Real Estate Price Prediction Based on Linear Regression and Machine Learning Scenarios

Tingjun Mao[1, *]

[1]DEPU Foreign Language School, ChongQing, China

*Corresponding author: mtj356685@icloud.com

Abstract. The changes in the housing market are not only related to human beings' daily life, but also have an important impact on the national economy. The prediction of housing price is one of the most widely concerned topics, which is linked to the formulation of national real estate policies and the analysis of the economic situation. In this context, this paper takes housing price prediction as the topic, selects the Eames housing price dataset in Iowa, and uses supervised multiple linear regression and machine learning algorithm to train and test the real estate price prediction model. Among them, there are 79 explanatory variables, which are related to housing attributes and the explanatory variable is housing price. 1460 data is included in the training set and 1459 in the test set. In the part of machine learning algorithm, PaddlePaddle deep learning framework is used in this paper to train and test the models with the help of AI Studio platform. The experimental results show that the scatter plots of the real values are clustered and distributed on both sides of the predicted line, and their direct differences are within 30 points. According to the analysis, the real estate price prediction model based on linear regression and machine learning is reliable and stable. This paper aims to provide some suggestions for the housing price prediction. These results shed light on guiding the reference direction for investors, so as to guide the formulation of relevant policies.

Keywords: Housing price prediction; multiple linear regression; machine learning

1. Introduction

Purchasing a house is one of the most important things in mankind's daily life, and housing prices are a key factor in determining whether people buy or not. Housing prices depends on a variety of factors, including location, size, and supply and demand in the real estate market. Changes in the housing market are also a crucial factor affecting the national economy. Therefore, housing price prediction can not only help buyers to find desirable and qualified houses, but also help professional economists to better analyze the current housing price situation. In this case, it helps government control housing prices quickly and reasonably and introduce real estate policies in a timely manner.

Contemporarily, there have also been a lot of studies on house price prediction. Kauko et al. studied neural network modeling through its application in the housing market of Helsinki, Finland [1]. Fan et al. proposed a variety of methods based on tree structure, which provided crucial tools for features selections [2]. Liu et al. proposed a neural network model based on hedonic price that is suitable for real estate price prediction [3]. Selim compared different scenarios and claimed that the artificial neural network model can be used as an improved choice for price prediction [4]. In another study, Kusan et al. proposed to predict the selling price of housing construction by using fuzzy logic model [5]. Azadeh et al. introduced an approach to solve the issue of predicting and optimizing housing market fluctuations [6]. Quang et al. concluded that the accuracy of RIPPER algorithm was always better than other models in housing price prediction by summarizing and studying the prediction methods mentioned by predecessors [7].

Apparently, house price prediction has always been a matter of great concern to people, and people have made great progress in house price prediction. This paper will provide a better multiple regression model for house price prediction by using a supervised how far linear regression model is used to predict real estate prices, and help buyers better choose reference. The rest of this paper is arranged as follows. The second section mainly analyzes and describes the data adopted and the research process; the third section presents the results obtained and explains their significance as well

as the design and analysis of the model; the fourth section describes the limitations of the research and the future development of the field; the last paragraph is a conclusion of the whole paper.

2. Data & Method

2.1 Data

The data analyzed in this study is the Eames housing price dataset [8] collected by Professor Dean DeCock and uploaded to Kaggle [9]. Mathematical model is used to make a reasonable prediction of the future housing price in this region. The dataset covers all recorded housing prices in Ames, Iowa, between January 2006 and July 2010. There are 79 explanatory variables, which cover almost every feature of a house; each house is described in detail with a variety of different attributes, including YearBuilt ("Original Construction Date"), YearRemodAdd ("Remodel Date") RoofStyle ("Type of roof") and several other similar attributes. There is also a target variable "SalePrice". The dataset also contains 1460 observations from the training set and 1459 observations from the test set.

2.2 Models and metrics

This paper adopts the selected Eames housing price dataset in Iowa, used supervised multi-distance linear regression and machine learning to train and test the real estate price model, so as to obtain a stable and reliable real estate price prediction model. Supervised learning is to use the samples of known categories to train and learn an optimal model. On this basis, it can achieve the required performance. Afterwards, the trained model is used to classify unknown data.

