Multi-model Combination Housing Price Forecasting Based on Web Search Data

Tian-bao XIE, Jia-ling GAO and Meng ZHAO

No. 58 Yanxiang Road, Yanta District, Xi’an City, Shanxi Province, China

*Corresponding author

Keywords: Baidu index, Housing price forecast, Gradient boosting decision tree, Model combination.

Abstract. In the background of large data, taking the housing price in Xi’an as an example, this paper extracts the relevant search words data from Baidu Index and builds the neural network, support vector machine and random forests model, according to the purchase information inquired by the consumers in Baidu. The gradient boosting decision tree model is established by forecast results of the three single model to achieve a combination of housing prices forecast to determine the final residential sales price index forecast. The result shows that the fitting degree of the combined model is 0.995, and the prediction accuracy is 10.4% higher than the optimal single forecasting model. This method can calculate the new commodity housing price index two weeks in advance than the National Bureau of Statistics. In future studies, this approach will be applied to many types of market forecasts to help companies and consumers make the best decisions.

Introduction

The real estate industry is an important impetus to the development of the national economy. Researchers from all walks of life have paid more and more attention to the study of the residential sale price. How to realize the forecasting of residential sales price with foresight and timeliness is a hotspot[1]. Traditional housing prices are usually based on economic indicators and use a single model of quantitative forecasting methods, such as time series model, exponential smoothing model, gray forecasting model and regression analysis. However, there is a common flaw in the traditional method of price forecasting, which is the hysteresis of the original data. With the advent of the era of large data, more and more research focus on the network search data which is a huge number and easy access. The search data which the search engine left are real-time, so they can solves the problem of data lag. Ginsberg[2] based on the Google index of keyword search volume, analysis of the key words and the relationship between the amount of influenza treatment, the establishment of the future trend of flu regression prediction model and achieved good predictive effect. Lynn Wu[3] found that the use of Google's index of housing prices related to the keyword search index can predict the future sales of residential sales market and sales prices. Based on the Google index, Yang Shuxin[4] the study found that the residential sales price index was the highest correlation with the related keyword search volume five months ahead of schedule. Dong Qian[5] using the Baidu Index where 16 cities in the new residential and secondary residential research, and the establishment of forecasting model, found that each city's optimal forecasting model is not the same.

Although the Web search keyword data can make up for the problem of data lag in the traditional house price forecasting, any single model cannot make the best forecast for all cities from the existing research. So, based on the network search data, this paper uses a gradient iterative decision tree model to fuse the results of a single model to obtain a better prediction value.
The Architecture of Housing Price Combination Forecasting

Multi-model combination refers to the use of some of the same or different individual learning machine learning, and then use a method to learn the results of the machine to be integrated to form an optimal result[6]. In general, multi-model combination will significantly improve the generalization ability and accuracy of the prediction model. The prediction accuracy of the combined model depends mainly on the prediction accuracy of a single prediction model. Therefore, selecting a single prediction model with good prediction accuracy is the key to building a combined forecasting model. In the single prediction model, we need to consider three aspects: accuracy, correlation and quantity. Through experiments and literature review, the most commonly used and effective neural networks (ANN), support vector machines (SVM), and random forests (RF) are used in the single prediction model. Gradient Boosting Decision Tree (GBDT) is the best one. Combination forecasting architecture can be divided into three parts, as shown in Figure 1.

First of all, based on the network search keywords to determine the indicators. Before the modeling, we need to select the index of network search keywords related to house price fluctuation, and select the high correlation index. Then, a single prediction model is established and optimized. Using the original training data to train a single forecasting model, and select the optimal parameters to optimize, with the built the best single prediction model output housing price index. Finally, a combined iterative decision tree model is proposed. The results of the single prediction model are fused by the gradient iterative decision tree to obtain the final combination forecasting result.

