Analysis of the Effectiveness of Predicting Housing Prices Based on Different Machine Learning Models

Wen Wang

School of Accounting, Wuhan Textile University, Wuhan City, Hubei Province, 430200, China
2116250220@mail.wtu.edu.cn

Abstract. With the growing housing price market, effectively predicting housing prices not only has an important impact on the economy but also is related to people's living standards. However, the fluctuation of housing prices is affected by many factors, and in most cases, there is a non-linear relationship between housing price fluctuations and housing factors. In reality, there are differences in the effectiveness of different machine learning models in predicting home prices. Therefore, this paper uses multiple machine-learning models to explore how effective different machine-learning models are for house price prediction. In this work, the authors searched the Kaggle website for a data set of housing prices and housing factors in Bangalore, India. The housing attribute data includes the housing price, number of hardware facilities (number of bedrooms, number of swimming pools, number of sofas, etc.), and number of service facilities around the house (stadiums, shopping malls, etc.). Then, this data set was used to evaluate the house price prediction method of random forest, ridge regression, and XGboost. The results showed that the mixed model showed the best fit. This study can be used to select appropriate machine learning model predictions for home prices

Keywords: Predicting Housing Prices, Machine Learning Models, Third Keyword.

1 Introduction

Housing price affects the stability of the social and economic system and plays an important role in the sustainable development of the whole macro society. Rising housing prices can stimulate economic development, increase residents' purchasing power, and thus promote consumption. Too high housing prices may make many families unable to afford the cost of buying houses, reducing the consumption level to save money to buy a house, limiting other types of consumption. With the further development of big data and artificial intelligence, the good advantages of machine learning in non-linear modeling make it more prominent in systemic financial risk warning [1]. Therefore, it becomes more and more common to use machine learning models to predict housing prices [2]. Most of the current housing price prediction models are single models. In actual research, it is found that a single model has disadvantages such as low accuracy, poor generalization ability, and easy overfitting. Giv-

B. Siuta-Tokarska et al. (eds.), *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, Advances in Economics, Business and Management Research 300,
https://doi.org/10.2991/978-94-6463-542-3_50

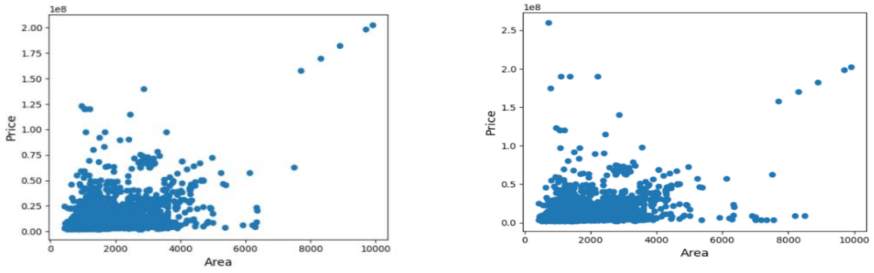
en these problems, the hybrid model based on feature selection and integrated learning can further improve the accuracy of house price prediction. The stacking ensemble learning model can take advantage of each different base learner to avoid the problem of poor generalization of the unitary learner and thus improve the prediction accuracy [3, 4]. However, at present, the academic field of using superposition integration model to predict housing prices is not extensive, and some scholars use this model for clinical intelligent diagnosis and fitting of OK lens, The experimental results also prove the maximum advantages of stacking in each machine model [5]. This paper explores the house price prediction effects of different machine learning models through Bangalore, India as a case study. Although the current study selects a variety of machine learning models, only one indicator (RMSE) measures different models to compare the prediction error.

This paper aims to improve the accuracy and stability of house price prediction, to contribute to the development and stability of the real estate market.