The first step in modeling is to examine and clean the data, train and test common collections of data engineering, and identify missing values. Further analysis of the federated database removes the number of missing values contained in each function. LotFrontage, FireplaceQual, Fence, Alley, and MiscFeatures had the most missing values [10]. Before calculating the various objects missing in a variable, it should be considered whether an object has been printed or a numeric object, and whether the value is intentionally or accidentally missing. Numerical objects missing from random values, such as "LotFrontage," are computed from homes in the same area via the Internet. Other numerical characteristics with mean values are mostly accompanied by certain rules, object types, some of which seem to contain missing values. However, N.A. values actually mean null value, e.g., a null value in PoolQC means that these variables do not have a pool attribute classified as "No_pool" [11].

The next step in estimating missing values is to look at the database to detect possible outliers. The batch quota function has an important function and these variables must be removed from the data set. Before designing features, they need to be analyzed in technical courses to better understand their diversity. Regardless of the relationship between functionality and standalone functionality, they all depend on the function of the SalePrice variable [12]. GarageCars are closely related to GarageArea, and GarageYearBuildilt is strongly related to Onaroly-2, most suitable functions can be obtained.

The new functionality was built on existing features, especially two-sided classification features This is a potential difference. It is worth noting that since this may mean that the second new feature is updated, the first is a two-dimensional classification feature, where the building date of the house is near the frame. In the same year, the second was digital function (YearRemodadd-Yearbuilt), which fixed the difference between the two, the first was y instead of 1980, whether they were built before 1960 or not, and whether they were also two new binary functions [13]. Later, Wasuilt was removed based on its loss conditions, its high relevance, and careful analysis of the new built-in features.

The object transform is the four objects used to correct the slope observed on their distribution map. After logarithmic transformation, the optimal distribution can be made, as shown in the following numbers that LotArea has a transformation. Another feature of the change is the "GroundLivingArea", whose distribution is free on the right and achieves a better distribution after log transformation, as shown in the following issues, in two other features, the square feet of the ground floor also underwent a logarithmic change, with the front plot being altered in shape [14].

After encoding each encoding object and removing the encoding variables, 201 columns were added to the data structure for a total of 288 functions. The joint data field is divided into training data field and test data field. The training collection includes 1,460 reviews and 1,459 test views. Item-based regressions are the first models Ridge predicts, and they use multiple linear regressions as they can adjust the diversity between features and reduce the prediction error [15]. Ridge regression has been tested using the number of folders required for cross-sectional accuracy. As the numbers show, the standard deviation increases as the number of folders increases, while the average error remains the same when using 10 folders:Graphs of ridge regression coefficients and regularization/ intensity line. Alpha = 10, Toll = 1E-05, and solvent = SVD were assumed to obtain a better true cross section score of 0.1132. Lasso regression-- Like multiple linear regression, the best estimate of cross-sectional accuracy of Lazo regression based on ultrafine optimization is 0.1147, alpha = 10, and max_iter = 25. The Elatic-Net model is a conditioning model that combines degradation with strength. As in the other models, α increases, decreases to 0. According to the graphical representation of the object, the optimal parameters of poisoning were l1_ratio = 0.001, alpha = 0.1, and the cross-sectional accuracy index was 0.11204. The tree models (training three tree models) are called Support Vector Machine (SVM), gradient spread regression function, and XGBoost. The tree model requires three major adjustments to show the optimal optimization Settings for the model, namely Gamma, Epsilon, and C， where C has the greatest impact. These are optimization options, e.g., epsylon helps define boundaries for allowable violations and substitution thresholds [16]. The best parameters of the network search regression are gamma=.000001, C=100 and Epsilon =0, which can also be observed as planned. One-dimensional linear regression is a major factor because the independent variable explaining the change in the dependent variable usually affects several important factors when studying real problems. At this point, one needs to use two or more influence factors as explanatory variables to account for changes in the dependent variable (also known as multiple regression) [17].

