Hindawi
Scientific Programming
Volume 2021, Article ID 7678931, 9 pages
https://doi.org/10.1155/2021/7678931

Research Article

Housing Price Prediction Based on Multiple Linear Regression

Qingqi Zhang

The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Correspondence should be addressed to Qingqi Zhang; qzhangbs@connect.ust.hk

Received 24 August 2021; Accepted 18 October 2021; Published 29 October 2021

Academic Editor: Punit Gupta

Copyright © 2021 Qingqi Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, the author first analyzes the major factors affecting housing prices with Spearman correlation coefficient, selects significant factors influencing general housing prices, and conducts a combined analysis algorithm. Then, the author establishes a multiple linear regression model for housing price prediction and applies the data set of real estate prices in Boston to test the method. Through the data analysis and test in this paper, it can be summarized that the multiple linear regression model can effectively predict and analyze the housing price to some extent, while the algorithm can still be improved through more advanced machine learning methods.

1. Introduction

The real estate industry has been one of the leading researches focusing on modern economics, for its significant implications on relevant industries and fields such as construction, investment, and public welfare. In general, purchasing and investing in any real estate project will involve various transactions between different parties. Thus, it could be a vital decision for both households and enterprises. How to construct a realistic model to precisely predict the price of real estate has been a challenging topic with great potential for further research [1]. It is generally believed by academia that correctly predicting the special price for a specific real estate is impractical since it involves plenty of factors exerting influence on the eventual cost.

There is a well-known saying about the appraisal of real estate by Li Ka-Shing, the most famous property tycoon in Hong Kong: "Three major factors are determining the price of a property, the first one is location, the second one is location, and the third one is still location." His word does not seem to make much sense from a statistical research perspective. Nevertheless, as a successful businessman in the property industry, what makes him attach so much significance to some specific factors like location when appraising a property is crucial. To what extent a particular factor like location plays an essential role in pricing a property is worth exploring by adopting a statistical model in real estate economics research.

The appraisal of real estate is traditionally conducted by a licensed professional, who would carry out a holistic survey based on several factors such as location, surroundings, areas, and facilities of a real estate. Nevertheless, the manual appraisal would inevitably have the possibility to involve the appraisers' factors and vested interest. This potential risk would likely cause a biased or subjective evaluation of a particular real estate, bringing loss for investors or households [2]. Thus, constructing a feasible algorithm and automated model which could appraise the real estate impartially and objectively has critical significance for any potential parties participating in these transactions.

According to economics principles, the market price of properties is attained when the demand and supply curves intersect with each other, which is subject to various factors, both subjectively and objectively. It is doubtful in practice that the market price of a property will equal the market value, as the market for real estate has been too unpredictable and fluctuating to be considered as an ideal market [3]. Affected by many subjective factors, it is significantly crucial for real estate appraisers to figure out the objective factors that account most for the pricing of properties.

In modern research on the property industry, advanced research methods such as machine learning and artificial

intelligence have been widely adopted in many aspects. Not only are they utilized in evaluating the price and value but also they are applied to figure out potential future applications and would-be challenges [4]. The comprehensive adoption of machine learning and artificial intelligence in the property industry has generally transformed this experience-driven industry with great arbitrage opportunities to an intelligent and data-driven enterprise [5].

Real estate appraisal has been divided into mainstreams at the current stage: mass appraisal and individual appraisal. The personal assessment is conducted when specific values are given for different characteristics of a given real estate. At the same time, mass appraisal adopts a systematic methodology to perform a precise assessment for a group of properties by adopting standardized procedures and rigorous testing in statistics [6]. Besides traditional techniques like the linear regression model, modern mass appraisal methodology has broadly incorporated computational intelligence approaches such as support vector machine (SVM), multilayer perceptron (MLP), and neural network. In practice, it has been widely employed in property appraisal for various purposes such as taxation and investment forecasting [4].

The current study of price modeling in real estate has mainly been based on the theory of hedonic prices, which was initiated by a famous economist Sherwin Rosen. His approach is generally regarded as feasible and could be applied to extensive research in the real estate field by academia. According to his theory, the property price could be characterized as a utility function of many relevant variables such as structural characters, neighborhoods, and the environment [7]. Based on this theory, pricing models for real estate are generally built through a multiple regression model, in which many general assumptions such as independence, homoskedasticity, and normal distribution of residuals must be fulfilled [8].

