Empirical analysis of I-GBDT to improve the accuracy of mass appraisal method

Jiaqi Wang\textsuperscript{1a}, Jingqi Sui\textsuperscript{1a,b}, Ziyan Zhang\textsuperscript{2b,d}, Jiabao Qi\textsuperscript{3c}, Nan Liu\textsuperscript{4b}, Junna Lv\textsuperscript{5b}

\textsuperscript{a}School of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing, 400074, China
\textsuperscript{b}School of Economics and Management, Chongqing Jiaotong University, Chongqing, 400074, China
\textsuperscript{c}School of Civil Engineering, Chongqing Jiaotong University, Chongqing, 400074, China
\textsuperscript{d}School of Management and Economy, Tianjin University, Tianjin, 300072, China
\textsuperscript{*}Corresponding author’s e-mail: suijingqi0468@163.com

Abstract. The transactions related to real estate in social and economic activities need to appraise the market value of commercial housing. In order to improve the accuracy and efficiency of mass appraisal model (MA), the paper proposed the Improved-GBDT algorithm based on hybrid feature selection, which combines the concept of supervised learning and unsupervised learning, selecting appropriate features based on the impact of features on results and the changing trend of feature data. The feature weights can be calculated by Random Forest and XGBoost. Considering grey correlation analysis, the correlation degree among features can be calculated, and appropriate features can be selected. In the paper, we took the housing sales data from May 2014 to May 2015 in USA into simulation analysis. The simulation results show that IGBDT model has better accuracy and stability than SVR, RF and XGBoost. Compared with RF and XGBoost model under optimal parameters, IGBDT not only reduced the average absolute error of prediction by 98% and 11%, but also make the maximum relative error drop by 97% and 12%. The results show that IGBDT can provide a powerful reference in evaluating the property market value.

1. Introduction
Real estate appraisal refers to a professional appraisers who follow the objective conditions of real estate and use professional appraisal methods like the market comparison approach to appraise the property market value. With the rapid increase in customer demand, the number of tax-based appraisals and bank credit appraisals has gradually increased, and real estate appraisal has begun to get more attention from scholars.

Formerly, Yu et al.\textsuperscript{[9]} modified the parameters of multiple linear regression through AHP based on actual trading cases. But this method could not obtain good evaluation results for situations where it is difficult to meet the distribution of actual data. Since then, Shi et al.\textsuperscript{[4]} The set-pair analysis method was supplemented, and an adaptive neural fuzzy inference method based on k-nearest neighbours was proposed to screen suitable real estate data as a basic data set for comparison to evaluate the value of real estate, but most of the real estate data at this stage presents The behaviour of small sample data is liable to fall into overfitting after screening, and it is difficult to obtain good evaluation accuracy. In
addition, Pang Fen\cite{7} proposed to use SVM model to identify the property category, and use Random Forest model to appraise the value of the property. But support vector machine shows poor classification performance on multi-category and high-dimensional data, making it difficult to improve accuracy. Li et al.\cite{5} propose using BP neural network to batch process real estate data sets Because the basic neural network is prone to overfitting and training is slow. Its processing results are not ideal.

Based on the above discussion, this paper proposes the GBDT algorithm based on hybrid feature selection. It considers the existing real estate data scale and integrates the concepts of supervised and unsupervised learning. It can filter out features that are helpful for improving the accuracy of the appraisal. Finally, the GA algorithm was used to tune the parameters of the GBDT model. The simulation results show that compared with the XGBoost model and the Random Forest model under the optimal parameters, the fitting degree and stability of the model on real estate data are significantly improved.

2. supervised learning analysis features

2.1 Random Forest Assessment Features
The model randomly selects a subset of attributes each time, and selects an optimal attribute for partitioning. The model guarantees randomness by controlling the number of $k$. It generally chooses $\log_{2}d$ as the value of $k$ ($d$ is the total number of attributes). The specific algorithm flow is described as follows:

- Establish a random forest model, according to $n$ Out of bag (OOB) performance test of each decision tree sub-model in a random forest, recorded as $\text{error}_i (i=1,2,...,n)$
- With the feature distribution of the rest of the decision tree unchanged, $N$ Number of out-of-bag data Group features add noise interference, and recalculate the $\text{Error}_i (i=1,2,...,n)$.
- The importance of the feature has a positive correlation with the average value of the error change after adding noise interference two times before and after. $j$ This attribute has the following formula:

$$\text{important}_j = \frac{1}{n} \sum_{i=1}^{n} (\text{error}_j - \text{error}_i)$$  (1)

From the formula (1), the importance of features can be obtained. It should be noted that the importance of features is measured by subtracting features and observing changes in accuracy, and random forest is implemented on the basis of decision tree integration. Similarly, each time the leaf node of the decision tree is split and the criteria for selecting the next node are different, the evaluation results of feature importance will be different. Therefore, a single supervised learning, that is, simply from the perspective of the results, cannot be effectively selected feature.