If y is the dependent variable and is independent of the variable, and the ratio of the independent variable to the dependent variable is linear. Assuming the term is a constant regression in the correction, then the effect of each additional unit on y is a partial regression y. The same principle is stable, and the effect of each unit on y is additional [18]. If two independent variables have a linear relationship, then the model can be described as follows:

$$y = b_0 + \sum_{i=1}^{k} b_i x_i + \epsilon \tag{1}$$

The parameters of the multiple regression model should be minimized based on the squared number of errors, and the least square method was used to solve the parameter as:

$$\begin{cases} \sum y = nb_0 + b_1\sum x_1 + b_2\sum x_2 \\ \sum x_1 y = b_0\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1 x_2 \\ \sum x_2 y = b_0\sum x_2 + b_1\sum x_1 x_2 + b_2\sum x_{2,}^2 \end{cases} \tag{2}$$

$$b == (x'x)^{-1} \cdot (x'y) \tag{3}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \end{bmatrix} \tag{4}$$

2.3 Metrics

Some regression models, such as unformed linear regression models, must perform the necessary testing and estimation after estimating the square of the minimum parameter. Therefore, there are coordinates that define certain linear regressions in the case that do not substitute for linear regressions. These linear regressions are common to the dependent variable. In the relationship between regression equations, the greater the degree of regression of the data point corresponding to each sample, and the closer it is to all independent variables (that is, the ratio of variables).

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \tag{5}$$

$$\sum(y-\hat{y})^2=\sum y^2-(b_0\sum y+b_1\sum x_1y+b_2\sum x_2y+\cdots+b_k\sum x_ky),\sum(y-\bar{y})^2=\sum y^2-\frac{1}{n}(\sum y)^2 \quad (6)$$

Standard errors are calculated as

$$S_y=\sqrt{\frac{\sum(y-\hat{y})^2}{n-k-1}},v_k=\frac{S_y}{y} \quad (7)$$

Here, k is the sum of explanatory variables in the multiple linear regression equation. The meaning of regression equation is to correct the value of the entire regression, or to explain whether the variables are closely related to the linear relationship between the dependent variables. F test can usually be used, and the formula of it is as follows [19]:

$$\mathrm{F}=\frac{\sum(\hat{y}-\bar{y})^2/k}{\frac{\sum(y-\hat{y})^2}{n}-k-1}=\frac{R^2/k}{\frac{(1-R^2)}{n}-k-1} \quad (8)$$

One dimensional linear regression (F-TEST) tests show linearity, but in multiple linear regression, this equation is incorrect. Each regression of the regression model should be tested separately for significant significance. Then, the degrees of freedom of the distribution table are viewed according to the value of A, as well as the critical value or value of the b_i regression, which varies widely from 0 to vice versa, with no significant difference of 0.

$$t_i=\frac{b_i}{s_y\sqrt{C_{ij}}}=\frac{b_i}{s_{bi}} \quad (9)$$

It is the jth element of the inverse matrix of bilinear regression in the celestial regression equation. The following formula can be calculated: $C_{ij}(x'x)$

$$C_{11}=\frac{S_{22}}{S_{11}S_{22}-S_{12}^2}, \quad (10)$$

$$C_{22}=\frac{S_{11}}{S_{11}S_{22}-S_{12}^2} \quad (11)$$

$$S_{11}=\sum x_1^2-\frac{1}{n}(\sum x_1)^2,S_{22}=\sum x_2^2-\frac{1}{n}(\sum x_2)^2,S_{12}=s_{21} \quad (12)$$

If the regression test fails, the matched parameters may not have a significant impact. The parameters should be excluded from the regression model; simpler regression models should be augmented or replaced with parameters. This may also be due to generalization of independent variables, where efforts should be made to reduce the influence of Collins [20]. Multiplication refers to the strong linear relationship between the independent variables in the case of multiple linear regression, so the estimation of the stable regression model and the official regression is inaccurate. As long as the variety is not too strict, whether there is a large multiplication in the multiple linear regression equation, then whether the solution between each variable between two independent variables can be calculated separately. It is also possible to calculate the number of independent variables (maximum eigenvalue, minimum eigenvalue) <100in the correlation matrix, and then collect without multilateral emphasis. If $100\leq \mathrm{k}\leq 1000$ independent variables have strong multiplicity, while $\mathrm{k}>1000$ indicating strong multiplication among independent variables. Absolute numbers are converted to logarithms or semantic numbers, or other explanatory variables are replaced. If there is a close relationship between the errors in the series, then there is no relationship between the error points, and a certain regression model cannot express the ratio of the actual variability to the dependent variable. The D.W. test is a series of errors, and the test method is similar to a full linear regression.