The definition of hedonic prices refers to the regression for the marginal contribution of properties and the neighborhood relations. This model and methodology have been widely adopted in property research since Rosen's research in 1974. A model was constructed to assess the value of properties and conducting urban analysis by considering many variables. This was seen as the initial development of the definition of the hedonic pricing model. As time went on, more statistical research and applications for Rosen's model have come into being. Stevenson, for example, has reexamined the heteroskedasticity in the hedonic price model, which, to a greater extent, consolidates the theory and veracity of the model [9]. By adopting modern information technology like Geographic Information System (GIS), Bin [10] has used precise geographical data to verify the preciseness of the model, proving that the semiparametric regression model is practical for both analyzing and predicting the property price [11].

In recent decades, unlike the hedonic regression model, artificial neural network (ANN) and fuzzy logic (FL) methodology have also been widely accepted. In Din et al.'s [12] research on the ANN method for property appraisal, it generally performs well and generates acceptable performance in some respects. Still, it also turns out that different input choices of variables would sometimes generate statistically different values of output, which indicates the instability and immaturity of the ANN methodology. In terms of fuzzy logic, it is widely believed to be a more promising and generic approach for evaluating properties. Liu et al. [13] have constructed a statistical model based on the fuzzy neural network prediction model, which incorporates the hedonic theory and a great database with relevant characteristics affecting the price of properties based on recently sold projects. The experimental outcome and analysis have shown that the fuzzy neural network prediction model has a promising ability for real estate price prediction given reliable input data with high quality. A comparative experiment has also revealed that multiple regression applications for property appraisal work well with given data [14].

However, theoretically, the models based on multiple regression seem to attach more significance to statistical inference rather than prediction due to its nature. For this reason, extensive research on this theory has put the focus on the significant factor that influences the model the most, evaluating the economic value of real estate with specific characteristics and identifies the causal effect relationship behind the regressors [3]. Despite the profound and far-reaching significance brought by this method, its potential for statistical prediction on the pricing of real estate has been in deficiency by its nature. The limit of the traditional hedonic regression model has a significant influence on the procedure of generating the model, making it hard to identify the appropriate variables to set up the eventual model. Therefore, adopting other methodologies to conduct rigorous and viable research to a different extent is indispensable.

2. Spearman Correlation Coefficient

With the development of technology, high-dimensional data have been widely adopted in various fields, including economics, finance, and engineering. High-dimensional data are an indispensable ingredient when processing relevant data and conducting research. In addition, Big Data is also a newly emerged concept, indicating the following two features: enormous sample size and high dimensions of data.

Spearman correlation coefficient has been a nonparametric rank statistic. It was initially designed as a measure of the strength of the association between two variables. As a significant measure of monotone association, it is widely adopted when the input distribution makes the ordinary Pearson's correlation coefficient misleading. Not a measure of the linear relationship between two variables, the Spearman correlation coefficient evaluates the extent that an arbitrary monotonic function can depict the relationship between two variables on the condition that there is no further assumption made about the frequency distribution. One advantage of the Spearman coefficient is that it does not require the linear assumption as Pearson's coefficient does. This significant advantage has made it widely adopted by many statisticians in analyzes.

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The correlation coefficient is a measure to depict the association between variables, which can delineate the association for two variables co-occurring. The correlation coefficient has been widely used in the scientific research field. We will mainly discuss the Spearman correlation coefficient in this part.

We first assume two-dimensional random vectors:

$$(X_1, Y_1), \left(X_{2,} Y_2\right), (X_3, Y_3), \tag{1}$$

are identically independently distributed (i.i.d.). Its joint distribution function is defined as H, and marginal distribution functions are defined as G and H, respectively. Considering the covariance among those two-dimensional random vectors, the Spearman correlation coefficient can be defined as follows:

$$\begin{aligned}\rho_s = 3\big(&P\{(X_1 - X_2)(Y_1 - Y_3) > 0\} \\ &-P\{(X_1 - X_2)(Y_1 - Y_3) < 0\}\big).\end{aligned} \tag{2}$$

The outcome ρ will not be affected by substituting (X_2, Y_3) with (X_3, Y_2).

