuition of the data, summarizing the data, etc., so that the analyst can see the data at a glance. Discover the data's hidden rules so you can be inspired to assist the analyst create a model that fits the data. Consequently, we do an exploratory examination of the data prior to modeling (Table 1).

Table. 1 Traditional analysis

	variables	mean	min	max	sd
1	Price	1864172.540	180000.000	39750000.000	2267282.958
2	Area_in_sq_ft	1712.974	274.000	15405.000	1364.259
3	Bedrooms	3.104	0.000	10.000	1.518
4	Bathrooms	3.104	0.000	10.000	1.518
5	Receptions	3.104	0.000	10.000	1.518

Indicating the variance across distinct houses, the variable Price has a mean value of 1864172.540, a minimum value of 180000, a maximum value of 39750000, and a standard deviation of 2267282.958. Price has a significant knowledge and understanding gap, which is intuitive. It could lead to issues with future model design, such poor model prediction accuracy. The mean value is typically skewed towards the minimum value when compared to the minimum and maximum value for the remaining three numerical variables. This demonstrates how the three variables' distributions are all right skewed. Additionally, it will result in time violations. a few model presumptions. A histogram of the distribution of numerical variables is shown in figure 1.

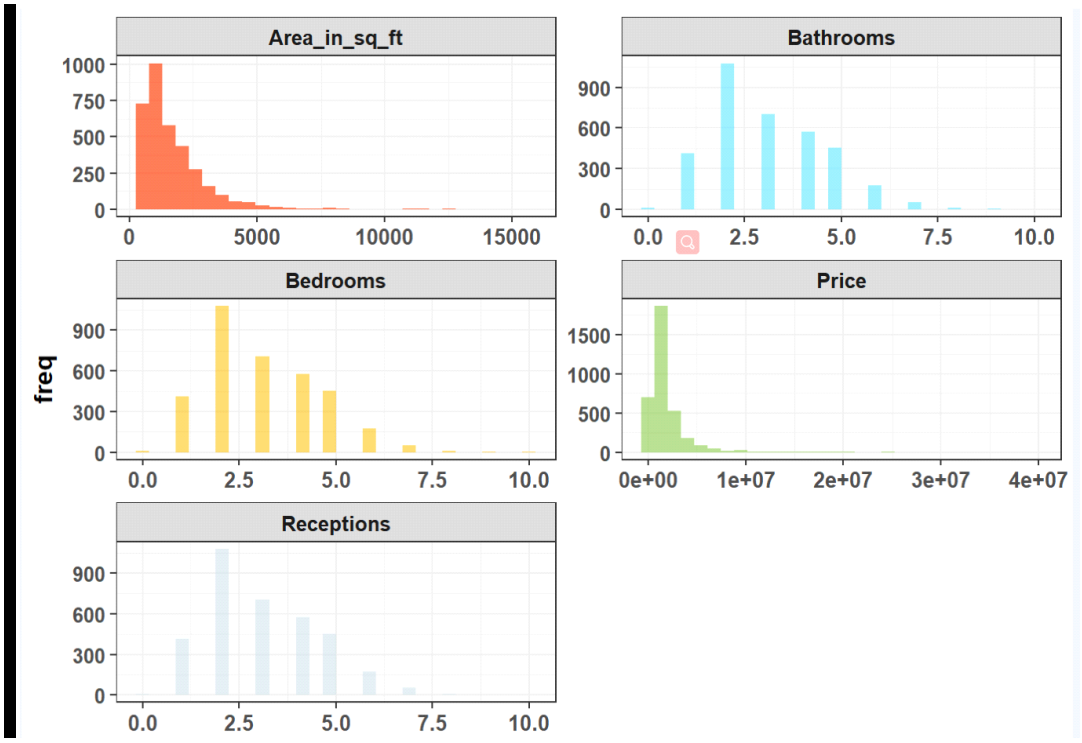

Fig. 1 Numeric Variables

The histogram makes it abundantly evident that the two variables Price and Area in sq ft have skewed distributions rather than following the normal distribution. Consequently, data transformation is required, and the others are There is no need for additional processing because the three variables' distributions roughly resemble a bell-shaped distribution.

