A novel combination machine learning model for regional GDP prediction: evidence from China

Yinghan Xia

School of Economics, Zhejiang University, Zhejiang, China
E-mail : XiayinghanZJ@126.com

Abstract: In recent years, the regional GDP prediction has become an efficient tool to coordinate economic development. This paper aims to study the regional GDP prediction and build a novel machine learning model with taking the entropy method into consideration to predict the future GDP values. This research uses the entropy method to calculate weights of the linear regression model and XGBoost regression model, then build a novel combination model to predict the GDP value of different regions. The model will combine the advantages of trend prediction from linear regression models and high fitting accuracy from XGBoost regression models. In the empirical analysis, the paper used the China's GDP values of different provinces in mainland and built the novel combination model with entropy method, linear regression model and XGBoost regression model. The results reveal that the proposed novel combination model outperforms the tow base models on the mean absolute percentage error evaluation metric.

Keywords: Regional GDP; Linear Regression; XGBoost Regression; Entropy method

1 Introduction

China, as a prominent global economy, has undergone substantial economic expansion throughout the past few decades. And the growth of regional GDP (Gross Domestic Product) within China exhibits disparities across provinces and regions due to variations in natural resources, levels of economic development, and governing policies. Generally, the eastern coastal provinces, namely Shanghai, Guangdong, Jiangsu, and Shenzhen, have witnessed remarkable economic progress, primarily attributed to their robust manufacturing and export sectors. In contrast, the western provinces, including Xinjiang, Tibet, and Qinghai, lag behind in economic development due to their remote geographical locations and inadequate infrastructure. In recent years, China's GDP growth rate has decelerated due to several factors, such as an aging population, escalating labor costs, and trade tensions with the United States[1]. Consequently, analyzing the diverse patterns of regional GDP growth assumes critical importance for the overall growth trajectory of the nation's GDP.

Accurately predicting regional GDP relies on several factors that can influence the precision of GDP forecasts, including data quality and the assumptions underlying the

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_101

models. Nonetheless, GDP predictions play a vital role for policymakers, businesses, and investors in making well-informed decisions regarding economic planning, investments, and risk management. In China, the government establishes official GDP growth targets annually as part of its Five-Year Plans, which guide the nation's economic and social advancement. These targets shape policy decisions, investment priorities, and serve as benchmarks for evaluating the performance of government officials. However, achieving high growth rates at the provincial level is also of paramount importance for economic development[2]. This paper focuses on the study of a combination model for predicting regional GDP values, aiding policymakers in designing effective policies and coordinating regional economic development.

2 Literature Review

Government departments typically adjust market economic policies and formulate plans to promote regional economic development based on economic forecasts for the upcoming year or even several years, with GDP forecasts playing a crucial role. Local governments, such as those in different provinces of China, formulate and adapt policies based on their own economic forecasts. Economic data forecasts often rely on historical data and statistical methods to predict future economic values. GDP forecasting, as a fundamental component of economic data, primarily utilizes historical data from various regions and employs statistical or machine learning models for prediction. Therefore, time series forecasting models are widely used in GDP forecasting. Some scholars incorporate additional non-time series data elements for a comprehensive analysis, thereby constructing more accurate forecasting models. For instance, Elkhan Richard[3] incorporated national income, environmental factors, and other data into historical time series data to build a new non-linear regression model. Some scholars opt for more complex models for analysis and prediction, particularly with the advancement of machine learning and other artificial intelligence technologies. Machine learning and deep learning models have emerged as primary methods for GDP forecasting. Wu and Chen[4], for instance, achieved significant results using Support Vector Machines (SVM) for GDP forecasting.

A single prediction model often only reflects and analyzes part of the information of the object being studied. In order to improve the accuracy of predictions, researchers often optimize in two ways. One is to optimize the model parameters. Yusof[5] combined the least squares support vector machine (LSSVM) and the artificial bee colony (ABC) algorithm for comprehensive prediction, proving that it can improve prediction accuracy. Long[6] optimized the parameters of the SVM using genetic algorithms to achieve more accurate predictions of the GDP of Anhui Province, China. Another way is to use different models for combination prediction to improve accuracy. Sa'adah and Wibowo[7] used both the long short-term memory (LSTM) and recurrent neural network (RNN) models for prediction, and their accuracy was significantly improved compared to a single model. With the advancement of deep learning, more researchers are focusing on the field of neural networks, constructing multi-node and multi-layer neural network models to predict future data for each region. However, neural networks are prone

to overfitting and other issues in prediction, so a cautious attitude is still necessary in industrial practice[8]. And some scholars also provided some combined model to predict or classification, Li Yiheng[9] has revealed that the combined model based on the entropy method and other single models had a better performance than single models. Some scholars have modified and improved model performance by adding or studying other variables. For example, Han and others[10] fully considered the factor of carbon dioxide to improve the accuracy of regional GDP prediction. Dai et al.[11] compared the effects of introducing nighttime light data into linear regression, exponential, and complex artificial neural network models in GDP prediction, demonstrating the impact of nighttime light data on GDP prediction.

In conclusion, when predicting regional GDP, it is essential to consider the influence of various factors. Introducing additional variables related to population and daily life, such as water and gas usage, can enhance the model's accuracy. Historical GDP data also serves as a crucial reference for predicting future GDP, warranting careful attention. From a modeling perspective, different models exhibit varying performance in terms of trend prediction and fit. Therefore, our research objective is to leverage the strengths of different models through specific algorithms, aiming to construct a combination model that strikes a balance between trend prediction and accurate fitting. In this study, we incorporate additional life and population data from diverse regions and consider combining the advantages of the linear regression model for trend prediction with the benefits of a tree model for accurate fitting.