2.4 Procedure

In this paper, supervised multi-distance linear regression and machine learning are used to train and test real estate price models. Firstly, the Eames housing price dataset in Iowa was analyzed and preprocessed. The independent variables and dependent variables were analyzed, and the data were

transformed into tensors to prepare for the training model. Secondly, the model of this paper is designed and the model suitable for this paper is defined. Finally, the model is trained with the data, and the trained model is tested with the test set data to verify its reliability [21]. The flow chart is shown in Fig. 1.

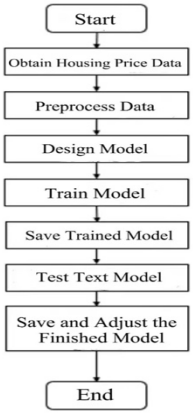

Fig. 1 Flow chart of this study

Supervised multi-distance linear regression models mainly include independent variable and dependent variable subject function, loss function, optimization function and verification function. The model structure diagram is shown in Fig. 2. Subject function mainly defines the relationship between independent variables and dependent variables. In this linear review model, the function is assumed to be $Y' = wx + b$, where Y' represents the predicted housing price of the model, and Y represents the parameters to be learned by the real housing price model, namely w and b. Loss function: A mathematical method is used to measure the error between the prediction result of a hypothetical function and the true value. Here, MSE (mean square error) function is used as the loss function. If N data are divided into i groups, the sample correction variance of group s_1^2 is:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{r}(n_i - 1)s_i^2}{N - r} \tag{13}$$

In the training process, the value of the loss function is gradually converged to obtain a set of weights that make the neural network fit the real model. Therefore, the ultimate goal of optimization function is to find the minimum value of loss function. The search is to fine-tune the values of the variables w and b to try out this minimum step by step. Adam function is adopted as the optimization function in this paper [22]. Adam optimization algorithm basically combines Momentum and RMSprop, and the formula is as follows:

$$V_{dw} = 0,\ S_{dw} = 0,\ V_{db} = 0,\ S_{db} = 0 \tag{14}$$

In the t iteration, the mini-batch gradient descent method is used to calculate dw and db. The exponential weighted average of Momentum is calculated and updated with RMSprop. The correction bias for Momentum and RMSprop is calculated and the weights are finally updated. The original data are mainly grouped, part of which is used as training set and part of which is used as validation set. In this paper, the K-fold cross validation function is adopted as the validation function.

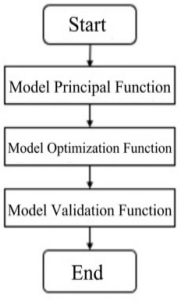

Fig. 2 Structure diagram of model in the study

3. Results & Discussion

3.1 Regression Analysis

The price data is first analyzed to draw price histograms. Seen from Fig. 3, the price data show a right-skewed normal distribution. 50% of the houses were clustered around $160,000, with an average price of $180,000. The minimum value of the house price is greater than 0, and the standard deviation is within the acceptable range, which means that the price data is available.

House area and housing decoration quality are selected to be analyzed with the independent variable house price. As shown in Fig. 4, the correlation between house area and house price. It can be seen from the figure that there is a close relationship between house area and house price, and the relationship is approximately linear. As illustrtaed in Fig. 5, decoration quality is closely related to housing price [23]. The data of level 1 and level 2 are too few, resulting in incomplete box chart. The house price of the other eight decoration quality is approximately normal distribution, and the influence of different decoration quality on the house price is more significant [24].

