A Comparative Study of Random Forest Regression for Predicting House Prices Using

Mohan Mao
Zhengzhou Foreign Language School New Fengyang Campus, Henan Province, Zhengzhou,450000, China
3576795694@qq.com

Abstract. Based on the rapid development of the real estate market, real estate prices in various regions of the world fluctuate greatly and are unstable, and we need to make some predictions for real estate prices. However, in reality, we pay too much attention to the relationship between past property prices and current property prices and often ignore the prediction of future house prices. Research on predictive models is lacking. Therefore, studying real estate forecasting models is one of the best solutions to solve the problems faced by the real estate market based on the thinking of the current situation. In response to this problem, I propose to use a random forest model, gradient boosting, and optional to build a reasonable predictive model. The final results prove that this predictive model can be used to some extent to predict changing real estate prices in the future market. It is hoped that the method in this paper can provide a reference for subsequent research on predictive models.

Keywords: random forest, gradient boosting predict model.

1 Introduction

1.1 Research background

For a number of practical issues, such as real estate prices, research cannot be ignored. For this reason, research on the forecasting of real estate price models is essential. By concentrating on the research pathway shown for this research, the goal is to establish an effective prediction model through the real estate market price change. Scholar' perspectives on studying predictive models are, of course, somewhat consistent and duplicated as a result of earlier research, but there are still a lot of approaches, options, combinations, and applications with practical applications that remain unexplored. Based on this, this paper establishes the research direction of establishing a real estate price prediction model.

1.2 Research Significance

This paper can, to a certain extent, make up for the shortcomings of the current prediction model establishment field for this research, expand the content of the prediction model establishment field, support the resolution of related issues, analyze

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_63

the perspective of the establishment of the predictive model, combine practical factors, and use an innovative approach to some issues arising from the current prediction of real estate prices and real estate market. This paper can, to a certain extent, make up for the shortcomings of the current prediction model establishment field for this research, expand the content of the prediction model establishment field, support the resolution of related issues, analyze the perspective of the establishment of the predictive model, combine practical factors, and use an innovative approach to some issues arising from the current prediction of real estate prices and real estate market.

(1)Market comparison technique: Choose real estate price cases with the same goals and other comparable market conditions, compare them to the real estate evaluation object's conditions, quantify each factor exponentially, and determine the real estate's value by accurately comparing and adjusting the indexes.

(2)Income reduction method: To determine the price of a home, first estimate the net income that real estate will generate over the course of each period. Next, use the appropriate reduction interest rate to bring that income up to date. The main objective of the income reduction approach is to ensure that real estate will generate more money in the future than it costs, and the calculation formula is typically real estate value equals to real estate net income reduction interest rate.

(3)Cost-based valuation: the cost of construction plus various taxes and normal profits determines the price of a home. This method is appropriate when there are not many real estate transactions and it is not possible to use the market comparison method. The formula for this method is: real estate price = full value of real estate rebuilt minus building depreciation.

(4)Route pricing method: For the same block of land, the value of the land use right is rather stable whereas the value of the land has a strong relationship with its location. By altering the width and depth of the frontage, the land value of the valuation item can be determined if the average cost of land in the block is known.

However, the field of house price forecasting has not yet been deeply integrated with the methods of machine learning, and although there have been some studies in the academic community, there are some aspects that can be improved.

1.3 Research content

This paper takes the data of property prices and a series of factors such as the number of bedrooms, size, number of living rooms, and size of a series of cities belonging to the Commonwealth of Australia in the fifth month of the Western calendar in 2014 as an example, analyzes this data and uses the regression techniques of random forest and gradient boosting to establish a predictive model as a research center. The purpose of this research is to create a reasonable method for establishing prediction models, and strive to improve the application probability of predictive models in solving real problems, apply theoretical results to time, and effectively improve the application time of prediction models.

1.4 Chapter Introduction

Section two is related work that carried out the relative research of the house price prediction. Section three is a methodology that accommodates the regression techniques that I used in the paper such as random forest, Gradient boosting, and so on. Section four is the experiment.Section five is the conclusion. The section six is a reference, it is about some relative papers that I used when I wrote this paper.

