American Journal of Information Science and Technology
2023; 7(2): 70-75
http://www.sciencepublishinggroup.com/j/ajist
doi: 10.11648/j.ajist.20230702.14
ISSN: 2640-057X (Print); ISSN: 2640-0588 (Online)

Predicting Real Estate Price Using Stacking-Based Ensemble Learning

Huiyi Zhao[1, *], Kainuo Wang[2]

[1]School of Urban Design, Wuhan University, Wuhan, China

[2]School of Resource and Environmental Science, Wuhan University, Wuhan, China

Email address:

2021302092032@whu.edu.cn (Huiyi Zhao), 2020302051136@whu.edu.cn (Kainuo Wang)

[*]Corresponding author

To cite this article:

Huiyi Zhao, Kainuo Wang. Predicting Real Estate Price Using Stacking-Based Ensemble Learning. *American Journal of Information Science and Technology*. Vol. 7, No. 2, 2023, pp. 70-75. doi: 10.11648/j.ajist.20230702.14

Received: March 14, 2023; **Accepted**: April 23, 2023; **Published**: April 27, 2023

Abstract: As one of the leading researches focusing on modern economics, real estate industry not only affects people's well-being but also has a close relationship with the national economy and social stability. Nevertheless, there are numerous complex factors that influence real estate prices, which makes house price forecasting remain a classic and challenging problem in the field of data analysis. The development of data mining and machine learning has greatly facilitated the analysis and extraction of useful information from complex data sets and the building of models to make predictions. In this study, a stacking-based ensemble model is proposed to identify potential links between property prices and various factors so that the more accurate prediction of property prices can be made. Some base predictive models, including linear regression, support vector regression, ridge regression, least absolute shrinkage and selection operator, machine language programs, random forest regression, and gradient boosting regression are trained to individually predict the estate price in the experiment. Then, the stacking-based ensemble model is obtained by integrating competent base predictive models and optimized using Grid search. The experimental outcomes indicate that the proposed model is superior to base predictive models and can be more accurate in predicting house prices.

Keywords: Ensemble Model, Real Estate Price, Stacking, Prediction

1. Introduction

The fluctuation of housing prices directly or indirectly affects the stable development of society, due to its significant implications on relevant industries and fields such as construction, investment, and public welfare. This makes house price forecasting become a hot topic in economics, sociology, and other areas. However, real estate prices are influenced by many complex factors including human, social, geographical, political, and so on. Finding an accurate method of predicting house prices can therefore give people a more objective view of the impact of various factors on house prices and help them to make informed decisions.

Traditionally, the appraisal of the real estate is conducted by a licensed professional, who may inevitably produce inaccuracy due to some bias and subjective assumption. Due to recent advances in artificial intelligence technology, the application of data mining and machine learning allows people to intuitively and accurately find patterns within complex data sets, making it possible to build appropriate predictive models to solve various problems including real estate price forecasting.

A widely accepted approach to improving the prediction performance of a classification task is to construct a set of base classifiers and combine their outputs [1]. In this study, a stacking-based ensemble model is proposed to identify potential links between property prices and various factors so as to forecast the property prices accurately. Firstly, some promising base predictive models, including linear regression (LR), support vector regression (SVR), ridge regression (Ridge), least absolute shrinkage and selection operator (Lasso), machine language program (MLP), random forest regression (RFR), and gradient boosting regression (GBR) are trained and compared. Then, several competent base predictive models are selected to form an ensemble learning

model, in which, Ridge plays the role of meta-regression model. Finally, Grid search is applied to optimize the process and find the best-fit ensemble model. The experiment result indicates that the stacking-based ensemble model outperforms the other base classifiers.

The remainder of this study is organized as follows. Section 2 reviews the previous works on real estate price prediction and ensemble learning methods. Details of the proposed ensemble model are explored in Section 3. Section 4 demonstrates the experiment results. Section 5 concludes the study along with some prospects.

