Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231031

Birth Rate Prediction System Based on Grey Prediction and Gradient Boosting Regression Tree Ensemble Model

Weibo Huang[1], Fei Lin, Yijia Qin and Li Ling
Guangdong University of Foreign Studies, Guangzhou, China.

Abstract. We selected several factors affecting the Birth Rate in Guangdong Province. Then we saw the contribution of each characteristic to the Birth Rate by the gradient boosting regression tree model. The grey prediction algorithm used each characteristic to predict its value in a certain year. Using variance ratio and small residual probability, we evaluated the prediction accuracy. We used the gradient boosting regression tree for predicting the Birth Rate in the next two years. This was done under the premise of knowing the data of each characteristic. The thinking of the factors affecting the Birth Rate and the prediction of the Birth Rate prompt us to think about the relevant realistic factors. They also assist the government in focusing on the adjustment of fertility policy. The aim is to promote population development and social progress to the greatest extent possible.

Keywords. Grey prediction, gradient boosting regression tree, combination model, birth rate prediction

1. Introduction

The development mode of different regions is different, different time stages will have different effects, and the change trend of Birth Rate is also different. At present, the research on the Birth Rate is mainly based on national theoretical and practical analysis. But the research on the analysis and prediction of the Birth Rate of each province need to be further studied. This paper takes the relevant data of Guangdong Province as an example to explore and analyze from a more microscopic perspective. Secondly, the current research on the Birth Rate mostly focuses on fertility desire and fertility behavior. It lacks a comprehensive analysis of more dimensions such as economy and education level.[1][2] While realizing the Birth Rate prediction, this work also demonstrates the contribution of each feature to the birth rate. It further analyzes the main factors that cause the birth rate change in multiple dimensions.

This paper analyzes data from the Guangdong Statistical Yearbook. To forecast the birth rate, we first identify the key factors that may influence the birth rate in Guangdong Province. We select these influencing factors and screen out nine factors that have a significant impact on the birth rate. Next, we employ the grey prediction algorithm to forecast the values of these nine indicators for the next two years. Finally, we use the

[1] Corresponding Author, Huang Weibo, Modern Education Technology Center, Experimental Teaching Center, Guangdong University of Foreign Studies, China; This study is financially supported by the 2021 Guangdong Higher Education Teaching Research and Reform Project. Email: hwb444@163.com

gradient boosting regression tree model with the parameter values to predict the birth rate of Guangdong Province for the next two years. However, we observed that the predictions provided by the random forest model, KNN, and SVR were inaccurate in forecasting the birth rate.[3]

2. Method design

Firstly, the system used gradient boosting regression tree for feature selection.[4] We screened out nine influencing factors that made a greater contribution to the birth rate. Secondly, we used the combination model of grey prediction and gradient boosting regression tree to predict the birth rate. The following will show the data indicators selected by the study and the required algorithms.

2.1. Indicator description

This paper selected the Birth Rate from 2000 to 2021 and 13 representative factors that may affect the Birth Rate in Guangdong Province. It included the Proportion of urban population, the Proportion of education level, the Average Household Size, the total Population of Household Registration at the end of the year, the total burden coefficient, the Consumer Price Index, the Retail Price Index, the Per Capita Disposable Income of all residents, the Local Fiscal Tax Revenue, the Sex Ratio, the Household Consumption Expenditure, the Medical Bed and the Medical and Health Institutions.

2.2. Gradient boosting regression tree feature selection

Feature selection is a process that involves selecting a subset of features from the original data. This selection aims to improve the model's performance and efficiency. The gradient boosting regression tree is an ensemble learning algorithm that is based on decision trees. This algorithm utilizes specific principles for feature selection, primarily focusing on two aspects. Feature selection is highly significant in gradient boosting regression tree. It helps improve the generalization ability and interpretability of the model. Additionally, it reduces the time and resource consumption during model training. The feature selection steps in the gradient boosting regression tree algorithm are as follows.

- To determine the importance of each feature to the target variable, there are two measurement approaches. One approach is to count the number of feature splits across all trees. The other approach involves calculating the average error reduction achieved through feature splitting.
- The model selects the first N features as input features, based on the order of feature importance.

The following is the derivation of the principle equation of feature selection.