2 Methodology

2.1 Data Sources

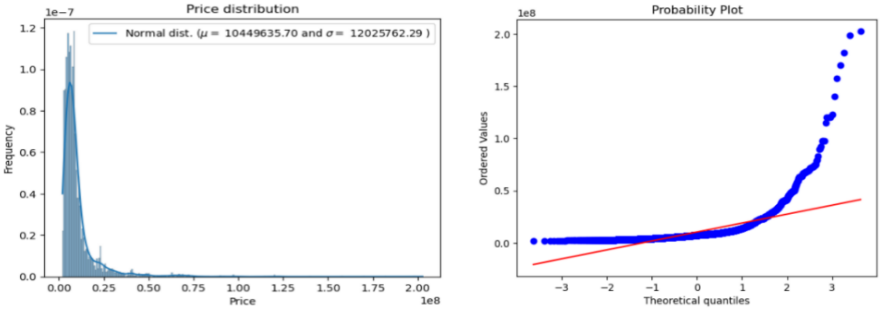
Using the Bangalore data as a dataset for model training and testing, the statistical dataset size was found as (6207,40). Of these, 6,207 training samples were used for model construction, and 40 test samples were used to evaluate the model. The initial characteristic dimension is 39 dimensions, which include the numerical characteristics of the area, living rooms, resales, maintenance, furniture, swimming pools, and other numerical characteristics. The initial data has no missing data at the time of testing, but there is still some abnormal data, and the data distribution does not conform to the normal distribution. Therefore, in the data preprocessing phase, the outliers of the target variable (price) are removed to be more consistent with the normal distribution. Fig. 1, (a) shows the raw data of the correlation between housing price and area in the original Bangalore housing price data set, and Fig. 1, (b) shows the processing results after feature engineering.



(a) (b)

Fig. 1. Distribution map (a) of the target variables; (b) removing the target variable outliers (Photo credit: Original)

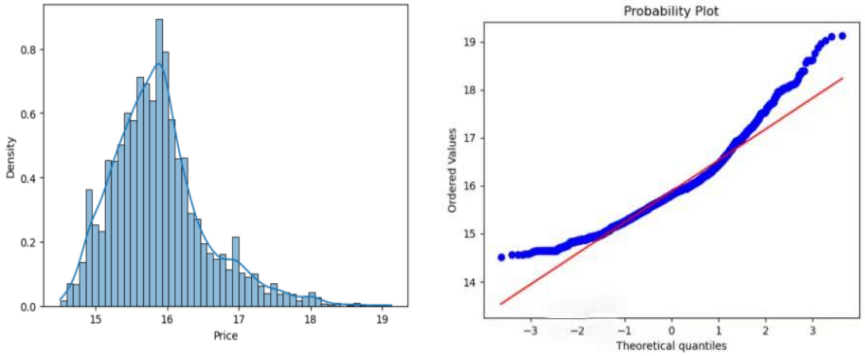
Next, use the Quantile-Quantile Plot and the frequency number histogram for the normality test for the target variables (Fig. 2).



(a) (b)

Fig. 2. Normality test (a) Price Distribution; (b)Probability Plot (Photo credit: Original)

It was found that the objective function did not conform to the normal distribution, so the $log\ (1\ +\ x)$ transformation was used to achieve a normal distribution(Figure 3).



(a) (b)

Fig. 3. Plot with a log(1+x), (a)Convert price distribution; (b)Convert probability (Photo credit: Original)

Further analysis of the correlation of the data using the heat map found that many characteristic variables were completely correlated with the target variables, and the problem of multiple contributions may appear in the subsequent prediction. Therefore, the accuracy and authenticity of the data set are in doubt. For the data of feature variables, normal distribution analysis used boxplots, of which 39 had up to 38 variables with non-normal distribution, indicating that this group of data should be treated using a non-normal distributed model. Then the log function was used, and more variables were still not distributed, which again showed that the data of this data set was extremely poor.

2.2 Data Methods

Random Forest.

The random forest belongs to the Bagging framework, which obtains different subsets by bootstrap sampling the data sets, and using these subsets to train base learners separately, belonging to a parallelized ensemble learning method. Random forest adds the process of feature selection to the Bagging framework, which makes the random forest have both sample and feature randomness, namely row sampling and column sampling [6].

Ridge Regression.

The random forest belongs to the Bagging framework, which obtains different subsets by bootstrap sampling the data sets, and using these subsets to train base learners separately, belonging to a parallelized ensemble learning method. Random forest adds the process of feature selection to the Bagging framework, which makes the random forest have both sample and feature randomness, namely row sampling and column sampling [6].

Extreme Gradient Boost (XGboost).

XGboost is a supervised machine learning algorithm, also known as Extreme Gradient Ascension. It belongs to the boosting branch of ensemble learning in the field of machine learning. XGboost is based on Gradient Boosting Decision Tree (GBDT), which belongs to the same branch and has been improved and optimized [7, 8].