Categorical variables must be processed after numerical variables have been treated. The first step is to count and sort the frequency of House Type in ascending order. The following table displays the findings. From the table 2, it can be found that the frequency of the two types of houses of Flat/Apartment and House are more than 1000. There are 1565 samples of Flat/Apartment type. 1430 samples are of the House type, and some of the frequencies of other types are even less than 10, such as Bungalow, which will make the training of this type of sample insufficient, thereby This leads to a decrease in the robustness of the model.

Table. 2 House_Type frequency

	Class	Freq
1	Flat / Apartment	1565
2	House	1430
3	New development	357
4	Penthouse	100
5	Studio	10
6	Bungalow	9
7	Duplex	7
8	Mews	2

4. Model Optional

By calculating statistics and creating uniform visualizations, we discovered the possible issues with the data set during the exploratory data analysis phase. This section describes various linear regression models that can be applied to this data set to create predictions. To begin applying the model, you must first convert the original data set such that it adheres to the model's presumptions. Here, we make the original data more in line with the predictions of the linear regression model by

using data processing techniques like recoding and logarithm. The original data set is then split into a training set and a verification set in order to reflect the prediction effect. After creating a multiple linear regression model using the training set, evaluate the model's hypotheses using residuals, standardized residuals, and other indicators while taking high leverage, outliers, collinearity, and heteroscedasticity into account. The best model should then be chosen after using stepwise regression techniques to assess the significance of various variables for the multiple linear regression model. Finally, the test set is predicted using the three models that were created from the training set, and the model's predictive power is assessed by computing the pertinent model prediction indicators.

The data for the multiple linear regression model must first be transformed. We applied the square root and logarithm transformations to Price and Area_in_sq_ft, and then drew the histogram once more. The outcome is depicted in the figure 2.