In another form for better calculation purposes, we can also offer a format for sampling for random samples:

$$\begin{aligned}&\{(X_i, Y_i), i = 1, \ldots, n\}, \\ &r = \frac{1/n\sum_{i=1}^{n}\left(RX_i - (n+1)/2\right)\left(RY_i - (n+1)/2\right)}{1/12\left(n^2 - 1\right)}.\end{aligned} \tag{3}$$

This equation can be viewed as a Euclidean distance for RX_i and RY_i. From a nonparametric point of view, it can be explained as follows: the smaller the distance is, the more likely X and Y will be positively correlated. The greater the distance is, the more likely X and Y are negatively correlated.

From the derivation above, it is apparent that the Spearman correlation coefficient is a rank correlation of coefficient. Therefore, it obtains some common merits of the rank correlation coefficient [15].

First of all, the Spearman correlation coefficient does not have a strict requirement for the sample normality as a nonparametric correlation coefficient. Thus, it has a more extensive adoption in statistical research.

Secondly, the Spearman correlation coefficient uses the rank of variables to compute the correlation between different variables. In this process, strict monotone transformation is employed so that the rank of variables will stay unchanged. Therefore, it has the merit of monotonousness. We can perform some transformation to generate various statistical models based on monotonousness, which extends this model to more variable filtering methods such as logistics model and box-cox transfer model.

Thirdly, the Spearman correlation coefficient can process discrete data practically. It will not be affected by dimension, which means it can precisely measure the correlation between different dimensions. Furthermore, the Spearman correlation coefficient is very unlikely to be affected by extreme values, which can minimize the negative impact of extreme values on our statistics.

Finally, the Spearman correlation coefficient can better depict some nonlinear relations. As pointed out by Lancaster, dependence or correlation measures indicate the strength of correlation, especially the mutually independent variables. Therefore, the correlation of variables does not necessarily mean the linear relationship between variables. It even cannot indicate a direct functional relation. In this sense, the Spearman correlation coefficient can better depict the correlation between variables when there exists nonlinearity [16].

3. Housing Price Prediction Based on Multiple Linear Regression

3.1. The Analysis of Main Factors Affecting Housing Price. Housing price is affected by multiple factors and features of a specific house. According to the previous research, some analysts have proposed several variables that significantly influence the overall housing price. According to Kusan et al. [9], these factors can be classified into three types: house factors, environmental factors, and transportation factors. Each factor and the effective primary mechanism are elaborated in the following text.

House factors can be divided into several types. The most influential type is residential factors, including residence, usability, and number of rooms. When people consider purchasing a house for living purposes, the factors above are the main determinants for the living quality. Buyers with family members would typically attach more importance to the essential feature of the house, like the living area and number of rooms, which have a significant impact on the overall living quality and experience in the house. Besides, the intangible features, like the view of residence and usability, also have a rather considerable influence on the housing price, through affecting buyers' experience on the house and willingness to pay.

The other influential types are the main factors related to building properties and floor factors. Building properties are mainly about hardware and basic facilities in the building, such as the elevator, generator, and garage. To depict an example of this, the number of containable cars within a garage is a rather important consideration. The rising trend of numbers of vehicles per household possessing generates a necessary demand for the quality and capacity of a garage in a house. Other affiliated facilities to the house like the swimming pool and backyard have also played an essential role in determining the housing price, as the demand for leisure and relaxation has been arising with the economic progress.

On the other hand, floor factors, like the number of floors, have also impacted the housing price significantly. Typically, household prefers the house with the number of the floors most suitable for their daily convenience. A family with children and elders tends to prefer a house with multifloor construction, which offers different family members separate living areas with appropriate privacy while living together.

Environmental factors mainly consist of two parts: regional environment and nearby pollution. Regional

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

environment refers to the overall living conditions in the surrounding community. Sanitation, as a significant indicator of the living quality, has been given more importance in the recent decades. A community with comprehensive sanitation services tends to attract buyers to pay a higher price. In addition, natural scenery, as an objective feature of the community where the specific house is in, also influences the housing price through various channels. Purchasers with a preference for a house with a mountain view or lake view may be willing to pay a higher price for a house near the natural scenery. Even for those buyers who do not have a specific preference, a decent and beautiful view would add more weight for the specific house and community when purchasers are choosing from various options.