2. Related Work

2.1. Real Estate Price Prediction

Researchers are interested in identifying patterns of real estate price volatility because of the significant social and economic effects of real estate pricing. According to the hedonic pricing theory put forth by renowned economist Sherwin Rosen, the real estate price can be characterized as a utility function of many connected variables [2]. Traditionally, licensed experts have calculated the value of real estate. Their work is frequently supplemented by computer systems known as Automated Valuation Models (AVM) and Computer Assisted Mass Appraisal (CAMA) [3-5].

Recently, an increasing number of researches have used machine learning models to predict the property prices due to the rapid development of data mining and machine learning. For example, Li et al. [6] used an SVR-based method to predict the real estate price. Sarip et al. [7] used a fuzzy regression model to predict the real estate price. However, individual machine learning models tend to produce the estimated results that are not robust and stable, because the fused models tend to perform better than the individual models [8].

2.2. Ensemble Method

Because they frequently outperform individual machine learning models, ensemble learning models, which combine some individual machine learning models, have drawn a lot of attention in the field of machine learning. The Bayesian averaging ensemble approach is the forerunner, but more contemporary methods, such as error-correcting output coding, bagging, boosting, and stacking, etc., have also been well developed [9].

As one of the popular ensemble learning approaches, stacking focuses on mixing several classifiers created using various learning algorithms on a single data set that is made up of examples or pairs of feature vectors and their classifications [10]. To predict the problems in various domains, the more adaptable structure and stable stacking model demonstrate the significant advantages. For instance, He et al. [11] created a stacking-based ensemble model for credit scoring and achieved improved predictive performance. Sanchis & Hanssen [12] proposed an enhanced local correlation stacking method to accelerate computing speed. Kardani et al. [13] proposed a hybrid stacking ensemble method based on finite element analysis and field data. The stacking-based ensemble method can undoubtedly be effective in the field of real estate given its outstanding outcomes in other fields.

3. Methodology

3.1. Data Set Exploration

The 1970s Boston house price data set [14] from Carnegie Mellon University StatLib repository, which covers housing data from 506 different suburbs in Boston, Massachusetts, is selected in this study. It includes 506 instances, and has 14 features, in which 13 independent attributes are common features and the median value of home owned is target variable. Details of features in data set are demonstrated in Table 1 shown below.

Table 1. *The description of fourteen features in data set.*

Features	Description
CRIM	Per capita crime rate by town
ZN	The percentage of the presidential land that covers more than 25,000 square feet
INDUS	Percentage of non-retail commercial acres in each town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	The average number of rooms per dwelling
AGE	Proportion of owned units built before 1940
DIS	Weighted distance to Boston's five job centers
RAD	Radial highway accessibility index
TAX	Full value financial tax rate per $10,000
PTRATIO	Student to teacher ratio by town
B	1000 (Bk-0.63)^2 where Bk represents the percentage of blacks by town
LSTAT	The lower percentage of the population
MEDV	Median Value of home Owned ($1,000)

3.2. Individual Models

To analyze the competency of base predictive models such as LR, SVR, Ridge, Lasso, MLP, RFR, and GBR in the ensemble model, they are trained and evaluated on the data set individually. Details of these base predictive models are as follows.

LR – Linear regression is a standard statistical procedure to create a linear model which uses the least mean square method to adapt the parameters of the linear function [15].

SVR - Support vector regression creates a "spacing band" on both sides of the linear function so that the model is optimized by minimizing the width of the band and the total loss [16].

Ridge - Ridge regression is a biased estimation regression method dedicated to collinear data analysis [17]. In essence, it is an improved least squares estimation method.

Lasso - Least absolute shrinkage and selection operator is a compressed estimate [18]. It retains the advantage of subset shrinkage and is a biased estimator dealing with data with complex collinearity.

MLP - Machine language program. It works primarily by searching for the best weights to train the model in order to learn the desired output [19].

RFR - Random forests regression explores the correlation reduction by means of injection of randomness [20].