Multi-model Combination Price Forecasting Based on Baidu Index

Determination of Keyword Index Based on Baidu Index

Figure 2. The framework of the relationship between search keywords and housing prices.
The relationship between search keywords and housing prices is shown in Figure 2, the macroeconomic factors affect the real estate market supply and demand, the real estate market supply and demand determine the residential price\(^7\). Real estate developers and buyers are the microcosmic main body of the real estate market. Performance in the real estate market is mainly residential sales and price changes, the performance of the Internet is the network search, browse the web and other indicators of change. First of all, when the investment and consumer demand, the developers and buyers will use the search engine to collect a lot of information; then, after the full collection of information, developers and consumers will make decisions that affect the real estate market supply and demand, thereby affecting the residential sales price.

Combined with the conceptual framework of the search keywords and housing prices, we will complete the initial selection of Baidu Index keywords, but there are a lot of key words in the primary selection, and there is some subjectivity in the selection method, which leads to the fact that the primary keywords are not necessarily closely related to the housing price, so we need to select the keywords from the correlation. The Spearman rank correlation test tests whether the variables are correlated by whether the two variables are synchronous or not. For \(n\) pairs of observed data shown as \((x_i, y_i)\), they are arranged in order from small to large. If the observed values are the same, then take the average rank and test whether the two variables are related. The calculation formula is shown in equation (1), which \(r_s\) indicates the degree of correlation and \(R_i\) and \(Q_i\) respectively expressed rank order.

\[
r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2
\]

Because the impact of investment and consumption behaviors of the market players on housing prices is lagging behind, the changes of the search data left on the search engine are immediate, resulting in the search keyword data and housing prices are different between the time lag\(^8\), so it is necessary to use the time difference correlation analysis method to filter out the first search keywords. The time difference correlation analysis is calculated as equation (2). Set the time series \(y\) as the benchmark index, the time series \(x\) is the analysis index, \(r\) is the time difference correlation coefficient, \(l\) indicates the number of lead lag which is called the number of time difference, \(l < 0\) indicates ahead, \(l > 0\) indicates hysteresis.

\[
r_l = \frac{\sum_{i=1}^{n_l} (x_{i-l} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n_l} (x_{i-l} - \bar{x})^2 \sum_{i=1}^{n_l} (y_i - \bar{y})^2} }, \quad l = 0, \pm 1, \pm 2, \cdots, \pm I
\]

**Establish and Optimize the Sub-forecasting Models**

The main influencing parameters of the neural network are hidden layer number, hidden layer node number and iteration number. It is generally believed that increasing the number of hidden layers can reduce the training error to a certain extent. Determining the number of hidden nodes is the key to building the model. The selection of the number of neurons in hidden layer often requires several training to determine the optimal number of neurons, equation (3) can be used to determine the number range of neurons. And then need to calculate the average absolute error of the model under different number of nodes to determine the optimal number of nodes, increase the number of iterations can reduce the prediction error, but too many iterations will lead to over-fitting.

\[
N = \sqrt{n + m + a}
\]

The main influence parameters of support vector machine are: kernel function, parameter gamma, and penalty factor C. SVM kernel functions include linear kernel function, polynomial kernel
function, radial Gaussian kernel function and sigmoid kernel function. The choice of kernel function
determines the accuracy of SVM mapping in high-dimensional space, while the determination of its
parameter gamma will affect the prediction accuracy, and the penalty factor C controls the complexity
of the SVM and the training error, if C is too large will appear "over learning", and too small will
appear "less learning".

The main influence parameters of random forest are: the number of variables pre-selected by tree
nodes and the number of trees in the forest. The number of pre-selected tree nodes determines the size
of a single decision tree, the number of trees in the forest determines the overall size of the whole
forest.

**Combination Forecasting**

Gradient iterative decision tree is a combination of Boosting and decision tree, which is a strong
learning of weak learner. The so-called Gradient Boosting is a framework for nesting a variety of
different algorithms. Boosting refers to an iterative process, the Boosting iterative sample is a
re-sampled sample, the Gradient Boosting iterates over the previous tree-optimized target, and a new
model is established in the gradient direction where the residual is reduced. The implementation of
the gradient iteration regression as follows:

<table>
<thead>
<tr>
<th>Algorithm 1: Gradient iterative implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: (\text{DATASET}D = (x_1, y_1), \ldots, (x_n, y_n)), Parameters: (L(y, f(x)), M)</td>
</tr>
<tr>
<td>Initialization: (f_0(x) = \arg \min_c \sum_i L(y_i, c))</td>
</tr>
<tr>
<td>for (m = 1) to (M) do</td>
</tr>
<tr>
<td>for (i = 1) to (N) do</td>
</tr>
<tr>
<td>(r_{mi} = -\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}</td>
</tr>
<tr>
<td>for (j = 1) to (J) do</td>
</tr>
<tr>
<td>(c_{mj} = \arg \min_c \sum_{x \in R_{mj}} L(y_i, f_{m-1}(x) + c))</td>
</tr>
<tr>
<td>(f_m(x) = f_{m-1}(x) + \sum_{j=1}^J c_{mj} I(x \in R_{mj}))</td>
</tr>
<tr>
<td>end for</td>
</tr>
<tr>
<td>end for</td>
</tr>
<tr>
<td>return: (\hat{f}(x) = f_M(x) = \sum_{m=1}^M \sum_{j=1}^J c_{mj} I(x \in R_{mj}))</td>
</tr>
</tbody>
</table>

In the establishment of GBDT combination forecasting model, the Boosting framework parameters
and weak learner parameters need to be optimized. The important parameters in the Boosting
framework parameters are: the maximum number of iterations (n-estimators), the learning (step size)
and the loss function of the weak learner (loss). The important parameters of the weak learner are the
max-depth of the regression tree and the max-features considered in the partitioning.

**Forecast and Analysis on the Combination forecasting of Xi’an New Commercial Housing
Price**

**Data Sources and Pretreatment**

The dependent variable of this experiment is the Xi’an new commodity housing price sales, which is
proportional to the sales price of new commercial residential buildings in Xi’an, and reflects the
fluctuation of sales price of newly built commercial residential buildings in Xi’an. The independent variables are the Web search keywords index which is related to Xi’an new commodity housing price sales.

From January 2011, the National Bureau of Statistics released in mid-monthly residential sales price index of large and medium cities. As the year-on-year data better reflect the changes in housing prices and search index over time, this paper selected year-on-year data.

Search keyword search volume from the Baidu Index, this paper converted the daily search data into monthly search data. At the same time, in order to avoid the uncertain factors that make the keyword search volume abnormal fluctuations, we make the monthly data to three moving average. In addition, in order to maintain consistency with the housing price statistics, we make monthly data to year-on-year treatment. Therefore, the data in this paper is from Jan. 2012 to Dec. 2015, which from Jan. 2012 to Apr. 2015 is training set and from May 2015 to Dec. 2015 is test set.

**Determination of Keyword Indicator**

The initial search index is determined by considering the duality of housing and based on the conceptual framework diagram of the search keywords and housing (Figure 1). When residential properties are consumer goods, for the future living environment, consumers will be more concerned about the following factors: residential price movements, residential units, location, mortgage interest rates, real estate and property companies. When residential properties are investment products, the investors from the residential rental and real estate market prices to profit, so they will be more concerned about housing prices, real estate traffic environment and market interest rates. Based on the above factors, this paper uses the subjective method to select the core keywords and the use the second search, semantic extension and long-tailed mining method to derivative the derived keywords. The initial keyword thesaurus contains 7 core keywords and 56 derived keywords, as shown in Table 1.

<table>
<thead>
<tr>
<th>Core keywords</th>
<th>Derivative keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Buying houses</strong></td>
<td>The policy of buying houses, matter of buying houses, the searching houses network of Xi’an, the rental network of Xi’an, the buying houses network of Xi’an, mortgage purchase, housing sales contracts, housing sales agreement, housing sales taxes and fees</td>
</tr>
<tr>
<td><strong>Housing price</strong></td>
<td>Housing price in Xi’an, the trend of Xi’an housing price, the future trend of Chinese housing price, the trend chart of housing price, the housing price of Xi’an</td>
</tr>
<tr>
<td><strong>Real estate</strong></td>
<td>Xi’an real estate, Xi’an real estate information network, Xi’an Vanke City, Xi’an real estate information network, Xi’an real estate network, the Xi’an housing authority, Xi’an real estate, housing resources, real estate brokerage firm, building taxes, real estate trading center, real estate transaction network</td>
</tr>
<tr>
<td><strong>Properties for sale</strong></td>
<td>Xi’an properties for sale, the property market in Xi’an, Washington, Xi’an metro map, real estate network, market trend</td>
</tr>
<tr>
<td><strong>Apartment Layout</strong></td>
<td>Small size decoration, size chart, decoration materials, the step of spruce up, Xi’an decoration, building materials</td>
</tr>
<tr>
<td><strong>mortgage</strong></td>
<td>Mortgage interest rates, loans, the latest mortgage interest rates, mortgage calculators, mortgage rate calculator, mortgage interest rate, building taxes, new tax of house, how to calculate the real estate tax, estate deed tax</td>
</tr>
<tr>
<td><strong>provident fund</strong></td>
<td>Individual housing provident fund inquiries, housing provident fund, housing provident fund inquiries, provident fund load amount, Xi’an provident fund, Xi’an provident fund inquiries, provident fund inquiries, Xi’an housing rental</td>
</tr>
</tbody>
</table>