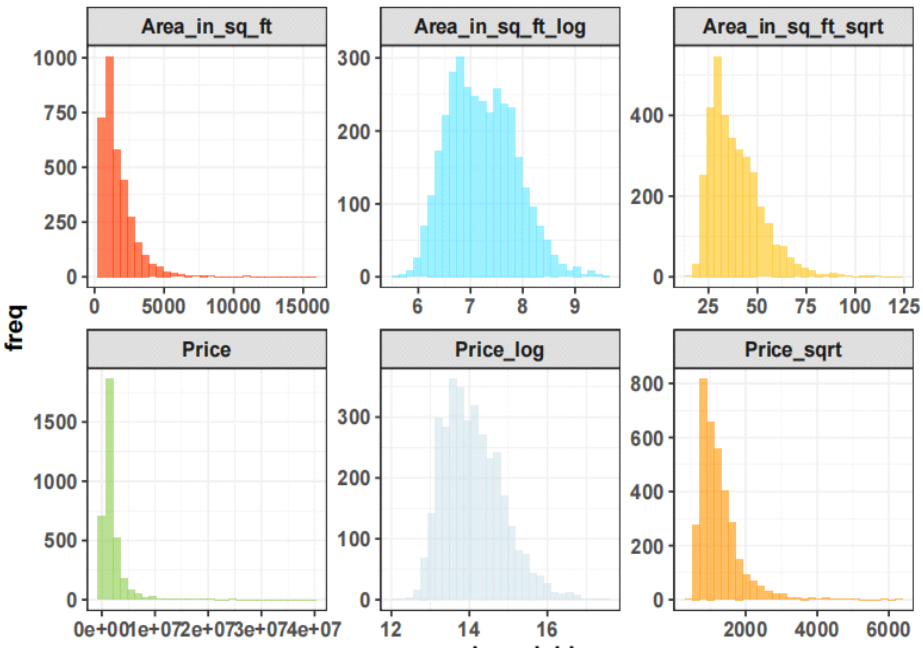

Fig. 2 Numeric Variables

The graph clearly demonstrates that whether it is logarithmic transformation or square root transformation, the transformed data approximately presents a normal distribution. Considering that the original data is economic data, this paper choses to use the logarithmic transformation to be more consistent Practical meaning. In this way, the new data set adds two variables which are Price_log and Area_in_sq_ft_log. For House_Type, given that the main categories are concentrated on types with a frequency greater than 100, here this work decided to encode all samples with a frequency less than 100 as others, which will make the modeling cost of the model is reduced a lot. But this also brings a problem, the model's ability to predict abnormal sample points of House_Type will be reduced. Also, for City_County, since the frequency is concentrated in the first few types, all types with a frequency less than 50 are encoded as others. Finally, to meet the modeling requirements, all these two-character variables are converted to factor types. To visually display the results after recoding, we still use box plots to show the distribution relationship between categorical variables and Price (figure 3).

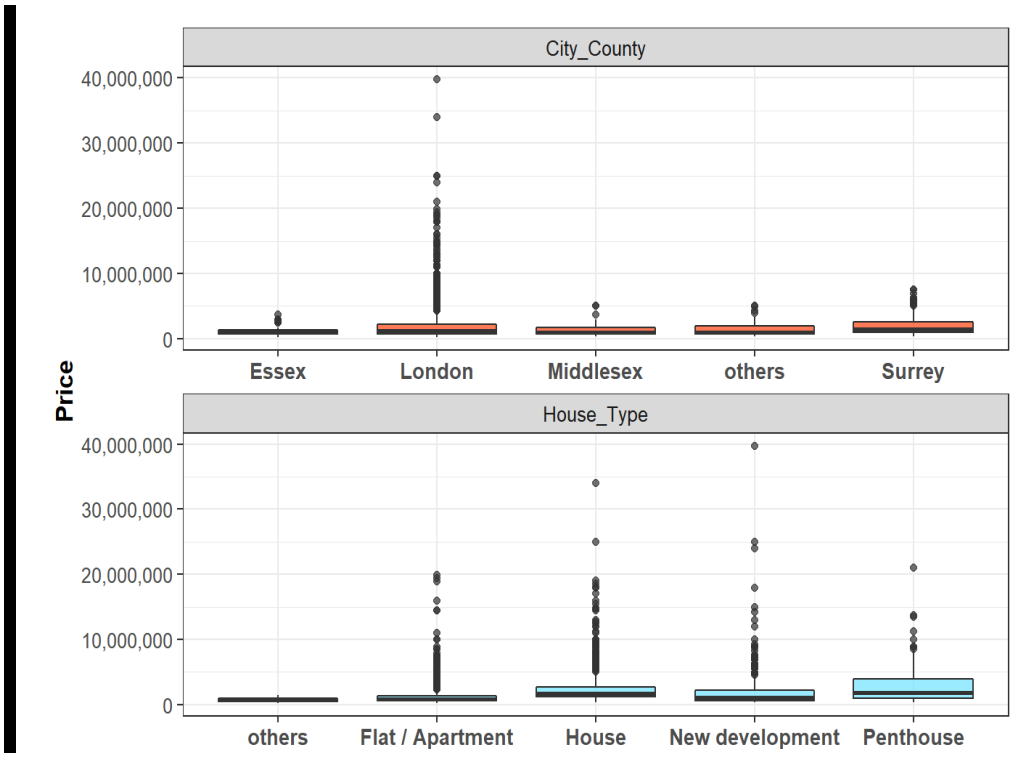

Fig. 3 Categorical variables

From the box plot, it is obvious that there are outliers in Price under the two categorical variables. Therefore, the logarithmic transformation performed in the previous article is very necessary. One can play a role in satisfying the model assumptions. The second one can reduce the variance.

Since the final decision to use the logarithm of Price and Area_in_sq_ft, and the transformed data must be highly correlated with the original data, only the logarithmic transformed data is used here to display relationship (figure 4).