In the t-th tree, the goal is to minimize the loss function $\mathrm{L}(y, f_t(x))$. The target variable y and the input feature x are involved in the process. To compute the loss function, the predicted $f_t(x)$ of the previous t-1 tree is added to the predicted value of the t-th tree. The training process of the gradient boosting tree utilizes the addition model, specifically described by Equation (1).

$$f_t(x) = f_{t-1}(x) + h_t(x) \quad (1)$$

$f_{t-1}(x)$ is the predicted value of the first t-1 tree and is the predicted value of the t-th tree. In order to minimize the loss function, we need to find the best. The model selects the first N features as input features, based on the order of feature importance.

2.3. Grey prediction algorithm

Grey prediction algorithm is a method to predict the system with uncertain factors. It is generally applicable to time series occasions.[5] We employed the Grey Prediction Algorithm GM(1,1) to predict the values of various indicators for the years 2022 and 2023. The following is the elaboration of prediction principle.

We Assumed that the feature $X^{(0)} = \{X^{(0)}(i), i = 1,2,\cdots,n\}$ was a non-negative original data sequence. Below, we showed the process of constructing the Grey Prediction Model.[6]

- First, we accumulated $< math >$ and obtained an accumulation sequence.
- We established the following first-order differential equation for $< math >$. That was GM (1,1) model, such as Equation (2).

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = \mu \quad (2)$$

- By solving differential equations, we obtained the predictive model, as in Equation (3).

$$\hat{X}^{(1)}(k+1) = [X^{(0)}(1) - \frac{\mu}{a}]e^{-ak} + \frac{\mu}{a} \quad (3)$$

- The GM (1,1) model was a cumulative amount. We simplified the data obtained from the GM (1,1) model as $\hat{X}^{(0)}(k+1)$. That was the grey prediction of $\hat{X}^{(0)}(k+1)$, as in Equation (4).

$$\hat{X}^{(0)}(k+1) = (e^{-\hat{a}} - 1)[X^{(0)}(n) - \frac{\hat{\mu}}{\hat{a}}]e^{-\hat{a}k} \quad (4)$$

2.4. Gradient boosting regression tree algorithm

The gradient boosting regression tree algorithm was an ensemble learning method based on decision trees. It performed model fitting by gradually learning residuals through a gradual process. The algorithm combined multiple decision trees, with each decision tree optimizing the error of the previous tree. Eventually, it obtained an integrated model with a reduced prediction error.[7][8]

The training process of the gradient boosting regression tree model proceeded as follows.

We denoted the number of decision trees as M. And we designated the gradient boosting regression tree model for the final output as $f_{(M)}(x_i)$.

- Initialization

We created the first regression tree $f_{(1)}(x_i)$, as Equation (5).

$$f_{(1)}(x) = arg_c min \sum_{i=1}^{N} L(y_i, c) \tag{5}$$

- Iteration

For the second to the m-th regression tree, we used the residual of the previous result, as shown in Equation (6).

$$r_{mi} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}] f(x) = f_{m-1}(x) \tag{6}$$

For the current m-th subtree, we traversed its feasible cut points and thresholds. We found the parameters corresponding to the optimal predicted value c. So it approached the residual as much as possible, as shown in Equation (7).

$$c_{mj} = arg_c min \sum_{x_i \in R_{mj}} L(y_i, f_{m\_1}(x_i) + c \tag{7}$$

R_{mj} referred to the collection of predicted values of the leaf nodes in the m-th subtree. It was obtained using all the partitioning methods. The collection represented the predicted values of the leaf nodes in the m-th subtree. The range of j was $\{1,2,\cdots,J\}$.

We renewed the m-th regression tree according to the following Equation (8).

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right) \tag{8}$$

I was a function. We set I to 1 if the sample fell on the node; otherwise, we set I to 0.

- Finally, we obtained the regression tree, such as Equation (9).

$$F(x) = f_M(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right) \tag{9}$$

The advantages of this algorithm were its ability to handle high-dimensional, nonlinear, and non-stationary data. It could also adaptively add new decision trees.

3. Results and discussion

In the implementation process of the system, we designed and adopted a combination model of grey prediction and gradient boosting regression tree. Feature selection model based on gradient boosting regression tree had a strong selection effect. It could select nine features with the highest correlation with the birth rate. Grey prediction demonstrates excellent performance in predicting small amounts of data. We have established a grey prediction model for the selected individual factor. We used this model to obtain the forecast values for the years 2022 and 2023.[9] Gradient boosting regression tree has the strong applicability and fault tolerance. We established a training model for historical data. This model was developed based on its robustness and versatility.[10] We substituted the data results of grey prediction into the trained model. By obtaining

more accurate prediction results, we indicated a higher level of prediction accuracy. This provided further evidence of the model's feasibility in delivering precise predictions. The following is a step-by-step description of the combination model of grey prediction and gradient boosting regression tree.