On the other hand, nearby pollution is also an environmental factor negatively related to the quality of the house. The most apparent factors are noise and air pollution. Noise is generally affecting the community through various channels such as the nearby factories, cars running in the central lane, and pedestrians crossing the community. Air pollution, compared with noise, is a somewhat measurable and quantifiable indicator of environmental pollution. The general measurement for air pollution and air quality is AQI, which stands for air quality index. Typically, houses located in a community with better air quality tend to attract buyers with a higher willingness to pay, thus generating higher housing prices in the market.

Transportation, as the principal channel for connecting the community with the outer areas, is worth our specific discussion when analyzing the external factors influencing the housing price. Transportation can influence housing via various aspects, including the distance to social and cultural centers, distance to trade and shopping centers, and distance to public transportation stations.

The distance to social and cultural centers can be a meaningful consideration for many household buyers. As children need to go schooling and enhance cultural and physical education, facilities such as libraries, schools, and sports complexes are the frequently visited places for those buyers, who has blended into their daily lifestyle. The commuting time is positively related to the distance between the house and the destination. Closer distance offers greater convenience for all the household members, thus contributing to a higher price when choosing various options.

Following the same logic, we can reasonably derive that the distance to the local shopping centers and public transportation stations is also crucial to the housing price. Shopping is one of the most frequent activities for daily lives in the United States. Residents tend to drive to the nearest supermarkets for daily consumption like grocery and to the comprehensive mall for higher-level consumption like clothing or luxuries. Public transportation, on the other hand, has been significant for residents as well. Although driving is the most common way of commuting in the United States, other ways like taking an airplane or metro are also important substitutes for transportation choices.

3.2. Multiple Linear Regression Model for Housing Price Prediction. Analyzing data is for extracting accurate estimation from basic information provided. The most important and common question is whether there is a statistical relationship between an explanatory variable (usually denoted by X_i) and a response variable (usually denoted by Y). To solve this problem, a typical way is to apply regression analysis to model and quantify this statistical relationship. Many types of regression are adopted in scientific research, depending on the feature and type of given data.

The most used model for conduction regression is called the linear regression model, which is used when the distribution of the response variable Y is continuous and approximately regular. Linear regression is the procedure that estimates the coefficients of the linear equation, with at least one independent variable that best predicts the value of a dependent variable. Our goal is to predict the outcome Y based on the given values of predictor variables X_i. The linear regression model allows us to evaluate the impact of multiple variables in the same model.

An appropriate model could be a straight line in an actual application, a higher degree polynomial, a logarithmic, or exponential. We may find a proper model by the forwarding method, in which we start by assuming a relatively simple straight line $Y = a + bX$. Next, we may find the most suitable estimator of the assumed model. If the model does not fit the data well, we may alternatively assume a more complicated model, like a second-degree polynomial model $Y = a + bX + cX^2$. On the other hand, the other method is called the backward method, in which we assume a complicated model first, and we then fit the model, trying to simplify it. Both methods could achieve the same goal for modeling the data appropriately, depending on the given situation and features of the data set [17].

A multivariate linear regression model is based on the assumption [18]:

$$y_i = \beta_0 + B' x_i + \varepsilon_i, \tag{4}$$

where $y_i = (y_{i1} \ldots, y_{id}, \ldots y_{iD})'$ and x_i are the D-dimensional vector of the output variables and the P-dimensional vector of the fixed regressor values for the i^{th} sample unit correspondingly. β_0 is a D-dimensional vector containing the intercepts for the D response; B is a matrix of the dimension $P \times D$ whose (p, d) element, β_{pd}, is the regression coefficient of the p^{th} regressor on the d^{th} response; eventually, ε_i symbolizes the D-dimensional random vector of the error terms corresponding to the i^{th} observation. The classical multivariate linear regression model also assumes that $\varepsilon_i (i = 1, \ldots, I)$ is the independently and identically distributed random vector. Its distribution is assumed to be multivariate Gaussian with a D-dimensional vector with mean value zero and a positive definite covariance matrix $\sum$ of dimension $D \times D$, which is

$$\varepsilon_i \sim \text{MVN}\left(0, \sum\right). \tag{5}$$

The proposed multivariate linear regression model is based on the aforementioned assumption and

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\varepsilon_i \sim \sum_{k=1}^{K} \Pi_k \mathrm{MVN}\left(\nu_K, \sum K\right), \tag{6}$$

where π_k represents positive weights that sum to 1 and ν_k represents D-dimensional mean vectors that satisfy the constraint:

$$\sum_{k=1}^{K} \pi_k \nu_k = 0. \tag{7}$$

4. Simulation

4.1. Spearman Correlation Coefficient Analysis. According to the reference research and our previous discussion in Section 2, an analysis for correlation coefficients of different variables with the housing price is conducted. With the data processing methodology using Python, Spearman correlation coefficients can be simulated as shown in Figure 1 using the housing price data set in Boston.