GBR - gradient boost regression uses the negative gradients of the loss function to solve the minimum value [21].

3.3. Stacking-Based Ensemble Model

The stacking-based ensemble learning method has attracted increasing attention in the data mining area due to its excellent performance in predictive analysis. Moreover, the greater the difference between the base predictive models that are integrated into the ensemble model through stacking, the better performance the ensemble model achieves [22]. Figure 1 shows the framework of the proposed ensemble model.

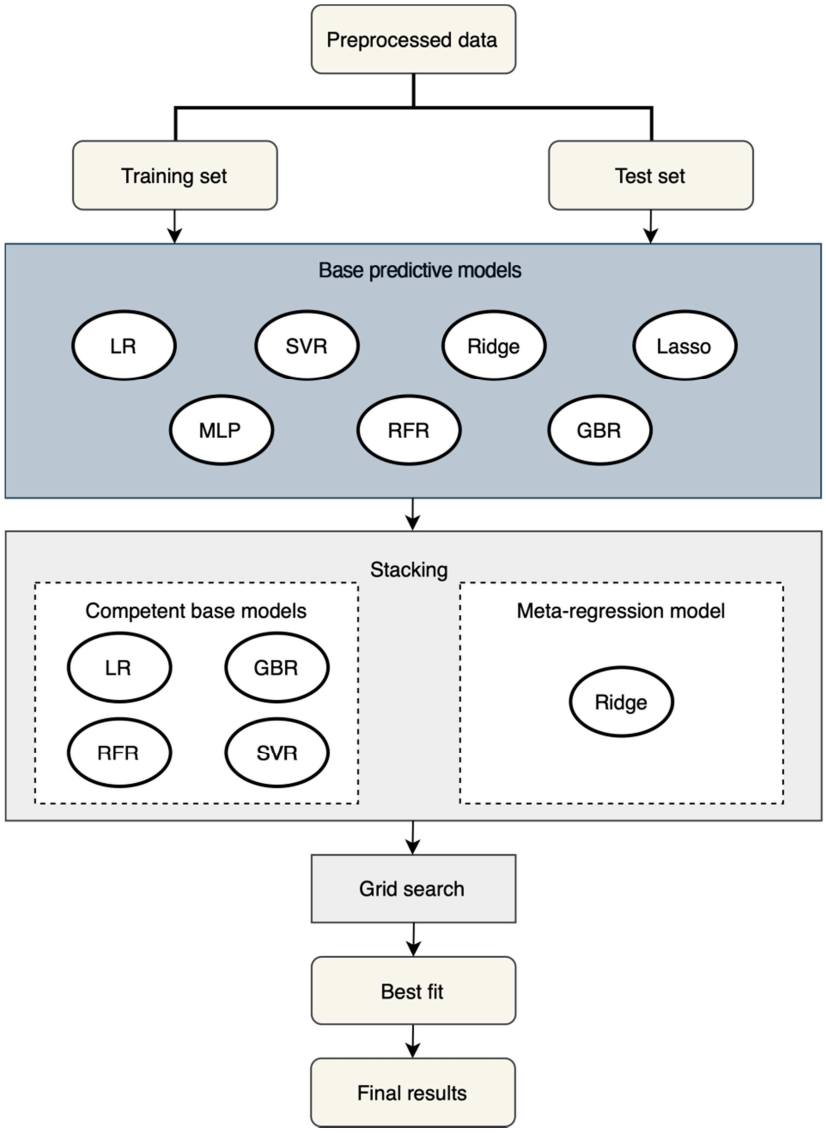

Figure 1. *Framework of the proposed model.*

Firstly, seven promising base predictive models including linear machine learning models (i.e., LR, Ridge, Lasso) and non-linear machine learning models (i.e., SVR, MLP, RFR and GBR) are trained using the data set and compared.

Then, four competent base predictive models with higher predictive accuracy - LR, GBR, RFR, and SVR are selected and integrated into the ensemble model through stacking.