2 Related Work

The use of machine learning for predicting home price growth has recently attracted a lot of relevant study. For instance, the genetic algorithm-based prediction model has the drawback of being highly dependent on the fitness function, not having a smooth and continuous optimization process, and having a large amount of sporadic data, necessitating more sample support; In order to perform classification and regression tasks, predictive models based on BP neural networks, one of the most popular artificial neural networks, are frequently utilized. However, this model has the drawbacks of being prone to local minima and training being dependent on initialization;Models based on time series analysis algorithms are frequently used to analyze time series data and forecast future trends, but they have the drawbacks of requiring prior determination of the type and parameters of the time series model as well as being sensitive to the data's initial value and noise. One such time series analysis model is the ARIMA model, which can forecast nonstationary time series. The basic idea behind it is to create an ARIMA model by moving average and autoregressively averaging the difference terms in the time series. This model's drawback is that it is simple to overfit the issue, and the model's parameters must be carefully chosen.

The random forest rule, gradient boosting, and optional tuning are all employed in this research. Gradient boosting makes good use of weak classifiers for cascade, fully taking into account the weight of each classifier, and can be applied to both regression issues and binary classification problems. It can also handle various types of data, including continuous and discrete values. The random forest algorithm, which is based on the decision tree algorithm, implements sample classification by randomly choosing some attributes and samples for the learning and voting of numerous decision trees. The advantage is that it can effectively avoid overfitting, and it has a good effect on the classification task of high-dimensional data and large amounts of data. Advantages of Optuna for hyperparameter tuning: (1) Easy integration and many functions: Simple installation is required, and then you can start using it. It is possible to handle a wide range of tasks and find alternatives with the best adjustments. (2) Instant dynamic search space: familiar Pythonic syntax, such as conditions and loops, for automatic search for the best hyperparameters. (3) State-of-the-art algorithms: quickly search large spaces and prune promising experiments faster for better and faster results. (4) Distributed optimization: Hyperparameter search can be easily parallelized with little change to the original code. (5) Good visualization: Various visualization functions can also be used to visually analyze optimization results.

Therefore, this paper uses these three methods to complement each other and propose a relatively efficient model.

3 Methodology

3.1 Random forest

Random forest tree We have already studied classification trees, and random forest contains a large number of them. The input vector is sorted by category. Each tree serves as a categorisation and represents one "vote" for the input vector. The forest is the tree that obtains the most votes. Now that there are N training sets, N random numbers must be located in the forest. Each training will be replaced with the original data. This diagram demonstrates how a forest's trees develop. - When there are M input variables, provide m M in such a way that m variables are picked at random from M at each node, and the best split on this m is used to split the node. During the expansion of the forest, m has a fixed value. Each tree realizes all of its potential. Never trim.(Looking at it is dull; in the article that follows, I will provide an example to illustrate it.) The relationship between any two forest trees, which was initially shown in the original study on random forests, is one aspect that impacts the forest error rate. - The robustness of every tree in the forest (trees with low error rates are robust classifiers); the forest error rate rises as correlation rises.By bolstering individual trees (more precise classification), the forest error rate is reduced. characteristics of random forest It can manage a very large number of input variables even without variable elimination. It offers an estimate of the crucial categorization variables. Internally and impartially, the generalization error is assessed as the forest expands. By accurately imputing missing data, accuracy is maintained even when the majority of the data is missing. It can be used to balance errors in taxon-imbalanced datasets. The forest that was created can be saved for use with future data. Make model computations that illustrate the connections between variables and categories. It computes proximity relationships between pairs of cases in order to cluster examples, identify outliers, or offer a fascinating view of the data (by zooming). By extending the aforementioned functions, unsupervised clustering, data visualization, and outlier identification can be done with unlabeled data. It proposes an experimental method to discover variable interactions. In a nutshell, this approach is very important and precise. Forest Operation at Random Understanding and utilizing the different options requires a deeper understanding of their computations. Two random forest-generated data pieces account for the majority of available options.While constructing the training set for the current tree using sampling with replacement, a third of the cases are left out of the sample. In order to generate a continuous, unbiased estimate of the classification error when new trees are added to the forest, "out-of-bag" (oob) data is used. Using it, estimates of variable importance are also generated. All of the information is surveyed once each tree has been built, and proximity is determined for each pair of cases. If two cases share a terminal node by one, they are closer together. At the conclusion of the run, the proximity is normalized by dividing by the number of trees. Proximity is used to fill in gaps, spot anomalies, and build low-dimensional

perspectives of the data. The thesis's information is disjointed and difficult to understand. Simply put, it is divided into the ensuing steps. Variable OOB error estimates The importance of lacking training set values in the interactive PROXIMITIES zoom prototype for Guinea values missing from the test setCases Without Labels Outlier-based unsupervised learning strikes a compromise between novelty and prediction error.Fig. 1 shows that the process of random forest.