The Spearman correlation coefficients between the house price and variables in the data set are shown above. These factors vary from house factors to environmental factors, all contributing to the formation of the overall price. This coefficient analysis reveals the general trend and significant factors on the housing price. To elaborate more about the exact definition of each variable, we can refer to the description in Table 1.

4.2. Multivariable Regression Analysis. As the methodology discussed above indicates, an empirical analysis based on the Boston housing price data set is conducted to test multiple factors and their impact on the median housing price as a response variable.

In the first place, data analysis is conducted on housing price, in the sense that the influence of the number of rooms on the overall housing price is analyzed, which can be seen in Figure 2. In Figure 2, the horizontal axis represents the average number of rooms per house, while the vertical axis represents the median price of self-owned houses in that region, measured in 1,000 US dollars. From Figure 1, there exists a positive, upward-sloping relationship between the number of rooms and overall housing price. With more rooms, the house is more likely to be a superior residence with higher quality and market value. This empirical evidence and trend have been a consistency of our common sense, which indicates a property will generally sell at higher prices when it has more rooms and space for living purposes. This empirical result has cross verified the features and trends that we have discussed in Section 3.

Next, we regressed the weighted average distance from the property to 5 employment centers in Boston, and we got the scatter plot as shown in Figure 3, in which the horizontal axis represents the weighted average distance from the house to 5 employment centers in Boston. From Figure 2, it can be summarized that the housing price is more concentrated in the middle and lower level when the distance ranges from 0 to 5 kilometers in the area close to the employment center. From a geographical perspective, this trend also shows that the sample has a higher density near the city center, in the sense that houses in Boston are more concentrated in a central location with an appropriate price. On the other hand, it can also be concluded that the house price distribution will be more sporadic as the distance increases, in the sense that the price variation between different houses may grow as the distance is getting further from the employment center in the city, and the price may fall into different ranges when the houses are in suburban regions.

From the empirical analysis based on Spearman correlation coefficient in using the previous data set for housing price in Boston, these factors including the proportion of lower-income group in the region, proportion of property land area larger than 25,000 square feet, the average number of rooms, etc., are among the primary factors that influence the housing price.

In this empirical analysis, with the data set of Boston housing prices, a multiple linear regression model is constructed to analyze different factors' impact on the housing price and predict the corresponding housing price based on the given input.

A typical standardization process in data analysis for the raw data is first conducted after obtaining the major influencing factors from previous analysis based on the Spearman correlation coefficient. By definition, the standardized value here will range between 0 and 1:

$$\text{standardized value} = \frac{(\text{characteristic value} - \text{minimum of characteristic value})}{(\text{maximum of characteristic value} - \text{minimum of characteristic value})}. \tag{8}$$

In the next step, as a tradition for regression analysis, the whole data set is divided into 2 parts: the training set and the testing set. A further comparison is made between the regression result based on the training set with the modeling result based on the test set. By this means, it could efficiently evaluate the accuracy and efficiency of the model empirically.

In the training process, a quantitative tool named gradient descent optimizer is optimized to calculate the parameters of the training model. The goal of this optimizer is to minimize the loss function in this analysis to find the optimum. The value of the loss function is obtained as shown in Figure 4. The vertical axis represents the loss value for each epoch, while the horizontal axis represents the epoch, which is the time of iteration.

In the next step, we train 100 epochs on the model parameters using the training data until the loss function converged, as the general methodology of multiple linear

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1: Description of each variable.