2.2 XGBoost appraises features
XGBoost (Extreme Gradient Boosting) is an integrated learning model based on boosting theory. This is a little different from the idea of Random Forest. First, we need to know the definition of the loss function of the model. There are formulas as follows:

$$\text{loss} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$  (2)

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$  (3)

$\hat{y}_i$ represents the predicted value of the model. $y_i$ represents the true value of the model. $K$ represents the number of the trees. $f_k$ represents the model of tree $k$. $T$ Indicates the number of leaf nodes in the tree. $w$ representing the score of each leaf node. $\gamma$ and $\lambda$ are hyper-parameters.

Of course, we have the following definition of the final objective function:

$$\text{loss}^{opt}(q) = \sum_i \left[ (\sum_{\omega_j} g_j) \omega_j + \frac{1}{2} \left( \sum_{\omega_j} h_j + \lambda \right) \omega_j^2 \right] + \gamma T$$  (4)

XGBoost calculates the gain of the structure score to select features as the segmentation points. It is
closely linked to the objective function. Then the importance of a feature is proportional to the sum of the number of its occurrences in all trees. This processing method is very simple, but it is difficult to accurately appraise some features that are not particularly important. This will affect our choice of features. So we need to develop a reference standard.

3. Unsupervised learning analysis features

3.1 GRA appraises features

GRA is a method based on unsupervised learning. It can quantitatively describe and compare the evolution of a system. The basic idea is to calculate the geometric similarity of several data columns through a set of reference data columns. Finally, the above results are used to judge whether they are closely linked. Before starting the calculation, it is necessary to do some preparation work like making features being dimensionless. Then, we need to calculate the correlation coefficient between the comparison sequence and the reference sequence. There is formula as follows.

\[
\rho = \frac{\min_{i} \min_{k} |y_{k} - y_{k}| + \rho \max_{i} \max_{k} |y_{i} - y_{i}| + \rho \max_{j} \max_{k} |y_{j} - y_{j}|}{\sqrt{\min_{i} \min_{k} (x_{k} - x_{k}) + \rho \max_{i} \max_{k} (x_{i} - x_{i}) + \rho \max_{j} \max_{k} (x_{j} - x_{j})}}
\]

\rho is called resolution coefficient. The smaller the value is, the greater the resolution will be. Through its calculation principle, we can find that this method of mining the correlation between features can make up for the shortcomings of supervised learning. Supervised learning is difficult to accurately appraise for less important features. But we can analyse the degree of correlation between these features and those very important features through GRA. So that those less relevant and less important features can be quickly removed. Then we can delete those features that are highly correlated with important features from those that are not particularly important features. Analysing the change in accuracy is the last step. Such a comprehensive concept of supervised learning and unsupervised learning will be more accurate for feature selection.

3.2 I-GBDT implementation process

The Improved-GBDT algorithm based on hybrid feature selection can be described from four aspects: data pre-processing, feature importance appraisal, parameter optimization, model testing and appraisal.

➢ Step1, we need to do suitable dimensionless processing considering the distribution of features. Dividing the data set by a certain proportion. Training set is called \(S'\). Testing set is called \(S''\). It is necessary to select a model from SVR, XGBoost, RF, GBDT that is more suitable for property appraisal in the initial situation.

➢ Step2, appraise and rank the importance of eigenvalues through RF, XGBoost. Calculate the degree of association between features by grey association analysis.

➢ Step3, considering the calculation results of the previous step, we choose different feature attribute sets to form different training and test sets. Then we need to generate multiple sets of two-dimensional vectors, which contain two hyperparameters of GBDT. Finally, GA finds optimal parameters with goodness of fit as the objective function.

➢ Step4, observe the changes in the model's goodness of fit, maximum error, mean square error, etc.