3.1. Data preprocessing results

Feature selection aimed to select the most representative feature subset from the original data. This improved the performance and efficiency of the model by enhancing its discriminative ability. Feature selection can reduce irrelevant features. It can also enhance the generalization ability and interpretability of the model. Moreover, it can also decrease the complexity and training time of the model. We initially employed the gradient boosting regression tree to train the 13 features gathered for the collected data. This training aimed to assess their potential impact on the birth rate. Subsequently, we utilized an interface to assess the contribution of each feature to the birth rate. Based on our observations, we could clearly see that the total population at the end of the year and the education level had a significant impact on the birth rate. These features played a crucial role in influencing birth rates. On the other hand, we found many features have a relatively small influence on the birth rate. We ranked the importance scores of each feature from high to low. And we selected the top 9 features for further analysis.

3.2. Model construction and evaluation

3.2.1. Construction and evaluation of grey prediction model

We based the prediction of the birth rate on 9 selected features through feature selection. The data for 2022 and 2023 were unknown. We needed to first predict the values of these 9 features that affected the birth rate. Here, we could use the gray prediction algorithm GM (1,1) model.

Finally, we predicted that the birth rate in Guangdong Province in 2022 would be 8.904‰ and in 2023 would be 8.741‰, as shown in **Figure 1**.

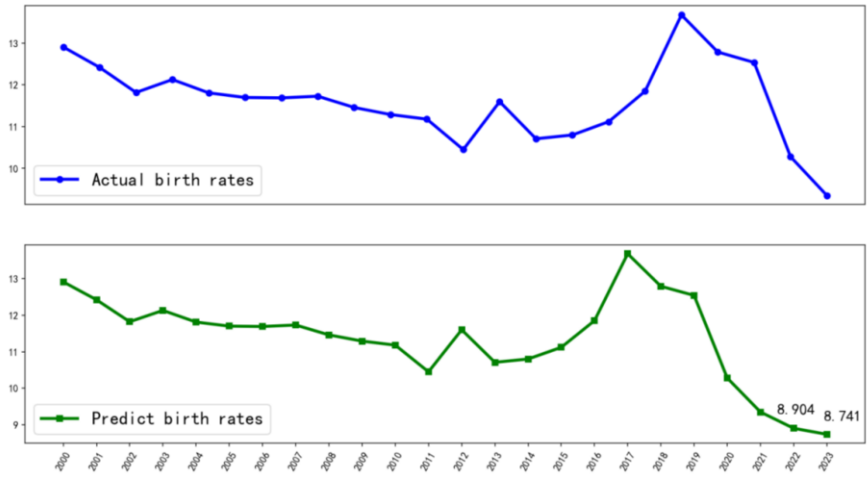

Figure 1. Actual birth rates and predict birth rates.

Calculate the root-mean-square error of the model, the mean absolute error, R^2, as Equation (10) to Equation (12).

$$\mathrm{MAE} = \frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\mathrm{y_i} - \mathrm{y'_i}\right| \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left((\mathrm{y_i} - \mathrm{y'_i})^2\right)} \tag{11}$$

$$\mathrm{R^2} = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}})^2} \tag{12}$$

Table 1. Precision index of Gradient boosting regression tree prediction algorithm

Mean Absolute Error (MAE)	Mean Square Error(MSE)	Root Mean Squared Error(RMSE)	R^2
0.001	4.279	0.001	0.99

From **Figure 1**, we can see that the predicted birth rate values and actual birth rate values have shown a decreasing trend over the past 20 years. The birth rate in Guangdong Province increased in 2016 and reached its highest value in 2017. This was due to the "two-child" policy implemented by the government in 2016. From 2019 to 2021, the birth rate in Guangdong Province has significantly declined, largely due to the impact of COVID-19. Assuming that there are no major social upheavals in the coming year, we can estimate that the birth rate. **Table 1** also reflects that the model has a good fit, and can explain more variability.

3.3. Results and discussion

We wrote the code in the Python language and mainly utilized the pandas and sklearn libraries, making it highly portable. We used all of the data in this paper from the Guangdong Statistical Yearbook, ensuring its reliability and validity. The entire testing program ran smoothly. We used a combination of mathematical models and machine learning algorithms for prediction. We integrated the Grey Prediction Algorithm with the Gradient Boosting Regression Tree Algorithm. By doing so, we obtained an improved data prediction model, thereby enhancing the experimental accuracy. At the same time, this solution also optimized the complexity of the algorithm. And it reduced unnecessary function parameters. The combination of the grey prediction and gradient boosting regression tree models played a mutually corrective role. Both models themselves are suitable for short-term forecasting with small sample sizes. They could effectively predict the birth rate. This model has high prediction accuracy and stability, and is reliable and effective.