Variable	Description
Lstat	The proportion of lower-income groups in the region
Indus	Percentage of nonretail business in the region
Nox	Nitric oxide concentration (per 10 million)
tax	Full property tax rate per 10,000 US dollars
crim	The average crime rate in the town
ptratio	The student-teacher ratio in the town
age	The proportion of the self-built property before 1940
rad	Radial accessibility index to highway
chas	Charles River (a dummy variable)
zn	The proportion of property land area larger than 25,000 square feet
dis	The weighted average distance from the property to 5 employment centers in Boston
rm	The average number of rooms per house

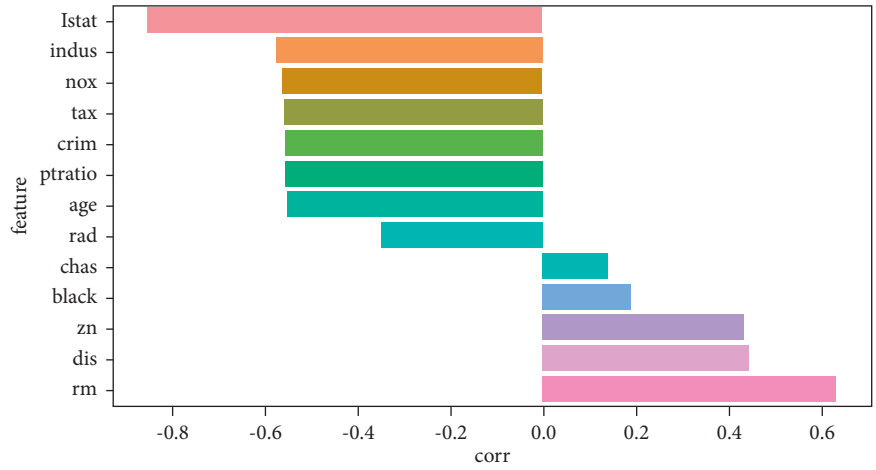

Figure 1: The correlation coefficients.

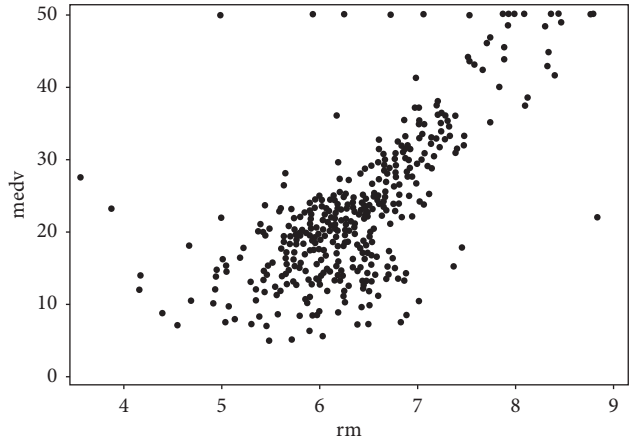

Figure 2: Housing price for the average number of rooms per house.

regression suggests. Based on this, the model parameters are used to forecast the housing price, and 100 samples are taken to obtain the prediction results, as shown in Figure 5.

As can be seen from Figure 5, the prediction results of the model are generally consistent with the trend of real value in the comparison set since the trend moves along the same direction with each other in most cases, and the overall deviation is generally acceptable in most cases. After training the model again for 500 epochs, the result is shown in Figure 6.

From Figure 6 result in the validation set, the prediction result is highly similar to that of Figure 5 in

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

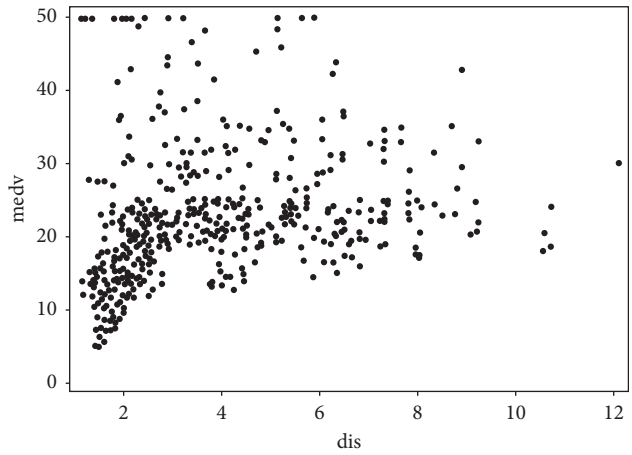

Figure 3: The house price distribution with different distances.

