ISSN 2616-5902 Vol. 5, Issue 2: 34-40, DOI: 10.25236/AJBM.2023.050207

Machine Learning-Based Analysis of Housing Price Predictors

Haoyue Zou[1,a,*]

[1]Faculty of Science, Hunan University of Technology, Zhuzhou, China
[a]Zzouhy86k@163.com
**Corresponding author*

Abstract: *With the rapid development of machine learning and related fields, it has been widely used in various industries. House price is a hot topic in the nation's livelihood, and the influencing factors are very complex. Therefore, a dataset with a large number of features is chosen as the data for processing in this paper. Firstly, the least-squares and random forest regression algorithms in machine learning are introduced; afterwards, the training and test set data are processed and analysed to identify the seven main factors affecting housing prices; finally, the two models are evaluated, and the results show that both algorithms can accurately predict housing prices.*

Keywords: *Housing price forecasting; Least-squares; Random Forest regression algorithms*

1. Introduction

1.1. Research significance

The real estate industry, as an industry closely related to people's lives, is the economic pillar of the country. Any slight change in it would tremendously impact the livelihood of the country and its residents. As the living standards of the country's residents improve and the population migrates gradually to the cities or towns, the movement of rural residents will inevitably bring an increasing housing demand. Any slight change in housing prices will tremendously impact the country's and its people's lives. Therefore, Use real home sales price data and machine learning models to build prediction models. At the same time, providing highly accurate and stable forecasting models can effectively predict house prices, offer government data support, facilitate the government to formulate macro-control policies and maintain social stability.

1.2. Research status at national and abroad

Nowadays, several machine learning-related housing price prediction models have been developed overseas, such as random forests, support vector machines and K-nearest neighbour methods. In 2001, Nguyen et al. compared neural networks and multiple linear regression models for predicting housing prices and made a conclusion that the accuracy of housing prices of netural networks was more accurate when the data set was more extensive. In 2010, K.C. La et al. used support vector machine methods to predict real estate prices. They demonstrated that support vector machine models outperformed multiple linear regression models and artificial neural network methods in forecast results. In 2012, Evgeny et al. compared four models, the random forest regression model, the multiple linear regression model, the decision tree and the artificial neural network. They concluded that the random forest regression model had the highest prediction accuracy. In 2017, Alfiyan et al. improved the accuracy of their prediction model by combining a particle swarm algorithm with a regression model to forecast house prices in Malang city.

The research on domestic house price forecasting gradually shifted from qualitative to quantitative analysis, and researchers in China-proposed using theories and models to forecast housing prices. Li Dongyue et al. used the grey model GM(1,1) to predict house prices. Meanwhile, In 2004, Yang Liming applied an improved artificial neural network to house price forecasting. In 2011, Qiu Qirong built a model to forecast house prices based on principal component analysis. In 2018, Tao Guyu compared basic linear regression, ridge regression, and Lasso regression modelling to predict house prices in Ames, concluding that the ridge regression and Lasso regression models mitigate the problem of models falling into overfitting.

1.3. Innovation

There are two main innovations in data and analytical methods.

Data: This article first pre-processes and analyses the data to build a prediction model using real housing data information, dramatically reducing the errors that missing data may cause.

Analytical methods: The simple and advanced models are combined as the analysis methods. A relatively simple OLS (Least-squares) model is used to construct a simulation fit first and then introduce random forest models to establish the regression model. Because this model has several advantages, such as no over-fitting, fast classification and few adjustment parameters, it provides better recognition accuracy and stability.

2. Preparatory knowledge

2.1. Least-square method

The least-square method is vital in estimating the unknown parameters in linear regression. The basic idea is to find the case where the sum of squares of the differences between the estimated and actual values of the model is the smallest and to use this value as the parameter estimate, i.e., to find the best function fit by minimizing the sum of squares of the errors.