Then, we use the eq.1 to calculated the Spearman rank correlation coefficient and the keywords which the correlation coefficient greater than 0.5 are performed the significance test. Using eq.2 calculate the lead time of the remaining 22 keywords. The remaining 13 first do significance test and p values are 0. The final screening of the keywords as shown in Table 2.
Table 2. Select keyword indicators.

<table>
<thead>
<tr>
<th>keywords</th>
<th>Correlation coefficient</th>
<th>P-value</th>
<th>keywords</th>
<th>Correlation coefficient</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>the rental network of Xi’an</td>
<td>0.55</td>
<td>0</td>
<td>Xi’an decoration</td>
<td>-0.59</td>
<td>0</td>
</tr>
<tr>
<td>housing sales contracts</td>
<td>-0.69</td>
<td>0</td>
<td>the latest mortgage interest rates</td>
<td>-0.81</td>
<td>0</td>
</tr>
<tr>
<td>Housing price</td>
<td>0.62</td>
<td>0</td>
<td>housing provident fund inquiries</td>
<td>-0.73</td>
<td>0</td>
</tr>
<tr>
<td>the trend chart of housing price</td>
<td>-0.60</td>
<td>0</td>
<td>Xi’an provident fund</td>
<td>-0.59</td>
<td>0</td>
</tr>
<tr>
<td>Xi’an Vanke City</td>
<td>0.72</td>
<td>0</td>
<td>Xi’an provident fund inquiries</td>
<td>-0.65</td>
<td>0</td>
</tr>
<tr>
<td>Xi’an real estate network</td>
<td>-0.60</td>
<td>0</td>
<td>provident fund inquiries</td>
<td>-0.76</td>
<td>0</td>
</tr>
<tr>
<td>Xi’an properties for sale</td>
<td>-0.50</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Establishment and Optimization of Single Predictive Models**

In the construction of neural network model, single hidden layer neural network has a strong ability of non-linear mapping, so this experiment uses a single hidden layer neural network. In determining the number of hidden layer nodes, the number of neurons in the hidden layer is in the range of [3,13] and when the number of neurons is three, the prediction error is the smallest, so the optimal hidden layer number is three. It is found that after 300 iterations, the error stability is not decreased rapidly. Considering the over-fitting problem, the final analysis determines that the number of iterations determined as 300. The comparison between the results of the optimal neural network model and the real housing price index is shown in Figure 4 (a). It can be found that the neural network model has a poor fitting to the peak value and the over-fitting problem leads to the prediction of the test sample. The deviation of the actual value is large.

The optimal absolute parameters are calculated by using the exhaustive method. The optimal parameters are as follows: the kernel parameters are radial, gamma is 0.25, and the penalty factor C is 1. As shown in Fig. 4 (b), it is found that the generalization of the model is better than that of the neural network, but the forecasting bias is still relatively large. At the same time, because of the lack of sensitivity to the data, the degree of fit to the peak is also insufficient.