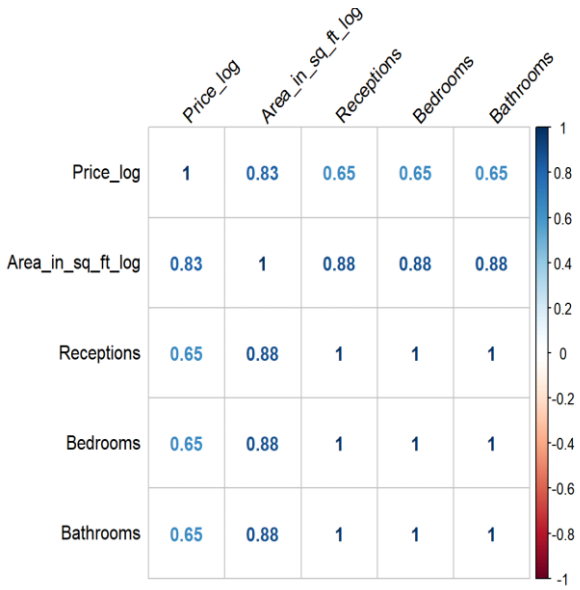

Fig. 4 The relationship of logarithmic transformed data

The graph demonstrates that all factors have a moderate level of connection with Price_log, and Area_in_sq_ft_log also has a strong and strong positive correlation with Receptions, which means that if you use Multiple linear regression models must consider the issue of multicollinearity. Since the correlation coefficients of the three variables of Bedrooms, Bathrooms, and Receptions are all 1, you can select only one variable to enter the model, here this paper chooses Bedrooms variable. At

the same time, Bedrooms and Area_in_sq_ft_log is highly positively correlated, so Bedrooms are not selected.

5. Model application

After that, formally create a multiple linear regression model; we have done so here. It is required to segregate the training set and the test set before developing the model. Here, this paper uses a training set of 75% of the data and a test set of 25% of the data (Table 3).

Table. 3 The training set

	Term	Estimate	std. error	Statistic	p.value
1	(Intercept)	5.293	0.114	46.241	0.000
2	House_TypeHouse	-0.225	0.021	-10.967	0.000
3	House_TypeNew development	0.112	0.025	4.476	0.000
4	House_Typeothers	-0.153	0.081	-1.896	0.058
5	House_TypePenthouse	0.107	0.046	2.328	0.020
6	City_CountyLondon	0.598	0.053	11.332	0.000
7	City_CountyMiddlesex	0.056	0.071	0.800	0.424
8	City_Countyothers	0.272	0.066	4.153	0.000
9	City_CountySurrey	0.104	0.058	1.780	0.075
10	Area_in_sq_ft_log	1.156	0.015	78.100	0.000

By outputting the model coefficients, it can be found that Area_in_sq_ft_log each additional unit will increase Price_log by an average of 1.156 units. For the variable House_Type, the benchmark level isFlat/Apartment. When changing from Flat/Apartment to House type, Price_log will decrease by 0.225 on average.

The outcomes of the analysis of variance reveal that all independent variables have P values that are very near to 0, which means that they all significantly affect the dependent variable (Table 4).

Table. 4 Analysis of Variance

	Term	df	sumsq	meansq	statistic	p.value
1	House_Type	4	294.938	73.735	540.670	0.000
2	City_County	4	30.483	7.621	55.879	0.000
3	Area_in_sq_ft_log	1	831.851	831.851	6099.673	0.000
4	Residuals	2600	354.578	0.136	-	-

6. Result

The model is repeatedly checked from the perspectives of residuals, normalized residuals, hat values, cooks.distance, etc. once the whole model has been built. The samples are tagged in row order to aid in the better identification of outliers and powerful influence points. From the figure 5, the three statistics of hat values, cooks. distance and dffits seem to be invalid due to the excessive number of samples and are affected by n. On the other hand, there are partial separation cluster points and strong influence points in the training set.