2.2. The principle of least-square

There are one-dimensional linear regression models:

$$Y_i = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{1}$$

ε is the random error term and $\beta_{0\cdots n}$, v are the regression coefficients.When the value of Q is minimal, the resulting $\hat{\beta}_0, \hat{\beta}_1 \cdots \hat{\beta}_n$ is called the least-square method estimator. As shown below:

$$Q = \sum_{i=1}^{n}(Y_i - \widehat{Y_\iota})^2 = \sum_{i=1}^{n}[Y_i - (\widehat{\beta_0}X_0 + \cdots + \widehat{\beta_n}X_n + \varepsilon)]^2 \tag{2}$$

Using the least-square method, it is easy to find the unknown data ($\hat{\beta}_0, \hat{\beta}_1 \cdots \hat{\beta}_n$) and minimise the sum of squared errors between the found data and the true data.

2.3. Advantages and disadvantages of the least-squares

Model advantages:

Least squares regression is relatively simple and easy to master. Meanwhile, the smallest sum of squares of the residuals is the optimal treatment and does not affect accuracy in the case of large amounts of data.

Model disadvantages:

The model can only handle linear regression problems and is highly ineffective in simulating non-linear ones. On the other hand, the model is susceptible to outlier values, which can easily cause extreme distortion and have a short application horizon.

2.4. Random Forest Algorithms

The Random Forest algorithm is a comprehensive decision tree algorithm. This algorithm creates a forest of a large number of unrelated decision trees, which can be used to compute either classification or regression problems. Classification problems primarily use decision tree voting, while regression problems use the average of multiple decision tree predictions as the regression result. The random forest algorithm's parallel generation of decision trees can effectively suppress overfitting and improve operational efficiency. At the same time, the number of decision trees can also be modified by changing the parameter values, which is the main reason why the Random Forest algorithm is currently an extremely popular machine learning algorithm.

2.5. Principles of Random Forest Regression Models

The random forest regression model is constructed as multiple regression trees by generating multiple datasets by the bootstrap method. The model trains several decision trees separately, and the learning process is parallel, which can effectively reduce the risk of over-fitting.

Mathematically this can be summarized as follows. Given a sample of data X and a set of predictions Y, a forest is created from a random variable θ to form a prediction tree $h(x, \theta k)$ that generates values, and the prediction tree is averaged over k to obtain a random forest predictor. [1-2]

2.6. Advantages and disadvantages of random forest regression models

Model advantages:

The random forest regression algorithm uses an integrated algorithm that offers more accuracy than most single algorithms. Also, random forest regression demonstrates some resistance to noise (Error) due to the randomization of samples and features. It is not easily over-fitted, making it a better model for outlier data sets. Most importantly, the algorithm is able to handle data with a large number of eigenvalues and can handle both discrete and continuous data without planning the data.

Model disadvantages:

The model has more branches on certain sample sets with more noise and tends to be over-fitted. Some aspects of random forest regression are similar to black-box models, which are challenging to interpret reasonably. Also, the space and time required for the model are proportional to the number of decision trees in the random forest. Therefore, too many decision trees can make the training time too long.

3. Experimental data processing

3.1. Data sources

This paper selected the data from the Kaggle competition on the 2016 website. The training data set provides 1,460 different housing price data samples, each containing 80 house characteristics, such as the type of home. The test data set provides 1,459 house samples and 79 house characteristics in addition to the house's selling price.

3.2. Data pre-processing

Feature engineering plays a critical role in improving the performance of a model. It takes the original dataset to clean, dimensionless, and normalize the feature values of the data in the original data set by processing and constructing the features of the data. More importantly, A good selection model can select better training feature values, allowing simple models to perform well and complex models to be more accurate. This section revolves around the process common to feature engineering.

To determine the relationship between each characteristic and the selling price of a house, we need to study and analyze the data. First of all, we use the *seaborn* library in *Python* to draw a histogram of the selling price of houses. As shown in Figure 1:

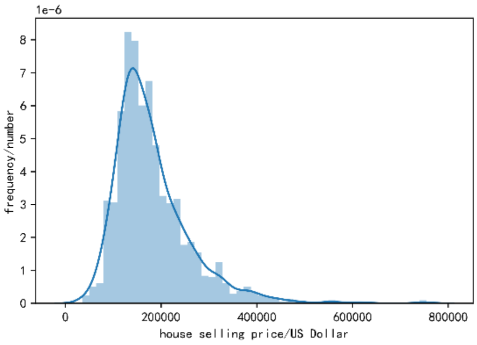

Figure 1: Histogram of housing prices

A kernel density estimation curve (KDE) can be seen in Figure 1, which can be used to estimate an unknown density function. The curve allows for a more intuitive view that the housing price data conforms to a normal distribution and allows for the next step in the discussion [3].

