Applying Machine Learning and Time Series to Predict Real Estate Valuations

Sijian Zhao*

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Wuhua District, Kunming City, Yunnan Province,650221, China

*Corresponding author.E-mail:Zhao_Si-Jian@hotmail.com

Abstract. This paper focuses on the combined application of time series analysis models and machine learning models in real estate valuation forecasting. When dealing with real estate valuation data, we use deep learning models to extract key features to achieve high accuracy prediction. The performance of four machine learning models (multiple linear regression, random forest, gradient boosting, and support vector machine) is also compared, and the results show that the random forest model has the best performance on the test set, and it possesses high prediction performance. Finally, the prediction accuracy was improved by combining ARIMA model and random forest for regression analysis of residuals. This paper demonstrates the potential of machine learning and time series analysis in real estate market value assessment, which provides new perspectives and valuable references for market analysis, investment strategy development and policy decisions.

Keywords: machine learning, deep learning, real estate valuation, random forest, ARIMA model

1 INTRODUCTION

1.1 Research Background

In recent years, with the rapid development of machine learning techniques and the advent of the big data era, more and more studies have begun to explore the use of these advanced techniques to value and analyze the real estate market. Alshammari T[1] pointed out that, although machine learning techniques have been widely used in a number of fields, there is not enough research on the prediction of real estate prices. Sisman S et al[2] applied a multiple regression analysis model to value house prices in several regions and used machine learning metrics as well as WtR, COD and PRD techniques to measure their valuation accuracy. Zhao Y et al. combined deep neural network with XGBoost method to forecast house prices[3].Then, the advantages of machine learning and ARIMA model were organically combined, and a new idea of integrated forecasting model was proposed; the model's ability to predict house prices was improved.

A. Haldorai et al. (eds.), *Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2024)*, Atlantis Highlights in Intelligent Systems 11,
https://doi.org/10.2991/978-94-6463-490-7_52

1.2 Research topic and significanc

This project intends to combine machine learning with ARIMA time series analysis to construct an integrated model that can accurately forecast house prices. The research results of this project will provide new ideas for the development of China's real estate market, as well as for China's real estate market research, investment decision-making and policy formulation. The methodology incorporates past sales data and can provide market players with more comprehensive and detailed valuations. This innovative approach will change how the real estate industry understands and responds to market changes, thus providing a better basis for decision-making in the real estate industry.

2 LITERATURE REVIEW

2.1 Current status of real estate appraisal

When forecasting house prices, one of the simplest machine learning models, a linear regression model, is usually used.[4]. For many nonlinear problems it will be analyzed by time series analysis methods [5].Some researchers have adopted the boosting algorithm to evaluate the house price of second-hand houses, and optimized the model to adapt it to the needs of actual put into use, thus playing a key role in the actual property evaluation work[6][7].Other authors introduced support vector machine (SVM) technology to the field of property valuation and developed a support vector machine-based real estate valuation model by taking advantage of its ability to handle small data sets and solve nonlinear problems[8].Yazdani et al. evaluated various types of machine learning and deep learning algorithms, including artificial neural networks, random forests, and K-nearest neighbor techniques, and applied them to a comparison of hedonic forecasting methods for house prices[9].These research results not only broaden the technical approach to real estate valuation, but also provide valuable references for subsequent research, showing the great potential and prospect of machine learning application in real estate valuation.

3 MODELING AND ALGORITHM STUDY

3.1 Modeling

In the field of statistical machine learning, constructing models is key to predicting accuracy.

Establish a multiple linear regression model. The model can be represented as[10]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + + \cdots + \beta_n X_n + \varepsilon_n \tag{1}$$

In this model, Y represents variables that act independently, $X_1, X_2, ..., X_n$ denotes variables considered as characteristics, β_0 denotes the intercept term, $\beta_1, \beta_2, ..., \beta_n$ denotes the model parameters, and ε corresponds to the error term.

Mathematical formulas for random forest regression models[11]:

$$H(X) = \frac{1}{T}\sum_{i=1}^{T} h_i(X) \tag{2}$$

In the model, $H(X)$ is the predicted value of this model for the input variables, N is the total number of decision trees, and $h_i(X)$ is the predicted value of the ith tree.

The gradient boosting algorithm model can be represented as[11]:

$$F(X) = F_0(X) + v\sum_{m=1}^{M} \gamma_m h_m(x) \tag{3}$$

$F_M(x)$ is the predicted value of the final model, $F_0(x)$ is the initial model, M is the number of iterations, v is the learning rate, γ_m is the optimal coefficient corresponding to the mth tree, and $h_m(x)$ is the contribution of the mth tree to the final model.