Table 2. Comparison of prediction accuracy metrics among different models.

Model	Prediction accuracy metrics R^2
Random Forest Model	0.89
K-Nearest Neighbors	0.36
Support Vector Regression	0.77
Gradient boosting regression tree	0.99

In addition, we selected R^2 as the measure of the regression model's goodness of fit. It measured the percentage of variability in the dependent variable explained by the model. We compared the R^2 of other regression models, such as random forest regression. And we supported vector machine regression and K-Nearest neighbor regression. **Table 2** showed the comparison results. This once again demonstrated the accuracy of using gradient boosting trees.

This work utilizes a combination of the grey prediction and gradient boosting regression tree models. So it is more effective in data prediction than other similar products. This model is suitable for short-term forecasting with small sample sizes, and can effectively predict the birth rate. The model also improves the accuracy and stability of prediction values, and is reliable and effective. At the same time, these two algorithms correct each other, greatly improving the accuracy of the data. Based on this, we obtained the predicted birth rate of Guangdong Province. The birth rate in Guangdong Province in 2022 is predicted to be 8.904‰, and the birth rate in 2023 is predicted to be 8.741‰.

4. Conclusion

In this paper, we combined the Grey Prediction model with gradient boosting regression tree. We used mathematical models and machine learning algorithms to predict future birth rates. This process prompts us to consider the factors that influence birth rates and improve relevant factors. Based on our findings, we could propose recommendations to the Guangdong provincial government. This could assist them in adjusting their birth policies. They could focus on household registration population, disposable income of residents, and educational level of residents.

The mathematical model designed and optimized in this study achieved good data prediction results. It improved the accuracy of the experiment. It can be applied to the analysis of birth rates and influencing factors in provinces across the country and even in other countries.

References

[1] Kearney Melissa S.,Levine Phillip B.,Pardue Luke. The Puzzle of Falling US Birth Rates since the Great Recession. Journal of Economic Perspectives.2022 Feb; 36(1): 151-76.
[2] Furtunov S.,Ruseva J.,Madjova V.. Analysis of some indicators and trends related to birth rates and optional abortions in Bulgaria and Europe and possibilities for their optimizing. General Medicine.2021 Nov;23(2):20-4.
[3] Yun Jungha,Kim Chae Young,Son SeHyung,Bae ChongWoo,Choi YongSung,Chung SungHoon. Birth Rate Transition in the Republic of Korea: Trends and Prospects. Journal of Korean medical science.2022 Oct; 37 (42): 304.
[4] Wang Wen Bao,Hu Yi Chung. Multivariate Grey Prediction Models for Pattern Classification Irrespective of Time Series. JOURNAL OF GREY SYSTEM.2019 Jan;31(2):135-42.
[5] Jiang Peng,Hu Yi Chung. Constructing interval models using neural networks with non-additive combinations of grey prediction models in tourism demand. Grey Systems: Theory and Application.2023 Jan;13 (1):58-77.
[6] Qiuping Wang,Subing Liu,Haixia Yan. The application of trigonometric grey prediction model to average per capita natural gas consumption of households in China. Grey Systems: Theory and Application.2019 Feb; 9 (1):19-30.
[7] Di Persio Luca,Fraccarolo Nicola. Energy Consumption Forecasts by Gradient Boosting Regression Trees. Mathematics.2023 Feb; 11 (5):1068.
[8] Park Soyoung,Jung Solyoung,Lee Jaegul,Hur Jin. A Short-Term Forecasting of Wind Power Outputs Based on Gradient Boosting Regression Tree Algorithms. Energies.2023 Jan; 16(3):1132.
[9] Xuejun Shen,Minghui Yue,Pengfei Duan,Guihai Wu,Xuerui Tan. Application of grey prediction model to the prediction of medical consumables consumption. Grey Systems: Theory and Application.2019 Apr;9(2):213-23.
[10] Shatnawi Amjed,Alkassar Hana Mahmood,AlAbdaly Nadia Moneem,AlHamdany Emadaldeen A.,Bernardo Luís Filipe Almeida,Imran Hamza. Shear Strength Prediction of Slender Steel Fiber Reinforced Concrete Beams Using a Gradient Boosting Regression Tree Method. Buildings.2022 Apr; 12 (5):550.