Analysis of the training set data reveals some house samples with missing features, such as pool quality and various features. This is shown in the figure 2.

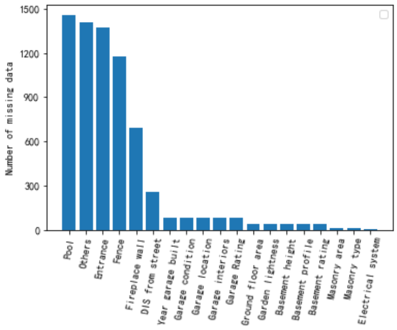

Figure 2: Missing data for housing features

The analysis of missing data is of paramount importance in pre-processing, where missing data can mean errors in the collected dataset or a reduction in sample size. If not handled properly, this can lead to significant errors in the subsequent modelling and analysis. The main methods commonly used are data padding and the removal of missing sample values. As shown in Figure 2, the test dataset has a total of 1459 house samples, and the features with more than 80% missing data are pool, entrance, fence and others. The features in this category can be considered invalid and dealt with by removing the disappeared sample values. The remaining features, such as the fireplace wall and a total of 19 features, have a low missing ratio, so the mean value of the feature is taken as its missing value.

3.3. Selecting the best features

Among the 80 characteristics, this part selected several high degrees of correlation with the selling housing price. As for the continuous variables, Pearson's correlation coefficient can be used as a statistical indicator which reflects the strength of the linear relationship between two continuous variables. Therefore, Utilizing the Pearson correlation coefficient can obtain the correlation strength and the housing price. Such as the joint distribution of living area, the Pearson correlation coefficient is 0.71, a strong positive correlation.

For categorical variables, correlations can be seen by the degree of change in the target variable on each categorical value. Generally, the higher the material and finish grade of the house, the more expensive the housing price. However, both methods must analyze each feature and are time-consuming and tedious. Thus, the following section uses the heat map approach to analyze the correlation between the target variables and features, i.e., a comprehensive analysis of the features in the dataset to find the best features.

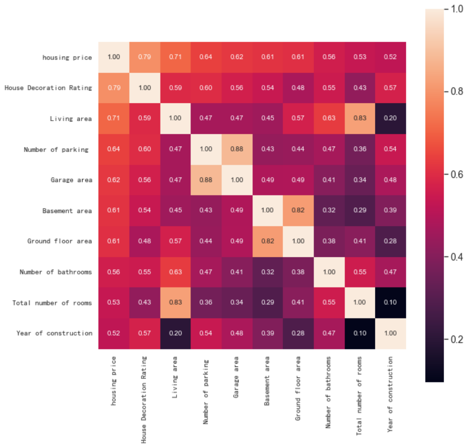

Figure 3: Relationship matrix heat map

The heat map shows the correlation coefficients between two out of 80 characteristics, resulting in the nine characteristics influencing home prices the most. In order: house decoration rating, living area, number of parking, garage area, basement area, ground floor area, number of bathrooms, the total number of rooms and year of construction. The heat map of their relationship matrix is shown in Figure 3 [4].

From Figure 3, the two variables, home decoration rating and living area, have a strong linear correlation with the house's selling price at 0.79 and 0.71, respectively. Therefore, it was selected for assessment characteristics.

There is also a strong linear relationship between the number of parking, garage area and housing price, but both are vehicle factors, and only one of them can be taken. Therefore, the former is used as the target variable since it is more strongly related to the target variable.

Likewise, the basement and first-floor areas can be taken as only one or the other. Choose the former.

Although the correlation between the number of bathrooms and the year of construction is slightly lower, the more the number of bathrooms and the year of the house affect the selling price to some extent.