Mathematical Modeling of Support Vector Machines[11]:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t.\, y_i(w^t x_i + b) \geq 1, i = 1,2,\ldots,m \tag{4}$$

where w is the normal vector to the hyperplane and b is the offset to the origin.

3.2 Implementation of the algorithm

3.2.1 Data set description and preprocessing

The dataset we use comes from the UCI database [12]. In the data preprocessing process, we first eliminated variables that are not relevant to prediction. For date data, it was converted into a form suitable for time series analysis. Data preprocessing lays the foundation for subsequent analysis to improve the data set availability.

3.2.2 Feature extraction and machine learning

In feature extraction, this project proposes to adopt the Keras framework, using the ReLU activation function to construct two hidden layers, which can better portray the nonlinear correlation and interactions between the data, and extract more representative features from them. We train with four machine learning models such as linear regression model, random forest, gradient boosting, and support vector machine. Each model can be used to make predictions about house prices based on the extracted features.

3.2.3 Evaluation Indicators

In terms of evaluation indexes, this paper selects the minimum mean square error and the minimum mean square error as evaluation indexes to reflect the forecasting effect of the model more directly. On this basis, using R^2, the interpretability of the model forecast is quantified to measure the overall fit of the model. The method provides a basis for the selection of real estate price forecasting models, thus ensuring the accuracy of real estate price forecasting.

3.2.4 Model combined with time series analysis

Finally, the effectiveness of the model in practical application was verified using the ARIMA model and the machine learning model of test optimization. Since it makes full use of the dynamics of real estate prices in different periods, the development trend of real estate prices can be better predicted using time series analysis.

4 METHODOLOGY APPLICATION AND CASE STUDY

The empirical case studies in this paper are selected from the real estate valuation market data of Xindian District, New Taipei City, Taiwan.

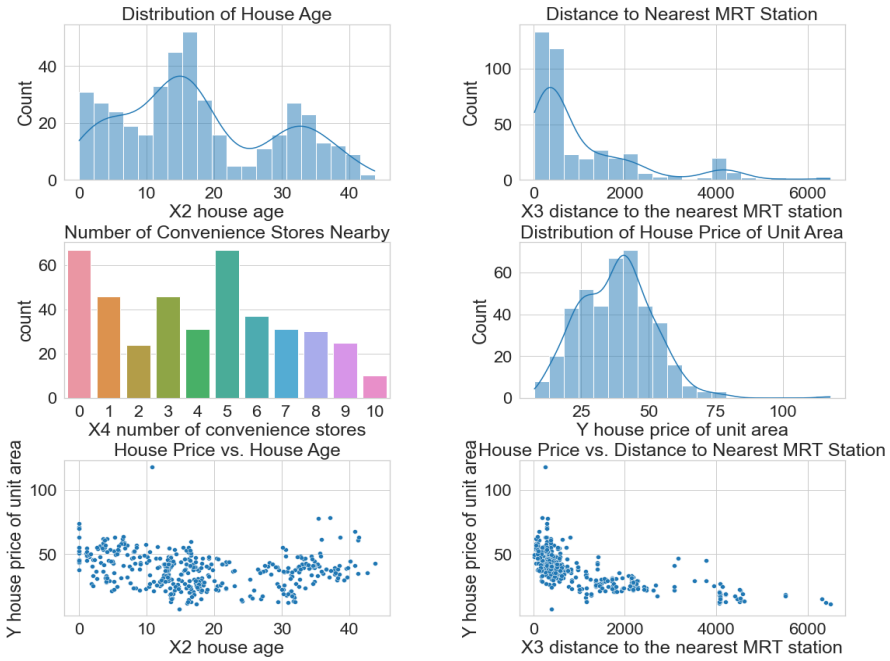

Fig. 1. Real Estate Market Analysis Chart Set

According to the analysis of the figure 1, there is a wide range of ages of houses, especially the number of new houses and houses about 30 years old is noticeable, which indicates the availability of houses in the market in all periods. The majority of homes are located within 500 meters of a metro station, showing the advantages of location, however, homes further away may be limited in price and appeal. In addition, the number of nearby convenience stores becomes a key indicator in assessing the ease of living, reflecting its importance in life. The wide distribution of prices per unit area illustrates the fact that house prices are influenced by a number of factors, including location, age and amenities. The correlation of house prices with age and proximity to metro

stations highlights the role of new housing and accessibility in increasing property values.