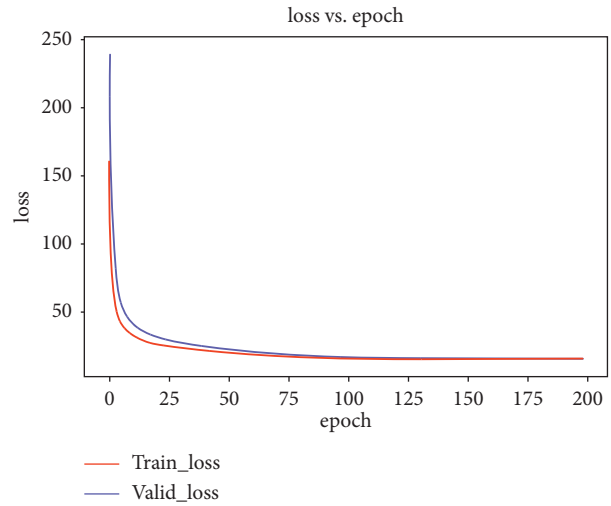

Figure 4: Loss function.

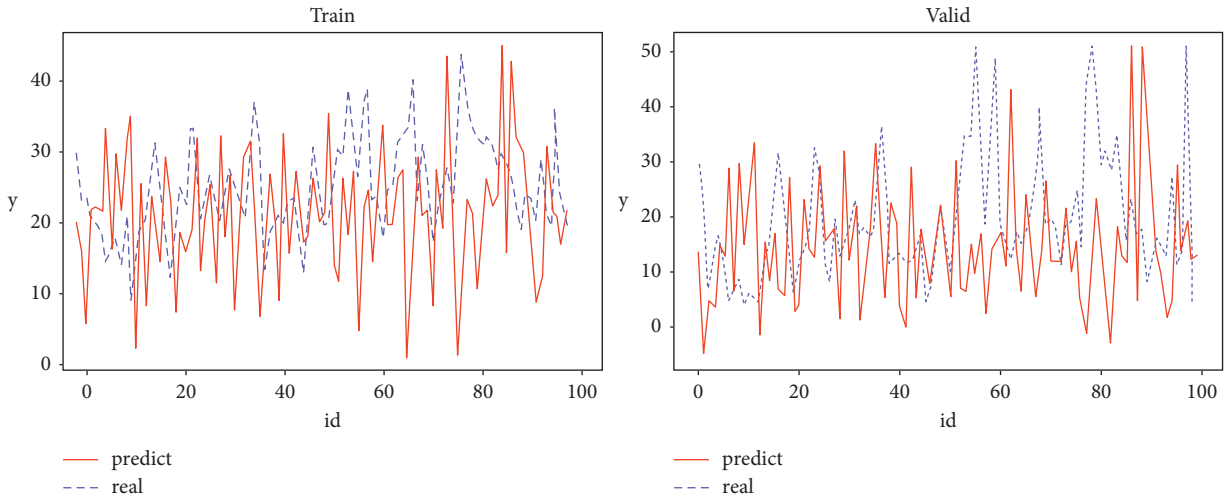

Figure 5: Prediction result.

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

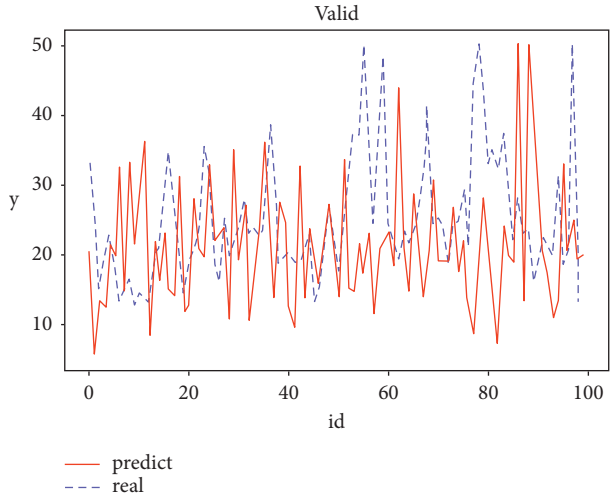

Figure 6: Prediction result.

general. This result can partially verify the accuracy and practicability of our empirical model constructed based on the Boston housing price data set. Therefore, with the analysis and discussion in this paper, it can be summarized that the multiple linear regression model can effectively predict and analyze the housing price to some extent. On the other hand, the prediction accuracy is still limited to some extent. In the further research, the application of machine learning algorithms and relevant methodologies in housing price prediction will be further optimized and widely researched.