Therefore, the following seven characteristics were selected: house decoration rating, living area, number of parking spaces, basement area, number of bathrooms, the total number of rooms, and year of construction.

4. Machine learning model building and evaluation

4.1. Least-square methods

In this module, the OLS library is used to model the seven features' previous section to derive the regression coefficients (retaining two valid digits), as shown in Table 1.

Table 1: Least-square method regression coefficients

Feature Name	Coefficient
Example text 1	Example text 2
Intercept	-771294.05
House Decoration Rating [T.2]	7237.34
House Decoration Rating [T.3]	-4403.67
House Decoration Rating [T.4]	12564.30
House Decoration Rating [T.5]	21290.46
House Decoration Rating [T.6]	31422.13
House Decoration Rating [T.7]	46231.52
House Decoration Rating [T.8]	86073.10
House Decoration Rating [T.9]	157768.64
House Decoration Rating [T.10]	154703.80
No. of parking spaces [T.1]	16198.83
No. of parking spaces [T.2]	25019.98
No. of parking spaces [T.3]	55873.97
No. of parking spaces [T.4]	30379.66
No. of bathrooms [T.1]	-95.45
No. of bathrooms [T.2]	-1255.69
No. of bathrooms [T.3]	41577.53
Living area	44.66
Basement area	14.66
Year of construction	406.82

4.2. Random Forest Regression Algorithms

In this part, a random forest algorithm is adopted to analyze the seven features, as shown in Table 2.

Table 2: Eigenvalues of Random Forest Regression Algorithms

Feature Name	Eigenvalues
House Decoration Rating	5
Living area	896
No. of parking spaces	1
Basement area	882
No. of bathrooms	1
Total number of rooms	5
Year of construction	1961
House selling price	94712

4.3. Model Evaluation

The decidable coefficient R^2 is used as the evaluation criterion of the model. Its formula is shown as follows:

$$R^2 = \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{3}$$

$y_i$is the true value, $\bar{y}$ is the$y$ mean, $\hat{y_i}$ is the predicted value of y, and m is the sample size. By solving in $python$ this can be obtained in Table 3.

Table 3: Decidability factor of two machine learning methods

Machine learning models	R^2
Least-square method	0.800645
Random Forest Regression Algorithms	0.850784

The value of the decidable coefficient R^2 is the value between $0\sim1$. The closer the value of $R^2$is to 1, the better the model fit, and vice versa, the closer it is to 0, the worse the model fit. As shown in the table above, $R^2$is greater than 0.8 for both the least-square method and random forest regression algorithms. Meanwhile, the accuracy of the latter algorithm is about 85.1% (retaining three as valid data), which is greater than the former by about 5.1% percentage points. Obviously, this latter algorithm is better than the former.

5. Reviews and Conclusions

Based on the review of relevant literature at home and abroad, the paper selects the 2016 Kaggle competition Boston house price data as the dataset. Then divides the data into a test set and a test set and trains them, and derives that the accuracy of the random forest regression model is higher than that of the OLS model. However, the OLS model inadequately considers all the factors, and the random forest model could have a significant error when there are too many branches. For this reason, further research of the thesis consists of three main points as follows.

(1) By combining the random forest regression algorithm, introduce other algorithms with higher accuracy and performance to enhance the accuracy of this algorithm. Meanwhile, the algorithm is improved and optimized to obtain a better algorithm, thus further improving the accuracy or reducing the time complexity.

(2) Subsequent experiments will use the logistic regression model in 'Sklearn' with Keras neural network to build a model for the multi-dimensional prediction of house prices.

(3) More detailed data pre-processing of the dataset to ensure that the missing data is entered closer to the actual situation.

In essence, the prediction of house prices is a classical regression problem. In this paper, we predict house prices based on least-squares and random forest regression algorithms and then demonstrate that the two machine learning methods can predict house prices more accurately through decidable coefficients. This reason shows that machine learning has a promising future in the real estate industry.

References

[1] Xu, G., Zhang, K., (2014). Property price evaluation based on random forest model. Statistics and decision making., 17:4

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

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

[4] Tian, R., (2019). Boston house price prediction based on multiple machine learning algorithms. China New Communications, 11: 228-230