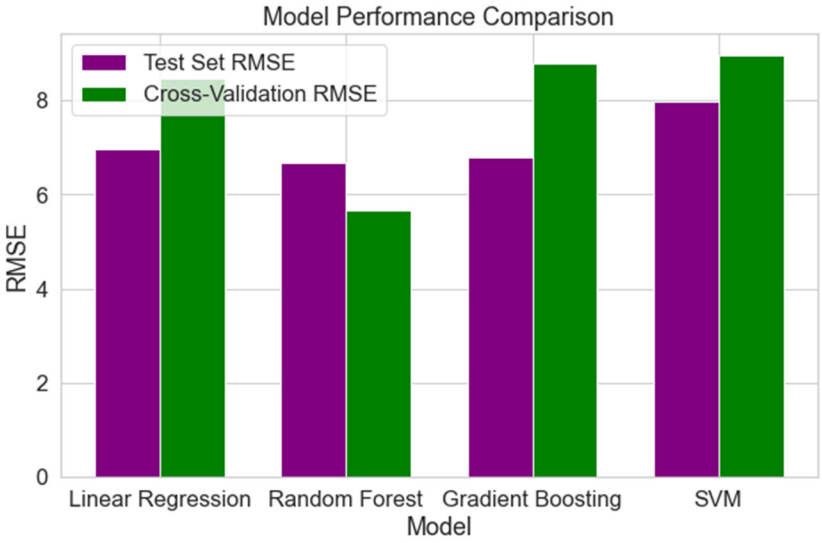

Fig. 2. Models Performance Comparison

As can be seen in Figure 2, the Random Forest model works best in predicting real estate valuations on the test set, indicating the highest prediction accuracy. Among them, the SVM model performs the worst, which is explained by the fact that it does not reflect the complexity of the data well. The gradient boosting model, on the other hand, is moderate, better than regression model and support vector machine, but not as good as random forest. As show in table 1.

Table 1. Model Performance in Real Estate Valuation Forecast

Metric	Value
RMSE	6.675
MAE	4.450
R^2	0.705

The results show that the Random Forest model has a better prediction effect on real estate prices, and its maximum mean square error reaches 6.675, which indicates that its prediction results are not much different from the real data; the maximum mean square error is 4.450, which indicates that this model has a high prediction accuracy and credibility. R^2=0.705, which indicates that this model can explain about 70.5% of the variation of the information, which indicates that this model has a strong fitting ability. Overall, the random forest model is a good method for predicting real estate prices. In conclusion, the random forest model has high prediction accuracy and good model fit, and it is also a popular model at present.

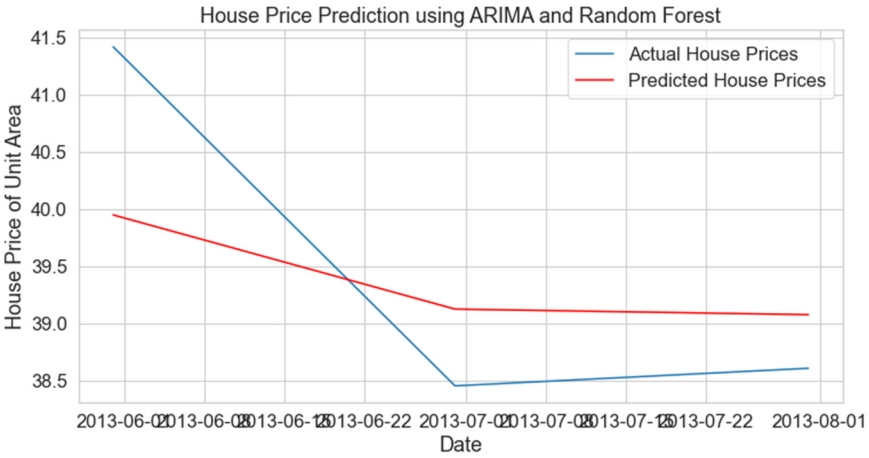

Fig. 3. Comparison chart of house price forecasts from time series analysis

we combines ARIMA with random forest regression to predict the trend of the real estate market. As show in figure 3, the actual trend of housing prices predicted by the model coincides with the reality. The model obtained a minimum mean square error of 0.946, indicating that the use of the combination of time series and machine learning can effectively reduce the mean square error, showing the importance of the model fusion technique to improve the accuracy of prediction.

5 CONCLUSION

5.1 Main conclusions

The results of this paper show that the method of combining time series analysis and machine learning for real estate price prediction has good application prospects. Among them, the random forest model performs better in many aspects. On this basis, the combination of ARIMA model and random forest technique can greatly reduce the mean square error of the model and can effectively improve the prediction accuracy. The research results of this project will provide a new perspective for the value assessment of real estate enterprises, and also lay the foundation for the prospect of using deep learning technology and traditional statistical analysis methods in the real estate market. Through further optimization and integration of the model, the method will play an increasingly important role in the analysis and prediction of the real estate market. The Figure 4 can facilitate readers to visualize the flow of the research method.