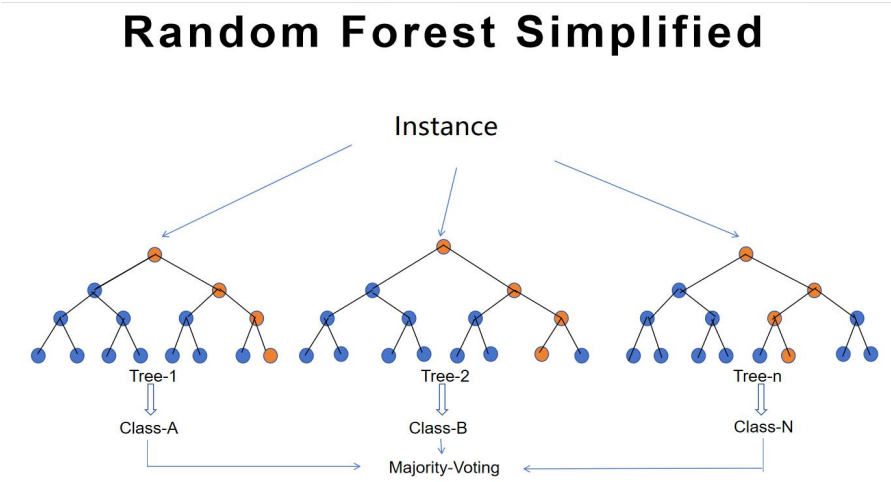

Fig. 1. The process of the Random Forest （Photo/Picture credit: Original)

3.2 Gradient boosting

The fundamental idea behind gradient boosting is to create multiple weak learners serially, each of which is intended to fit the negative gradient of the loss function of the prior accumulation model, in order to decrease the cumulative model loss after adding the weak learner in the direction of the negative gradient. Consider a sample with a true value of 10, a learner fit result of 7 for the first learner, a residual of 10-7=3, a residual of 3 used as the fitting target for the next learner, a learner fit result of 2 for the second learner, a boosting model combined with the two weak learners predicting the sample to be 7 + 2 = 9, and so on to add the weak learner to improve performance. This is a pretty, really simplistic image, but it shows the basic idea behind gradient boosting. Gradient boosting is often referred to as gradient descent on the function space. The gradient descent that we are more acquainted to is often the gradient descent of the value on the parameter space when training a neural network, for example, computing the gradient of the current loss on the parameter in each iteration and updating the parameter. Gradient By fitting the loss function to the gradient of the previous cumulative model, Riving develops a weak learner for each iteration, which is then added to the cumulative model to gradually lower the loss. Therefore, gradient descent of parameter space uses gradient information to change parameters as opposed to gradient descent of function space, which uses gradient fitting to reduce loss. Fig 2 shows that the process of the gradient boosting.

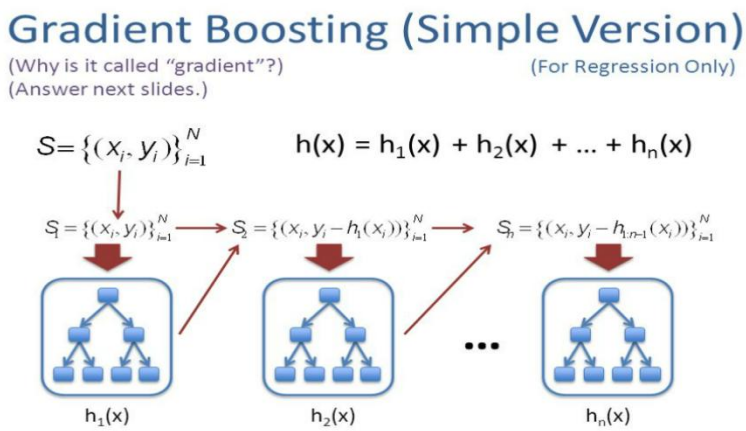

Fig. 2. The process of Gradient boosting (Photo/Picture credit: Original)

3.3 Optuna

Optuna is a framework for automatic hyperparameter tweaking that was created specifically for machine learning, deep learning, and user APIs with functionality for scripting languages. Optuna's code is extremely modular as a result, allowing customers to dynamically create a search space for hyperparameters that suits their needs. PS: The Optuna framework was created for rapid and automated research.