5. Conclusion

The author constructs a fundamental algorithm based on the multiple linear regression method to predict housing prices and combines it with the Spearman correlation coefficient to determine the influential factors affecting housing prices. To train and test the parameters of this multiple linear regression model, the author applies the data set of the housing prices in Boston for model construction. From the simulation results shown above, it can be concluded that the proposed multiple linear regression model can effectively analyze and predict the housing price to some extent. Admittedly, the prediction accuracy is still limited at specific points, and the universality of the model still needs to be improved in further research. In further research into the corresponding models, the author will further study machine learning in the application of housing price prediction, as well as constructing a more robust algorithm based on a more advanced machine learning methodology.

Data Availability

The raw data sets used for this work are available upon request from the corresponding author.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

The author would like to express his deepest gratitude to his parents for supporting him in conducting the study and research. It is their spiritual and material support that contributes to the completion of this humble paper. The author would also like to extend his sincere gratitude to Dr. Owen HU for his assistance and guidance in the data processing part of this paper. His patient guidance enlightens the confidence and enthusiasm in conducting further research on this topic.

References

[1] S. Borde, A. Rane, G. Shende, and S. Shetty, "Real estate investment advising using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 3, p. 1821, 2017.

[2] B. Trawinski, Z. Telec, J. Krasnoborski et al., "Comparison of expert algorithms with machine learning models for real estate appraisal," in *Proceedings of the 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Gdynia, Poland, July 2017.

[3] V. Kontrimas and A. Verikas, "The mass appraisal of the real estate by computational intelligence," *Applied Soft Computing*, vol. 11, no. 1, pp. 443–448, 2011.

[4] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3–17, 2014.

[5] J. R. Barr, E. A. Ellis, A. Kassab, C. L. Redfearn, N. N. Srinivasan, and K. B. Voris, "Home price index: a machine learning methodology," *International Journal of Semantic Computing*, vol. 11, no. 1, pp. 111–133, 2017.

[6] W. J. McCluskey, M. McCord, P. T. Davis, M. Haran, and D. McIlhatton, "Prediction accuracy in mass appraisal: a comparison of modern approaches," *Journal of Property Research*, vol. 30, no. 4, pp. 239–265, 2013.

[7] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.

[8] E. Lughofer, B. Trawiński, K. Trawiński, O. Kempa, and T. Lasota, "On employing fuzzy modeling algorithms for the valuation of residential premises," *Information Sciences*, vol. 181, no. 23, pp. 5123–5142, 2011.

[9] H. Kusan, O. Aytekin, and I. Özdemir, "The use of fuzzy logic in predicting house selling price," *Expert Systems with Applications*, vol. 37, no. 3, pp. 1808–1813, 2010.

[10] O. Bin, "A prediction comparison of housing sales prices by parametric versus semi-parametric regressions," *Journal of Housing Economics*, vol. 13, no. 1, pp. 68–84, 2004.

[11] Y. Kang, F. Zhang, W. Peng et al., "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land Use Policy*, vol. 2020, Article ID 104919, 2020.

[12] A. Din, M. Hoesli, and A. Bender, "Environmental variables and real estate prices," *Urban Studies*, vol. 38, no. 11, pp. 1989–2000, 2001.

[13] J.-G. Liu, X.-L. Zhang, and W.-P. Wu, "Application of fuzzy neural network for real estate prediction," *Advances in Neural Networks - ISNN 2006*, vol. 3973, pp. 1187–1191, 2006.

[14] I. V. Lokshina, M. D. Hammerslag, and R. C. Insinga, "Applications of artifificial intelligence methods for real estate valuation and decision support," in *Proceedings of the In*

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hawaii international conference on business, Honolulu, Hawaii, USA, January 2003.

[15] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data," *QUAGEO*, vol. 30, no. 2, pp. 87–93, 2011.

[16] T. Oladunni and S. Sharma, "Hedonic housing theory - a machine learning investigation," in *Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, USA, December 2016.

[17] L. Petrella and V. Raponi, "Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress," *Journal of Multivariate Analysis*, vol. 173, pp. 70–84, 2019.

[18] G. Soffritti and G. Galimberti, "Multivariate linear regression with non-normal errors: a solution based on mixture models," *Statistics and Computing*, vol. 21, no. 4, pp. 523–536, 2010.

5192, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2021/7678931 by Bellarmine University, Wiley Online Library on [16/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License