Next, the predictive outputs from stacking integration are used as the new input features of Ridge that acts as the meta-regression model due to its excellent flexibility and high

efficiency [17], which makes the robust ensemble prediction.

Last but not least, Grid search method [23] is used to fine tune the "best fit" outcome of ensemble model through parameter optimization.

4. Experiment

This section illustrates the statistical metrics for evaluating all models including base predictive models as well as the stacking-based ensemble model. Then, the experiment results about real estate price prediction are also analyzed. All models and methods in the experiment are implemented with Python programming language.

4.1. Evaluation Metrics

Three widely used statistical metrics are adopted to measure the performance of base predictive models and the stacking-based ensemble model, including mean square error (MSE), explained variance score (EVS), and coefficient of determination (R^2). The calculation process of these three metrics is defined first. If N sample data is divided into r groups and the sample variance of the ith group is s_i^2, then the evaluate metrics of the whole group is shown in below equations (1-3).

$$MSE = \frac{\sum_{i=1}^{r}(y_i-1)s_i^2}{N-r} \quad (1)$$

$$EVS = 1 - \frac{Var(y_i-\hat{y}_i)}{Var(y_i)} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=0}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{N}(y_i-\bar{y})^2} \quad (3)$$

Where, N represents the size of the data set, while y_i ($1 \leq i \leq N$) is the practical real estate price at time. $\hat{y}_i$ represents the predicted estate price at time, $\bar{y}$ represents the average estate price of the N samples, and Var represents the variance. Among these evaluation metrics, the lower MSE value indicates the better performance. The best values of EVS and R^2 are both 1.0 and the smaller the values, the worse the quality.

4.2. Experimental Results Analysis

Table 2 demonstrates the evaluation results of seven base machine learning models (i.e., LR, SVR, Ridge, SVR, MLP, RFR, and GBR). It shows that the ranking of the better results in MSE index is GBR, RFR, Ridge, LR, Lasso, and MLP. In EVS index, the ranking of the better results is RFR, GBR, Ridge, LR, SVR, Lasso, and MLP; while in R^2 index, the ranking of the better results is RFR, GBR, Ridge, LR, SVR, Lasso, and MLP. Based on the above results, LR, SVR, RFR, GBR and Ridge are found to be more competent among these 7 base models, so LR, GBR, RFR, and SVR are selected as competent base predictive models and Ridge is selected as meta-regression model in the stacking-based ensemble model.

As indicated in Table 3, the best experimental results in MSE, EVS, and R^2 indexes among the seven base predictive models are found to be 9.3826 (GBR), 0.8669 (RFR), and 0.8668 (RFR) respectively, which are worse than that of the ensemble model (i.e., 8.7513, 0.8839, and 0.8841, respectively).

Table 2. *Evaluation results of seven base predictive models.*

Model	MSE	EVS	R^2
LR	17.1186	0.7632	0.7443
SVR	17.9265	0.7322	0.7322
Ridge	17.1155	0.7633	0.7443
Lasso	21.4957	0.7063	0.6789
MLP	24.2206	0.6401	0.6382
RFR	13.0010	0.8669	0.8668
GBR	9.3826	0.8593	0.8592

Table 3. *Comparison between the best base model and ensemble model.*

Model	MSE	EVS	R^2
Best base model	9.3826 (GBR)	0.8669 (RFR)	0.8668 (RFR)
Ensemble model	8.7513	0.8839	0.8841

Figure 2 shows the significant advantages of the ensemble model over the seven base predictive models in the form of a bar chart.

5. Conclusion and Future Work

Finding a suitable way to forecast real estate prices is of critical significance for investment financing and social stability. In this study, a stacking-based ensemble model is proposed to predict the real estate prices. Specifically, the ensemble model is integrated using four competent base predictive models, and one meta-model. The experimental results show that the stacking-based ensemble model outperforms the base predictive models.