4 Experiment

The dataset used in this article is the specific price of each property in cities belonging to the Commonwealth of Australia, such as Melbourne and Sydney on May 4, 2014, as well as the information contained in a single property, such as the number of bathrooms, the number of master bedrooms, whether there is a balcony, basement space, etc., a total of 14,600 pieces of data. The main purpose of this paper is to use this data set to correlate the property price with the specific information it contains, such as the number of bathrooms, the number of master bedrooms, whether there is a balcony, basement space, etc., and establish a prediction model, and finally calculate the error to verify the accuracy of this model.

Data processing: Describe the data with some visual style by applying T = Transpose and style and check the target variables for distribution. The result can be seen in Figure .3.Then check the missing values and drop those values and who are less than 50% we will try to fill them. Finally, we tried multiple models but the model was with the lowest RMS" XGB Regressor" and use Optuna to adjust the parameters

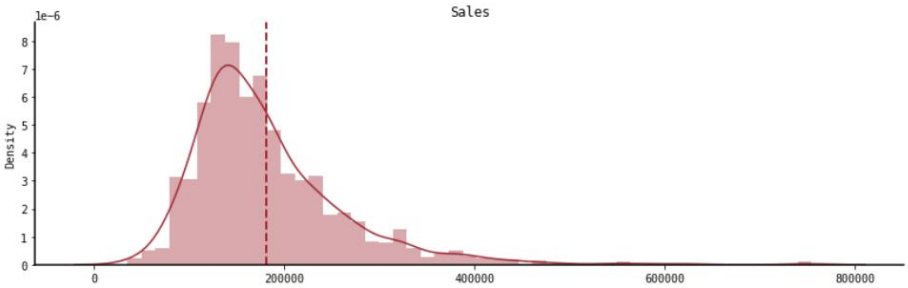

Fig. 3. Picture after data processing (Photo/Picture credit: original）

5 Conclusion

In recent years, the socio-economy and development continue to deepen, the society's demand for reliable future prediction means has also deepened, at present, the academic research on predictive models has been quite rich, domestic and foreign academic researchers from different research perspectives, for the establishment of more complete, accurate prediction models to make suggestions. But combined with specific research results, these research contents focus more on the technical and technical level. In the process of research, through the perspective of many cases of the Australian Federation, this paper introduces the theory of regression techniques like random forest and gradient boosting and puts forward all-around countermeasures for the establishment of more complete and accurate prediction models. This also reflects the innovation of this paper at the level of theoretical and practical problems in the research. The shortcomings in the research of this paper are mainly reflected in the insufficient application of theory. In this study, this paper uses Regression Forest theory and Gradient Boosting theory to propose countermeasures for the establishment of more comprehensive and accurate prediction models. However, it should also be fully recognized that the ideological connotation of the above theory is rich and limited by its academic ability, so in the application of this theory, it is inevitable that there will be situations such as insufficient application and deep understanding. This is the inadequacy of the study in this article. In the field of establishing predictive models, there is still a problem of insufficient combination of theory and practice, and in this regard, there is still a lot of room for progress in the research of establishing predictive models. Accordingly, subsequent research can be further improved in terms of combining theory and practice.

References

1. The prediction model of secondary housing prices based on XGBoost [J]. Zhang Zhifeng, Cui Yadong, Cui Xiao. Digital technology and applications. 2019(11)

2. Machine Learning Predictions of Housing Market Synchronization across US States: The Role of Uncertainty [J]. Rangan Gupta, Hardik A. Marfatia, Christian Pierdzioch, Afees A. Salisu. The Journal of Real Estate Finance and Economics. 2021 (prep)
3. The prediction model of secondary housing prices based on XGBoost [J]. Zhang Zhifeng, Cui Yadong, Cui Xiao. Digital technology and applications. 2019(11)
4. House price prediction model based on multivariate linear regression [J]. Li Shengda. Scientific and technological innovation. 2021(06)
5. Housing price prediction: parametric versus semi-parametric spatial hedonic models [J]. Montero José María,Mínguez Román,Fernández Avilés Gema. Journal of Geographical Systems. 2018 (1)
6. Research on Ensemble Learning-based Housing Price Prediction Model [J]. Bowen Yang,Buyang Cao. Big Geospatial Data and Data Science. 2018 (1)
7. Modeling and prediction analysis of GDP in China [J]. Fan Jinrong Fan. Business information. 2020(11)
8. Establishment and prediction analysis of the market prediction model of small package edible oil [J]. Zhao Yang, Liu Yang, Zou Xin. Grain, oil, and food science and technology. 2015(01)
9. A Boost to Renting[J]. Wang Jun. Beijing Review. 2017(36)
10. Integrated learning invasion detection method based on random forest [J]. Sheng Zhan, Chen Lin. Computer knowledge and technology. 2022(19)