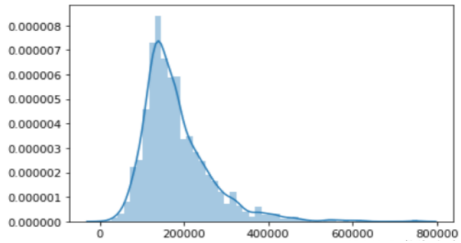

Fig. 3 Normal distribution of prices

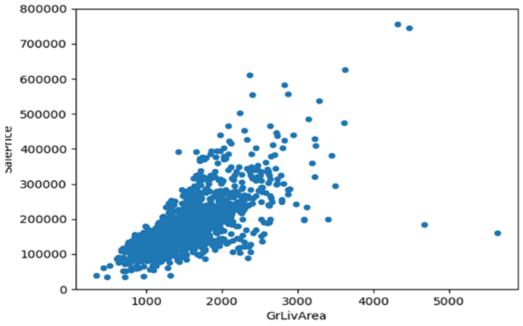

Fig. 4 Relationship between house area and price

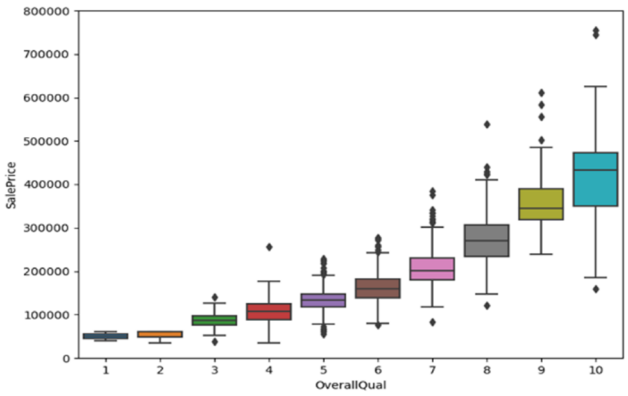

Fig. 5 Relationship between decoration quality and housing price

It can be observed simply and intuitively that the two dependent variables have obvious linear correlation with the house price. However, there are more than 70 dependent variables in the data set adopted in this paper, and the selection of simple two dependent variables is subjective and not representative. Therefore, the correlation matrix is used for further analysis, which can comprehensively reflect the correlation degree between any two column variables in the training set. Fig. 6 shows the correlation matrix of independent variables and dependent variables [25]. Apparently, different independent variables have different correlations with housing prices. From the analysis of

the whole proof chart, it can be seen that there are few blue independent variables, and most of them are dark red and red. This proves that many dependent variables have a strong correlation with the price of housing, and they affect the price trend of housing. Moreover, as the blue dependent variable may have abnormal values, outliers and their missing values should be processed and optimized before training. Thus, more accurate relationship between dependent variable and independent variable can be obtained, which makes the trained housing price prediction model more accurate.

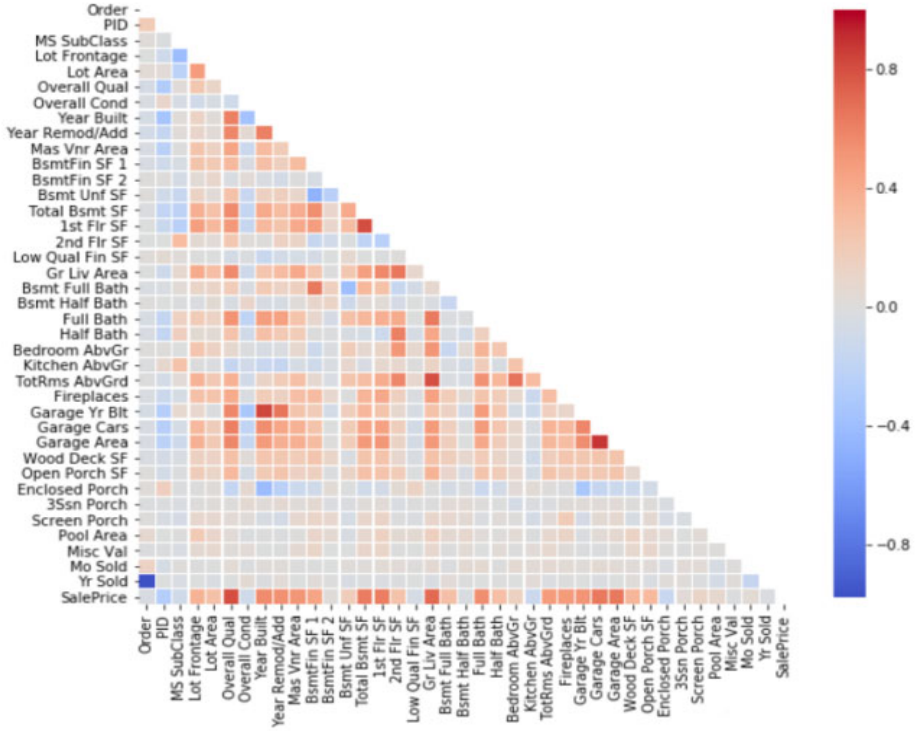

Fig. 6 Matrix diagram of related dependent variables and housing prices

3.2 Prediction results of the model

In this paper, PaddlePaddle deep learning framework is adopted to train and test the model on the AI Studio platform. After importing the data set and configuring the model, the model is instantiated and trained. Finally, the trained model is tested using the data of the test set. In order to observe the reliability of the model in this paper more conveniently, the real values of housing prices and thecomparison with the test are plotted, as shown in Fig. 7. The real value of the prediction results of the model in this paper is not linear. The line drawn by the predicted value crosses the scatter distributed by the true value. The scatter plot of the true value is clustered on both sides of the predicted line, whose direct gap is within 30 points into the clustered distribution. Intuitively, the real estate price prediction model based on linear regression and machine learning is reliable and stable.