When the random forest model is constructed, the number of variables selected by the decision tree node branch is determined by adding the variables one by one. When the three variables are selected for branching, the model explains the most variables and the least squares means, so the optimal model mtry = 3. After determining the optimal number of nodes in the decision tree in the random forest model, it is necessary to further determine the number of decision trees in the model. In this paper, we use visual analysis to determine the number of decision trees. When the number of decision tree is greater than 800, the error of random forest model tends to be stable, so the number of decision tree is 800. As shown in Fig. 4 (c), both the random forest generalization ability and the fitting to the peak value are optimized, but when the noise is large, the random forest will produce over the problem of fitting, stability is poor.

**Implementation of GBDT Model**

After a single prediction model is built, the predicted value of the single prediction model for the training sample is used as the training sample of the GBDT combination. Since this experiment is a regression problem, the most commonly used mean square error: \( \phi(y, f) = \frac{(y - f(x))^2}{2} \). Other parameters use default values. The GBDT combination model was established and the test set was predicted. The fitting degree of the test set was 0.943.

Next, the GBDT combinatorial model will be optimized. First, the step size and the iteration number are optimized. The step size is [0,1], the step size and the iteration number are the best decision factors. Generally, the more iterations are required and the smaller the step size is to 0.1 and the number of iterations is optimized. The results are shown in Figure 3. It is found that the loss
function does not decrease significantly and the region is stable after more than 50 iterations. Therefore, the number of iterations is determined to be 60.

Figure 3. Relation between iteration number and loss function.

After determining the appropriate number of iterations, the decision tree will be adjusted parameters. The parameters max-depth and max-features in the decision tree are more suitable for the experiments with eigenvalues less than 50. Because the eigenvalues of this experiment are 13, the default values are not to limit the depth of subtree and give full consideration to the number of features. In summary, the optimal parameter in GBDT is set: $n_{-estimator} = 60$, $learning = 0.1$, $max_{-features} = None$. The fitting degree of the training set is 0.995, which is 5.5% higher than that of the GBDT model with all the default values.

**Comparative Analysis of Combination Forecasting Model and Single Forecasting Models**

After the Neural Network, Support Vector Machine, Random Forest and GBDT combinatorial model are built, the forecasting ability of them will be compared as shown in Figure 4. Using the data extrapolated from the first 40 periods, after the vertical line for the test set that is extrapolated results. It can be seen from Fig. 4 (d) that the predicted value of GBDT combination model is in good agreement with the actual value, while the other three sub models have a large gap between the predicted value and the actual value, especially the Neural Network model.

In order to compare the prediction ability of the model, the normalized mean squared error (NMSE) and mean square error (MSE) are used to test the fit and stability of the fusion model and three single models. The smaller the value of NMSE, the fitting degree of the model is better. The smaller the value of MSE, the stability of the model is better. The results are shown in Table 3 and it can be seen that the NMSE and MSE of the GBDT combined model are the minimum, especially in the test set, which indicates that the prediction ability of the GBDT fusion model is significantly better than that of the single prediction model. The fit degree of the GBDT fusion model is 0.995, while the fitting degree of the random forest which is the best one among the three sub-models is 0.901. The prediction accuracy of GBDT combined model is improved by 10.4%.

Figure 4. Comparison of predicted and actual values for each model.
Table 3. Model performance comparison results.

<table>
<thead>
<tr>
<th>Models</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NMSE</td>
<td>MSE</td>
</tr>
<tr>
<td>ANNs</td>
<td>0.02</td>
<td>0.55</td>
</tr>
<tr>
<td>SVM</td>
<td>0.04</td>
<td>0.99</td>
</tr>
<tr>
<td>RF</td>
<td>0.02</td>
<td>0.43</td>
</tr>
<tr>
<td>GBDT</td>
<td>0.01</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Conclusion

The innovation of this paper is to use the method of combination forecast to forecast the price of house in order to solve the problem that the single forecasting models cannot make the best forecast for housing price of all cities. The experimental results show that the predictive ability of the combined model is much better than the single predictive models. The fitting degree of the test set is 0.995, which is 10.4% higher than that of the single optimal model. It indicates that the single model can achieve the complementary effect of the advantages and disadvantages by combining forecasting to get the optimal value. The research ideas and methods of this paper can be extended to other official statistics such as transaction volume of newly built commercial housing, price and turnover of second-hand house, consumer price index, resident income, unemployment rate.

References