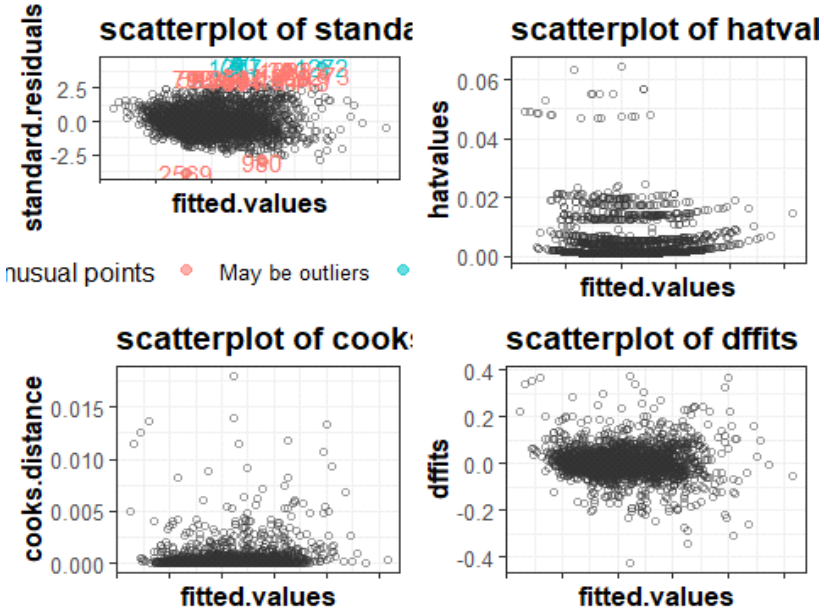

Fig. 5 The statistics of three different parameters

7. Constant Error Variance

In relation to homoscedasticity, heteroskedasticity exists. To make sure that the regression parameter estimator has good statistical features, the so-called homoscedasticity is used. The overall regression function's random error term adheres to homoscedasticity, which means they all have the same variance, and this is a key tenet of the traditional linear regression model. The linear regression model is heteroscedastic if this presumption is false, i.e., the random error factor has distinct variances.

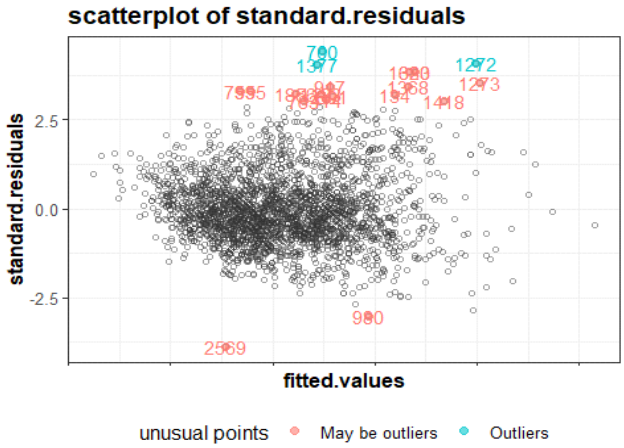

Fig. 6 The result of scatterplot

Although there are obviously abnormal sample points in the fig.6, considering that there are too many sample points, it can be considered that this model does not have serious issues with heteroscedasticity.

8. Error Normality

The residual normalcy is subsequently tested. The prior essay employed the statistical test approach. This is merely a repetition of the graph test approach to provide a more understandable result. The residual histogram and the residual QQ graph are the graph test techniques that are employed.