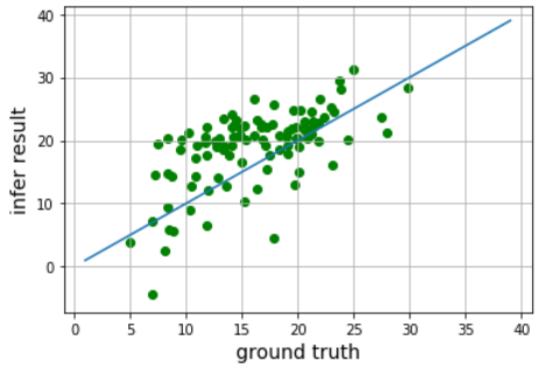

Fig. 7 Comparison between real value and predicted value.

3.3 Interpretation and significance of model results

Based on supervised multi-distance linear regression model for real estate price prediction, the data set selected in this paper is confirmed to be feasible [26]. Furthermore, the correlation is further confirmed by analyzing the relationship between them through matrix diagram, the abnormal data and missing data are optimized, and the model was trained and tested using the data [27]. The experimental results show that the predicted house price is correlated with the real house price, and the scatter plot of the real house price is clustered and distributed on both sides of the predicted house price. It further proves the reliability of the model in this paper, and also illustrates the significance of the study. These results can provide a better multiple regression model for the prediction of housing prices and give buyers a better reference [28].

4. Limitation and Prospects

This paper trains the real estate price prediction model based on linear regression and machine learning, and verifies the reliability of the model, which can provide a better prediction model for housing price prediction. However, the limitations of this study can still be clarified. Firstly, due to the lack of capacity, the research in this paper fails to adopt more extensive prediction mechanisms such as deep learning and these emerging science and technology can better guide the prediction of the model. Secondly, the open data set used in this paper is the Eames housing price data set, which has certain limitations and cannot represent the trend of the whole real estate industry. Thirdly, the social and natural factors can have an impact on the housing price every year, which makes the model not always applicable to the housing price prediction of all cities [29].

In the future, advanced deep learning models should be used to train and test the housing price prediction model. Moreover, a broader set of data should also be used to predict house prices in different cities. Furthermore, data should be updated dynamically, and the latest data sets should be adopted to collect big data to achieve better real-time dynamic housing price prediction.

5. Conclusion

In conclusion, this study investigates a better prediction model for housing price prediction based on linear regression and machine learning, and verifies the reliability of the model. To be specific, this paper selects Eames housing price dataset in Iowa and uses supervised multiple linear regression and machine learning algorithms to predict real estate prices. Primarily, the data is transformed into a tensor, after which the price data is analyzed by histograms to confirm the availability of the data. In order to further analyze the relationship between independent variables and dependent variables, this paper uses matrix graph to conduct correlation analysis, and optimize the abnormal data and missing data. Finally, the model is trained and tested. The experimental results show that the real housing price scatterplots are clustered and distributed on both sides of the predicted housing price, which further proves the reliability of the proposed model. However, this paper also has certain limitations, e.g., other algorithms are not used for comparison, and the universality of the model is not further verified. In the future, more advanced deep learning models and more extensive and dynamically updated data sets should be adopted for research, in order to achieve real-time stable prediction of housing prices. The main significance of this study is to provide suggestions for housing price prediction and provide better reference for investors to choose. Overall, these results offer a guideline for provide a better multiple regression model for the prediction of housing prices and give buyers a better reference.

References

[1] Kauko T et al. Capturing housing market segmentation: An alternative approach based on neural network modeling. Housing Studies, 2002, 17(6): 875–894.

[2] Fan G et al. Determinants of house price. A decision tree approach. Urban Studies, 2006, 43(12): 2301–2315.