4. Empirical analysis

4.1 Experimental data

In order to verify the accuracy and stability of the GBDT appraisal model based on hybrid feature selection established in this paper, we used the house sales data of King County in the United States from May 2014 to May 2015 for simulation analysis. The basic situation of specific data are shown in Table 1 below.
Table 1 Basic information table of Property valuation

| Explanatory variables | Variable name     | Mean          | Standard deviation | Max            | Minimum         |
|-----------------------|-------------------|---------------|--------------------|----------------|-----------------|
| Dependent variable    | selling price     | 542874.93     | 372907.12          | 688500.00      | 75000           |
| Building factor       | Number of bedrooms| 3.37          | 0.89               | 10.00          | 0               |
|                       | Number of bathrooms| 2.12          | 0.77               | 7.75           | 0               |
|                       | House area        | 2082.49       | 922.82             | 9890.00        | 390.00          |
|                       | Parking area      | 15352.73      | 45773.94           | 1651359.00     | 572.00          |
|                       | Floor number      | 1.50          | 0.54               | 3.50           | 1.00            |
|                       | House rating      | 7.66          | 1.17               | 13.00          | 3.00            |
|                       | Construction area | 1791.47       | 829.41             | 8860.00        | 390.00          |
|                       | Basement area     | 291.01        | 446.62             | 4820.00        | 0               |
| Independent variable  | Time factor       |               |                    |                |                 |
|                       | Construction year | 1971          | /                  | 2015           | 1900            |
|                       | Year of restoration| /             | /                  | 2015           | 0               |
|                       | Geographical factor | latitude     | 47.56              | 0.14           | 47.78           | 47.16          |
|                       |                   | longitude     | -122.21            | 0.14           | -1.21           | -122.52        |

The dependent variable is represented by $y$, and the independent variable is represented by $x_1$ to $x_{12}$ respectively and unified use of the international system of units.

4.2 Select the base model
For different data distribution situations, the SVR model, Random Forest model, XGBoost model and GBDT model have different appraisal precision. So we need to choose the right model based on the data distribution. 8000 cases were chosen as training data. The following Figures 1 and 2 can describe the changes in the error of the SVR model and the random forest model after GA adjustment parameters.
It is found that the error values of GBDT are less than 100,000, and the XGBoost model has some examples with errors exceeding 100,000. We observed their performance on the test set. We can use Figure 5 to show their errors on the test set and Figure 6 to illustrate their relative error on the test set, setting GBDT as a reference.

After comprehensive consideration, we found that the maximum error of the GBDT model is smaller than the XGBoost model, which has better accuracy and stability.

4.3 Feature importance appraisal
First, we draw a learning curve to observe the model training. As shown in Figure 6, it is clear that the model is overfitting.

Therefore, considering the influence of features on the overall prediction accuracy, the features need to be deleted. \( x_1, x_2, \ldots, x_8 \) get dimensionless by dividing by the maximum. Then calculate the year gap between \( x_9 \) and 2016. Normalize \( x_9 \) based on these gaps. Remove the integer part of \( x_{11}, x_{12} \) and normalize \( x_{11}, x_{12} \). The appraisal results of the Random Forest model and XGBoost model are shown below.
Fig. 7 Histogram of the importance of features Fig. 8 Histogram of the importance of features

Obviously, both models have a higher appraisal of feature $x_3$, $x_6$, $x_9$, $x_{11}$, $x_{12}$. But the accuracy of the model with only five features is not convincing, as shown in the figure 7 and figure 8. We can find that models with only these five features have large errors both in the training and test sets. We need to add features that are not very important and suitable for the model to improve the accuracy of the model.

At this time, we obtain the degree of correlation between various factors through gray correlation analysis. Because feature $x_{10}$ is not significant in both appraisals, considering its own distribution characteristics, we try to ignore it. And calculate its correlation coefficient matrix and express it with a heat map, which is shown in figure 10.

Conclusions can be drawn from the chart. Features $x_3$, $x_6$, $x_9$, $x_{11}$, $x_{12}$ are necessary as well. In addition feature $x_1$ has the strongest correlation with the best feature and the importance of appraisal is low so that we can delete it. Then, we just need to know whether need to add $x_2$.

4.4 Appraisal and analysis of results

The model with $x_1$, $x_{10}$, removed as the optimal model at this time adjust its parameters by GA, and result is shown in figure 11.
XGBoost model and I-GBDT are all better than RF model. The best score of XGBoost is 0.99983657 and the best score of I-GBDT is 0.99984833. In order to visually show the improvement of the model's superiority, the graph of the error contrast fluctuation is shown in the figure 12 and figure 13.

If the value is non-negative, it means that the prediction gap of XGBoost is larger than that of GBDT model, otherwise it means that the prediction gap of XGBoost is smaller than that of GBDT model. Obviously, GBDT is better than XGBoost model in both training and test sets.