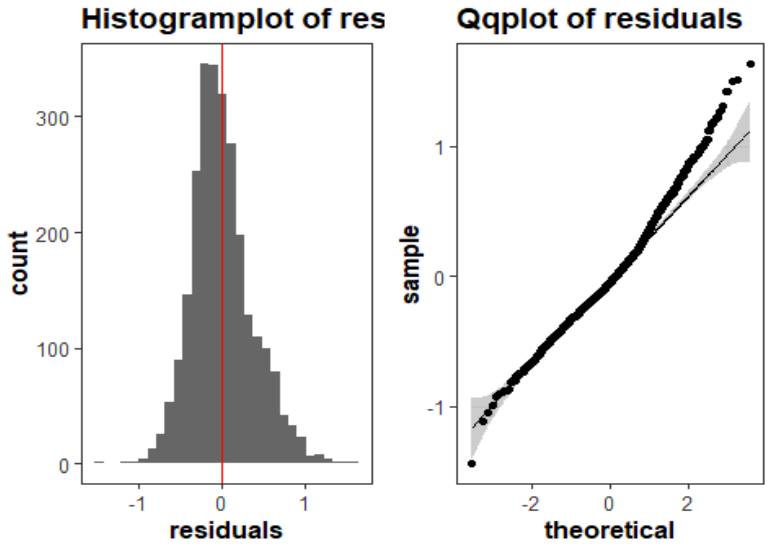

Fig. 7 The residual histogram and the residual QQ graph

First analyze the histogram. From the figure 7, the mean of the residuals is very close to 0 and combined with the distribution shapes on both sides of the mean, the residuals are roughly normal distribution; then look at the QQ chart, although there are some samples The points deviate from the confidence zone of 95%, but the population is distributed within the region. Therefore, it can be determined that the residuals are normally distributed through two graphs.

9. Collinearity

Multicollinearity is the term used to describe how precise or high correlation can cause the explanatory variables in a linear regression model to be skewed or difficult to estimate precisely. There is generally a link between the explanatory variables in the design matrix because of the inappropriate model design caused by the lack of sufficient economic data. Complete collinearity is an uncommon occurrence. There is typically some approximation of collinearity, or collinearity. The model's multicollinearity is examined in this instance.

Table. 5 Variance of independent variable

	Variables	GVIF	Df	GVIF^(1/(2*Df))
1	House_Type	1.846	4.000	1.080
2	City_County	1.161	4.000	1.019
3	Area_in_sq_ft_log	1.745	1.000	1.321

As can be observed in table 5, none of the independent variables have a variance expansion factor more than 5, hence multicollinearity is not a concern. This is due to the early-stage deletion of several highly associated independent variables.

10. Variable selection

A variety of techniques can be used to compare models based on different criteria. All fit measures rarely agree on a single model as the "best" model. Here, we choose variables by using the best subsets (Table 6-8).

Like coefficient of determination, adjusted R square is a model effectiveness metric that takes model complexity into account. The full model is the best from the standpoint of adjusting goodness of fit, and its goodness of fit reaches 0.777, which is a commendable result.

The goodness of fit metric that accounts for complexity is Mallow's Cp. Similar to this, Mallow's Cp demonstrates that the whole model is ideal.

The goodness of fit metric that takes complexity into account is the Akaike Information Criterion. AIC demonstrates that the entire model is the best as well. In conclusion, we may select the entire model as the ideal model.

Table. 6 R Square

	Mindex	n	Predictors	adjr	cp	aic
1	7	3	House_Type City_County Area_in_sq_ft_log	0.765	-2.000	2218.839
2	4	2	City_County Area_in_sq_ft_log	0.746	207.699	2415.143
3	5	2	House_Type Area_in_sq_ft_log	0.718	518.740	2688.989

Table. 7 Mallow's Cp

	mindex	n	predictors	adjr	cp	aic
1	7	3	House_Type City_County Area_in_sq_ft_log	0.765	-2.000	2218.839
2	4	2	City_County Area_in_sq_ft_log	0.746	207.699	2415.143
3	5	2	House_Type Area_in_sq_ft_log	0.718	518.740	2688.989

Table. 8 Akaike's Information

	mindex	n	predictors	adjr	cp	aic
1	7	3	House_Type City_County Area_in_sq_ft_log	0.765	-2.000	2218.839
2	4	2	City_County Area_in_sq_ft_log	0.746	207.699	2415.143
3	5	2	House_Type Area_in_sq_ft_log	0.718	518.740	2688.989

11. Make a Forecast

Building a model that can forecast the new data set is the ultimate objective. We don't need to process the test set separately because the data in both the training set and the test set are processed using the logarithm and other variables. The prior multiple linear regression model is used directly here to make predictions.