The GB in XGboost refers to Gradient Boosting, which is a type of boosting method. Each of its base models is established to reduce the residuals of the previously established models. To achieve this goal, new models are usually established in the gradient direction where the residuals are reduced, and the trained new models are combined with the existing models cumulatively [9]. However, although the XGBoost algorithm can measure the feature importance, it cannot accurately judge the relationship between each feature and the final prediction results [10].

3 Empirical Analysis

3.1 Fitting Effects of a Single Decision Tree Model and a Random Forest Model as Measured by Expected Value

First, four variables, the number of bedrooms, area, the number of gymnasiums, and the number of pools, were selected as characteristic variables, and the sklearn library and expected value were used to test their cross-validation scores. The results showed that the explained variance was only 0.08, indicating that the effect of the single decision tree model was relatively general. Generally, the closer the expected value is to 1, the better the model is.

After that, the authors considered the random forest model for fitting. Ranking of the importance of feature variables by mutual information regression. According to

the importance order of the feature variables ranked by mutual information, the feature variables in the top ten are first tested and found that with the increase of tree nodes, its prediction effect tends to be about 0.3. It can think that the EV of the random forest model with 10 features is around 0.3. After testing the feature variables in the top 20, their EV was stable at around 0.35 and the degree of change was very low, indicating that the prediction effect of the random forest model was also general, so other models or methods were considered.

3.2 Fitting Effects with the Stacked Mixture Model

To reduce the risk of overfitting, machine learning models such as ridge, random forest, and gradient boosting were used and stacked with XGboost. After redistinguishing the test set and the validation set, the cross-validation and loss functions are set, and the use of various models (the smaller the range, the better the fit), and finally stacking to form a mixed model. The test results showed a root mean squared error of 0.514 for XGboost, 0.568 for ridge regression, 0.516 for random forest, and 0.466 for the mixed model.

4 Conclusion

This paper selects the data of housing prices and related factors in Bangalore, India, and uses the machine learning model method of mixed model (random forest, XGBoost, ridge regression, gradient lifting algorithm, stacked five model mixture) and a single model such as single decision tree model and random forest model. The results showed that the final predictive value of the mixed model was significantly higher than that of a single model. The results show that the combination of various models of machine learning can effectively pool its advantages and improve the accuracy of machine learning predictions. For the prediction of this group of data, the prediction effect of a single model is not as good as the fit of the mixed model. The mixture of various machine models is beneficial to overcome the shortcomings of a single model and make the prediction of experimental results more accurate.

References

1. Jiagen, X., Jing, L., Peiwen X.: Study on the impact of housing price and stock price fluctuation on macroeconomic stability. East China Economic Management, 32(03): 5-13 (2018).
2. Hongquan, L., Liang, Z.: Systemic financial risk monitoring and early warning based on machine learning technology. Operations Research and Management, 32(11): 212-219 (2023).
3. Yin, F., Du, J., Xu, X., et al.: Depression detection in speech using transformer and parallel convolutional neural networks. Electronics, 12(2), 328 (2023).
4. Chung, D., Yun, J., Lee, J., et al.: Predictive model of employee attrition based on stacking ensemble learning. Expert Systems with Applications, 215, 119364 (2023).

5. Jiaming, G., Kangmei, L., Jun, H., et al.: Stacking Integrated learning is applied to the clinical fitting of orthokeratology lens in myopia correction (English). Journal of Donghua University (English Edition): 1-11 (2024).
6. Liangjin, Z., Mingyang, Z.: Price analysis of second-hand houses in Shenzhen based on random forest. The China market, (26): 68-71+133 (2022).
7. Shuxiang, C., Shiqi, X., Jun, L., et al.: Cost prediction model of expressway bridge project based on ridge regression optimization algorithm. Building economy, 44(S2): 225-229 (2023).
8. Jiawen, Z., Yanling, X., Zheng, L., et al.: A regression estimation model based on wheat yield. Journal of Wheat Crop Sciences: 1-10 (2024).
9. Rong F.: A Study on batch evaluation of second hand housing prices based on XGBoost model. Central South University of Economics and Law (2022).
10. Menghua, D., Tianshu, Z., Junfei, C.: Study on the influencing factors of ecological compensation payment willingness based on the XGBoost-SHAP model. Water Conservancy Economy, 42 (02): 44-50 (2024).