[3] Liu J et al. Application of fuzzy neural network for real estate prediction. LNCS, 2006, 3973: 1187–1191.

[4] Selim H. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications, 2009, 36(2): 2843–2852.

[5] Kusan H. et al. The use of fuzzy logic in predicting house selling price. Expert Systems with Applications. 2010, 37(3): 1808–1813.

[6] Azadeh A. et al. A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. Expert Systems with Applications, 2012, 39(1): 298–315.

[7] Cock D. D. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project" Journal of Statistics Education. 2011, 19(3): 11-13.

[8] Cock D. D. House Prices - Advanced Regression Techniques. Retrieved from: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

[9] Truong Q., et al. Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 2020, 174: 433-442.

[10] Zauhar R., et al. As in Real Estate, Location Matters: Cellular Expression of Complement Varies Between Macular and Peripheral Regions of the Retina and Supporting Tissue. Front Immunol, 2020, 13: 519.

[11] Moro M. F., et al. COVID-19 pandemic accelerates the perception of digital transformation on real estate websites. Qual Quant, 2022, 13: 1-17.

[12] Soundararaj B. et al. Using Real-Time Dashboards to Monitor the Impact of Disruptive Events on Real Estate Market. Case of COVID-19 Pandemic in Australia. Comput Urban Sci, 2022, 2(1): 14.

[13] Medlock A. E., et al. Prime Real Estate: Metals, Cofactors and MICOS. Front Cell Dev Biol, 2022, 10(12): 89.

[14] Lee C. C. et al. The Effects of Leader Emotional Intelligence, Leadership Styles, Organizational Commitment, and Trust on Job Performance in the Real Estate Brokerage Industry. Front Psychol, 2022, 13: 88.

[15] Cohen J. P. et al. The impact of the Coronavirus pandemic on New York City real estate: First evidence. Reg Sci, 2022, 62(3): 858-888.

[16] Bao W., et al. Real Estate Prices, Inflation, and Health Outcomes: Evidence from Developed Economies. Front Public Health, 2022, 10(8): 51.

[17] Guenego A., and Fahed, R. Stroke Prognostication Obeys the Same Rules as Real Estate: Location, Location, Location!. Neurology, 2022, 98(11), 429-430.

[18] Gong W., and Kong, Y. Nonlinear Influence of Chinese Real Estate Development on Environmental Pollution: New Evidence from Spatial Econometric Model. Int J Environ Res Public Health, 2022, 19(1).

[19] Bachmann M., et al. The Increasing Investment of Real Estate in the Health System-A Comparison between the USA and Europe. Healthcare (Basel), 2021, 9(12): 122.

[20] Wang Zhou, et al. Effect of Regret Aversion and Information Cascade on Investment Decisions in the Real Estate Sector: The Mediating Role of Risk Perception and the Moderating Effect of Financial Literacy. Front Psychol, 2021, 12(7): 36.

[21] Balemi N et al. COVID-19's impact on real estate markets: review and outlook. Financ Mark Portf Mang, 2021, 35(4): 495-513.

[22] Pujals M., et al. HMGA1, Moonlighting Protein Function, and Cellular Real Estate: Location, Location, Location! Biomolecules, 2021, 11(9): 21-23.

[23] Wang, C., et al. Does real estate bubble affect corporate innovation? Evidence from China. PLoS One, 2021, 16(9): 25.

[24] Steegmans J., and de Bruin, J. Online housing search dataset: Information flows of real estate platform users. Data Brief, 2021, 38: 10.

[25] Lesame K., et al. On the Dynamics of International Real-Estate-Investment Trust-Propagation Mechanisms: Evidence from Time-Varying Return and Volatility Connectedness Measures. Entropy (Basel), 2021, 23(8): 11-13.

[26] Grybauskas A. et al. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. J Big Data, 2021, 8(1): 105.

[27] Sellwood M., et al. What biomedical education might learn from real estate tours. Biochem Mol Biol Educ, 2021, 49(5): 681-682.

[28] Paul T. K., et al. Multi-attribute decision making method using advanced Pythagorean fuzzy weighted geometric operator and their applications for real estate company selection. Heliyon, 2021, 7(6): 340.

[29] Gauger F., et al. Linking real estate data with entrepreneurial ecosystems: Coworking spaces, funding and founding activity of start-ups. Data Brief, 2021, 37(10): 71.