In terms of future development, the model proposed in this study still has much room for development. Firstly, the preprocessing of the raw data can be enhanced using the data balancing algorithm to deal with the data imbalance. For example, a hybrid optimal ensemble classifier (HOEC) method that combines sampling and cost-sensitive classification learning was proposed by Yang et al. [24] to generate a balanced data set. Besides, more promising base predictive models can be explored to optimize the composition of the ensemble model, thus further improving the accuracy of the prediction. For example, artificial neural networks (ANN) and support vector machines (SVM) were applied in stock forecasting by Kurani et al. [25] to improve the forecasting accuracy. Moreover, more integration methods of base predictive models can be explored. For example, a heterogeneous Boosting-based ensemble model, Hboost, was proposed by Kadkhodaei et al. [1] to reduce the bias error. In addition, the ensemble model proposed in this

experiment can be deployed to more domains to solve predictive problems.

Figure 2. *Evaluation results of seven base predictive models and ensemble model.*

References

[1] Kadkhodaei, H. R., Moghadam, A. M. E., & Dehghan, M. (2020). HBoost: A heterogeneous ensemble classifier based on the Boosting method and entropy measurement. *Expert Systems with Applications*, 157, 113482.

[2] Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82 (1), 34-55.

[3] Jahanshiri, E., Buyong, T., & Shariff, A. R. M. (2011). A review of property mass valuation models. *Pertanika Journal of Science & Technology*, 19 (1), 23-30.

[4] McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, 30 (4), 239-265.

[5] Kauko, T. J., & d'Amato, M. (2017). *Advances in Automated Valuation Modeling: AVM after the Non-agency Mortgage Crisis*, Springer.

[6] Li, D. Y., Xu, W., Zhao, H., & Chen, R. Q. (2009). A SVR based forecasting approach for real estate price prediction. In *Proceedings of 2009 International Conference on Machine Learning and Cybernetics*, Baoding, Hebei, July 12-15, Vol. 2, pp. 970-974.

[7] Sarip, A. G., Hafez, M. B., & Daud, M. N. (2016). Application of fuzzy regression model for real estate price prediction. *Malaysian Journal of Computer Science*, 29 (1), 15-27.

[8] Nalić, J., Martinović, G., & Žagar, D. (2020). New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers. *Advanced Engineering Informatics*, 45, 101130.

[9] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, Cagliari, Italy, June 21-23, pp. 1-15.

[10] Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. *Machine Learning*, 54, 255-273.

[11] He, H. L., Zhang, W. Y., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, 98, 105-117.

[12] Sanchis, C., & Hanssen, A. (2011). Enhanced local correlation stacking method. *Geophysics,* 76 (3), 33-45.

[13] Kardani, N., Zhou, A., Nazem, M., & Shen, S. L. (2021). Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data. *Journal of Rock Mechanics and Geotechnical Engineering*, 13 (1), 188-201.

[14] Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5 (1), 81-102.

[15] Zhang, H. Y., Jin, Y. L., Shi, J. X., & Zhang, S. (2021). Predicting PM2. 5 concentrations using stacking-based ensemble model. *Applied and Computational Mathematics*, 10 (6), 156-162.

[16] Vapnik, V. N., & Lerner, A. Y. (1963). Recognition of patterns with help of generalized portraits. *Avtomat. i Telemekh*, 24 (6), 774-780.

[17] Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, 29 (1), 3-20.

[18] Ranstam, J., & Cook, J. A. (2018). LASSO regression. *Journal of British Surgery*, 105 (10), 1348-1348.

[19] Gaines, R. S. (1965). On the translation of machine language programs. *Communications of the ACM*, 8 (12), 736-741.

[20] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, 804-818.

[21] Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324.

[22] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14, 241-258.

[23] Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telecommunication Computing Electronics and Control*, 14 (4), 1502-1509.

[24] Yang, K., Yu, Z., Wen, X., Cao, W., Chen, C. P., Wong, H. S., & You, J. (2019). Hybrid classifier ensemble for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 31 (4), 1387-1400.

[25] Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Annals of Data Science*, 10 (1), 183-208.