Then, we utilize the model developed from the training set to predict the test set by performing an inverse transformation on the fitting value acquired by the training set model to obtain the fitting value of Price. The model training set and test set's rmse, rsq, and mae are displayed in the table below.

According to the table, the root mean square error of the multiple linear regression model for the training set is 1486728, meaning that each house's average fitting difference on the training set is 1486728; goodness of fit in real-world applications, it is already extremely close to ideal at 0.631, and the average absolute error is 610136, which is just somewhat larger than the root mean square error.

Table. 9 Model training set

	key	.metric	.estimate
1	mlr_train	rmse	1529861.634
2	mlr_train	rsq	0.609
3	mlr_train	mae	631746.405
4	mlr_test	rmse	1292455.266
5	mlr_test	rsq	0.561
6	mlr_test	mae	617063.363

12. Conclusion

As a kind of time series data, housing prices are affected by many factors. This article takes housing data in London as an example, and tries to predict housing prices from the perspective of the geographic location factors and specific housing configuration information using multiple linear regression models

Only a portion of London's area data are included in this data set. The robustness of the model will be improved, and the corresponding prediction accuracy and reliability will also be improved, if a bigger region or more different sorts of samples are obtained. However, there are several issues with this sample as well. The data set used in this project only contains housing price data for London and the surrounding areas, the data collection time is not yet known, and whether the model can be maintained for a long time, all of which may prevent the model from being able to predict on very large data. Additionally, the multiple linear regression model's sample range is limited, and pre-processing will only handle some categorical variables with a frequency less than a certain value in crude fashion. Future study may therefore need to take into account more time-sensitive data, which will enhance the model's usefulness for society, or data with broader dimensions, such as year. Built and other variables, which can also enhance the fit. Unsurprisingly, the data will also need to be updated to reflect new circumstances as house prices alter in the future. Future researchers will be responsible for this duty, nevertheless.

References

[1] Shafiei Chafi, Z.,; Afrakhte, H. Short-term load forecasting using neural network and particle swarm optimization (PSO) algorithm. Mathematical Problems in Engineering. 2022 26: 362-371.

[2] Amjady N. Short-term hourly load forecasting using time-series modeling with peak load estimation capability. IEEE Transactions on power systems, 2001, 16(3): 498-505.

[3] Wang Y, Zhao Q. House Price Prediction Based on Machine Learning: A Case of King County, 2022 7th International Conference on Financial Innovation and Economic Development. Atlantis Press, 2022: 1547-1555.

[4] Madhuri C H R, Anuradha G, Pujitha M V. House price prediction using regression techniques: a comparative study, 2019 International conference on smart structures and systems (ICSSS). IEEE, 2019: 1-5.

[5] Jie F, Zhong LY., Dong ML, et al. Moderation effect analysis based multiple linear regression. Journal of Psychological science, 2015 (3): 715.

[6] Zhang S, Cheng D, Deng Z, et al. A novel kNN algorithm with data-driven k parameter computation. Pattern Recognition Letters, 2018, 109: 44-54.

[7] Anggraini N, Tursina M J. Sentiment analysis of school zoning system on Youtube social media using the K-nearest neighbor with levenshtein distance algorithm. 2019 7th International Conference on Cyber and IT Service Management (CITSM). IEEE, 2019, 7: 1-4.

[8] Ho W K O, Tang B S, Wong S W. Predicting property prices with machine learning algorithms. Journal of Property Research, 2021, 38(1): 48-70.

[9] Madhuri C H R, Anuradha G, Pujitha M V. House price prediction using regression techniques: a comparative study. 2019 International conference on smart structures and systems (ICSSS). IEEE, 2019: 1-5.

[10] Liu X. Spatial and temporal dependence in house price prediction. The Journal of Real Estate Finance and Economics, 2013, 47(2): 341-369.