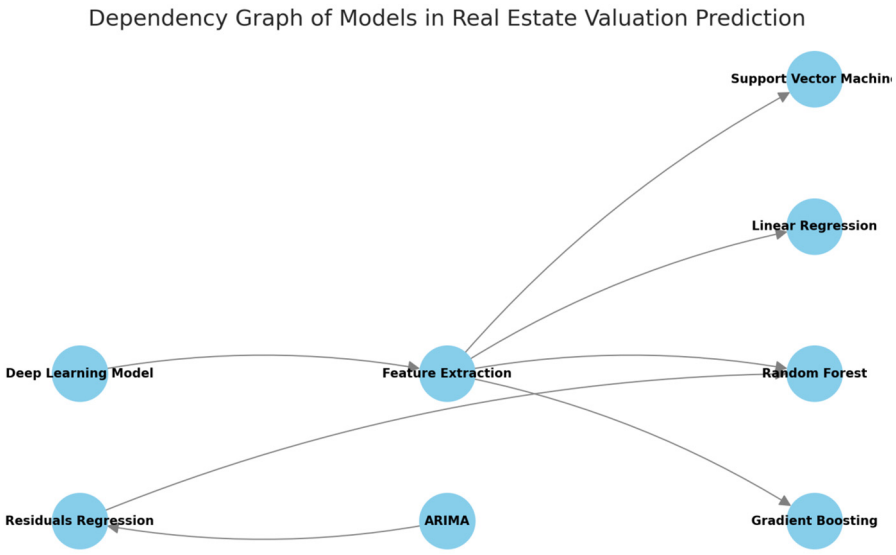

Fig. 4. Dependency Graph of Models in Real Estate Valuation Prediction

5.2 Future research directions

the methodology of this study in the field of real estate valuation prediction is expected to be further extended and deepened. Future research can be conducted in the following areas:

(1) Dataset and feature expansion: future research can explore more diversified and segmented datasets, such as considering macroeconomic factors, regional development policies, and changes in supply and demand in the real estate market, to further enrich the predictors of the model. Meanwhile, further mining and creation of new features through deep learning techniques may identify more key factors and improve the accuracy and interpretability of the predictions.

(2). Model fusion and optimization: considering that different models may perform prominently on different data subsets, future research can explore more model fusion techniques, such as stacking and hybrid models in integrated learning, to maximize the advantages of various algorithms. In addition, the application of parameter tuning and automated machine learning techniques are important directions to improve model performance.

REFERENCE

1. Alshammari T. Evaluating machine learning algorithms for predicting house prices in Saudi Arabia[C]//2023 International Conference on Smart Computing and Application (ICSCA). IEEE, 2023: 1-5.

2. Sisman S, Aydinoglu A C. Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis[J]. Land use policy, 2022, 119: 106167.
3. Zhao Y, Chetty G, Tran D. Deep learning with XGBoost for real estate appraisal[C]//2019 IEEE symposium series on computational intelligence (SSCI). IEEE, 2019: 1396-1401.
4. Ghosalkar N N, Dhage S N. Real estate value prediction using linear regression[C]//2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 2018: 1-5.
5. Xu, Xiaojie, and Yun Zhang. "House price information flows among some major Chinese cities: linear and nonlinear causality in time and frequency domains." International Journal of Housing Markets and Analysis 16.6 (2023): 1168-1192.
6. Sibindi R, Mwangi R W, Waititu A G. A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices[J]. Engineering Reports, 2023, 5(4): e12599.
7. Ragapriya N, Kumar T A, Parthiban R, et al. Machine Learning Based House Price Prediction Using Modified Extreme Boosting[J]. Asian Journal of Applied Science and Technology (AJAST), 2023, 7(1): 41-54.
8. Goel Y K, Swaminathen A N, Yadav R, et al. An Innovative Method for Housing Price Prediction using Least Square-SVM[C]//2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2023: 928-933.
9. Yazdani M. Machine learning, deep learning, and hedonic methods for real estate price prediction[J]. arXiv preprint arXiv:2110.07151, 2021.
10. Wu, X., & Zhang, M. (2020). Applied Regression and Classification: Implementation Based on R and Python, 2nd Edition. Beijing: Renmin University of China Press.
11. Zhou, Z.-H. (2016). Machine Learning. Beijing: Tsinghua University Press.
12. UCI Machine Learning Repository. (Year). Real Estate Valuation Data Set. Retrieved from https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set.