As shown in the figure 14, figure 15, we need to examine the error of the model with different features deleted on the training set. If the value is non-negative, it means that the prediction gap of the model is larger than that of I-GBDT model, otherwise it means that the prediction gap of the model is smaller than that of I-GBDT model. Compared to the model without deleted features, the mean square error of the model is reduced by 1%. Compared to the model with one feature deleted, the mean square error of the model is reduced by 1%. Compared to the model with three features deleted, the mean square error...
of the model is reduced by 16%. Then we observe the performance on the test set.

![Comparison diagram of testing error fluctuation](image)

Fig. 16 Comparison diagram of testing error fluctuation

We gradually increase the number of regressions in the model. The results are shown in the figure 16. Apparently, as the number of regressors increases, the error of the model on the test set gradually decreases. Compared with other models, the I-GBDT model has the lowest error ratio on the test set and is more stable.

5. Conclusion
This paper innovatively proposes a hybrid feature selection method based on a comprehensive consideration of supervised and unsupervised learning concepts. This concept comprehensively considers the impact of features on results and the degree of correlation between features to screen for more important features. Empirical analysis shows that the mean square error of this model is 7% and 90% lower than the mean square error of XGBoost and RF, respectively. The mean square error of model I-GBDT is 1% lower than the mean square error of deleting one feature model. The mean square error of model I-GBDT is 16% lower than the mean square error of deleting three features model. Obviously, this way of selecting features is more effective.

Acknowledgments
This study is particularly grateful for the support of the national natural science foundation of China (project no. 71701028), the humanities and social sciences foundation of the ministry of education (project no. 15yjczh105), the chongqing college students innovation training program (project no. 201810618007) and the college students entrepreneurship fund.

References
[1] Guo, M.Q. (2019) A Summary of Research on Home Mortgage Loans at Home and Abroad—Based on CtieSpace Word Frequency Analysis. Fujian Quality Management, 19: 106-108.
[2] Chen, J.Q., Yang, J.J., Lou, Z.Z. (2019) Research on new short-term load forecasting model based on XGBoost algorithm. Electric Measurement & Instrumentation, 56 (21): 23-29.
[3] Chen, C. (2018) Research on the improvement of the real estate valuation market comparison method——Taking Nanjing Jindingwan Garden XX as an example. Shandong: Qingdao University of Science and Technology.
[4] Shi, D.H. (2014) Application of improved ANFIS in real estate evaluation. Computer Engineering and Applications, (12): 247-251, 256.
[5] Li, J., Du, K. (2015) Application of BP neural network in batch evaluation of houses. Value Engineering, (3): 163-164.
[6] Wang, M.H., Liang, X.C. (2019) Personal Credit Evaluation Based on CPSO-XGboost. Computer Engineering and Design, 40 (07): 1891-1895.
[7] Pang, F. (2017) Application of random forest model in batch evaluation of second-hand houses. Chongqing: Chongqing Jiaotong University.
[8] Yu, D.W., Luo, Y.P. (2019) Short-term power load forecast based on GBDT-Ridge. New Industrialization, 9 (6): 23-26,33.
[9] Yu, B.W., Xia, Y. (2016) Research on batch evaluation of linear regression based on AHP. Science and Technology Plaza, (05): 128-133.
[10] Gao, C., Xu, H.L. (2018) Unbalanced text classification method based on support vector machine. Modern Electronics Technology, 41 (15): 183-186.
[11] Yao, Q., Lin, M.T., Lan, Q.H., Huang, Z.W., Zheng, Y.W., Jiang, X., Zhu, Y.D., Kou, L.F., Xu, H.L., and Zhao, Y.z., (2020) In vitro and in vivo appraisal of didymin cyclodextrin inclusion complexes: characterization and chemosensitization activity. Drug delivery, 27(1).
[12] Hou, Z., Li, C.C., Liu, Y.C., Zhang, M., Wang, Y.T., Fan, Z.F., Guo, C., Lin, B., Liu, Y.. (2020) Design, synthesis and biological appraisal of carbohydrate-based sulphonamide derivatives as topical antiglaucoma agents through selective inhibition of carbonic anhydrase II. Journal of enzyme inhibition and medicinal chemistry, 35(1).
[13] Marko, G., Aljoša, B. (2020) Appraisal of the paraoxonase-1 kinetic parameters of the lactonase activity by nonlinear fit of progress curves. Journal of enzyme inhibition and medicinal chemistry, 35(1).