3 Research Methodology

3.1 Linear Regression

Linear regression is a statistical methodology employed to model the association between a dependent variable and one or more independent variables. The primary objective is to ascertain the optimal linear equation that can effectively elucidate the relationship among these variables. The formula for a simple linear regression is expressed in Formula 错误!未找到引用源。:

$$y=b\_0+b\_1*x+e(1)$$

where y is the dependent variable, x is the independent variable, b_0 is the intercept, b_1 is the slope, and e is the error term. And the formula for the multiple linear regression is shown as below:

$$y=b\_0+b\_1*x\_1+b\_2*x\_2+\cdots+b\_k*x\_k+e(2)$$

where y is the dependent variable, $x_1, x_2, \ldots, x_k$ are the independent variables, b_0 is the intercept, $b_1, b_2, \ldots, b_k$ are the slopes, and e is the error term[12].

3.2 XGBoost Regression

XGBoost, an abbreviation for Extreme Gradient Boosting, is a highly popular ensemble machine learning algorithm designed for regression tasks. It serves as an extension of the gradient boosting algorithm, employing a gradient descent optimization strategy to minimize a specified loss function. We define the XGBoost model by specifying the hyperparameters, such as the number of trees, learning rate, and maximum depth[13]. The main formula of the decision tree is the splitting function shown as below:

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i+\lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}\right] \quad (3)$$

And the g_i and h_i are shown in formula **错误!未找到引用源。**:

$$g_i = \nabla C(x_j - q)$$
$$\mathrm{h}_i = \nabla^2 C(x_j - q) \quad (4)$$

And calculate the loss of errors, the loss function formula shown as formula **错误!未找到引用源。**.

$$\sum_{i=1}^{n} L(y_i, p_i) = \frac{1}{2}(y_i - p_i)^2 \quad (5)$$

The XGBoost will uses the loss function to build trees by minimizing the below equation:

$$\sum_{i=1}^{n} L(y_i, p_i) + \frac{1}{2}\lambda o_v^2 \quad (6)$$

Where, o_v^2 is the output value. The XGBoost regression is a powerful and widely used machine learning algorithm that can handle complex nonlinear relationships between variables.

3.3 Entropy Method

Within this paper, the proposed approach for predicting next year's GDP value relies on the utilization of the entropy method. The entropy method serves as an objective weighting technique, which leverages entropy values associated with each indicator or model to determine their relative significance or weight. Entropy, stemming from probability theory, quantifies the degree of uncertainty within information. By calculating the entropy value of each indicator or model, considering the amount of information it contributes, these values are then employed as weights within the composite model for predicting GDP values.

Suppose y_{it} is the predicted value of the tth variable by the ith evaluation method, where i (i=1,2,…,m) represents different evaluation methods and t (t=1,2,…,n) is an independent variable[9]. The entropy method includes the following 4 steps:

Step 1: Calculating the proportion of error.

$$p_{it} = \frac{e_{it}}{\sum_{i=1}^{n} e_{it}} (7)$$

Step 2: Defining the entropy value of the prediction error of the evaluation method as H_i.

Step 3: Defining the coefficient.

Step 4: Calculating the weighting coefficients of the evaluation methods:

$$w_i = \frac{1}{m-1}(1 - \frac{d_i}{\sum_{i=1}^{m} d_i}) (8)$$

where $\sum_{i=1}^{m} w_i = 1$.

This study utilizes the aforementioned approach by combining the accurate trend prediction of Linear regression with the relatively precise fitting of XGBoost regression, and assigning different weights to each, thereby constructing a predictive model that incorporates the strengths of both.

4 Empirical results

In this paper, the author considered various factors to predict the GDP value of different regions. These factors include human factors, live factors, and historical values of GDP. The factors used to predict the next year's regional GDP include the natural growth rate of the population, per capita water resources, total population at the end of the year, population mortality rate, total population density, total population using gas and water, and birth rate (as described in Table 1). Additionally, the GDP value of the previous year in different regions such as Guangdong, Shandong, Tianjin, and others were also considered.

Table 1. Abbreviation of the factors

Factors	Abbreviation
human natural growth rate	ngr
per capita water resources	$pcwr$
total population at the end of the year	tp
mortality rate of a population	mrp
total population density	tpd
total population using gas	$tpgas$
population using water	$tpwater$
birth rate of a population	brp

Source: www.stats.gov.cn[14]

The author of the paper collected data from the national statistical yearbook for 31 regions in mainland China from 2004 to 2018[14]. The Python 3.10 and Jupyter software was used to analyze the dataset and processed the raw data by shifting the GDP to last year to be an independent variable, label encoding to map the raw names of the factors to abbreviations, and one-hot encoding to change the region names into independent variables. These preprocessing steps were likely done to make the data more suitable for training the prediction model.

The range of GDP values across different regions was assessed using a boxplot, grouping the values by region. With the help of the Figure 1, we found that several regions like Jiangsu and Guangdong have a numerous vary of GDP in the past 20 years and most of the regions have a flat trend. Conversely, most regions exhibited a relatively stable trend. Consequently, when selecting regression models to predict future GDP values, it is crucial to consider both exceptional regions and prediction accuracy. Notably, the inclusion of tree models should be contemplated to enhance prediction accuracy for standout regions, while evaluating the efficacy of linear regression models in forecasting overall trends across the majority of regions.

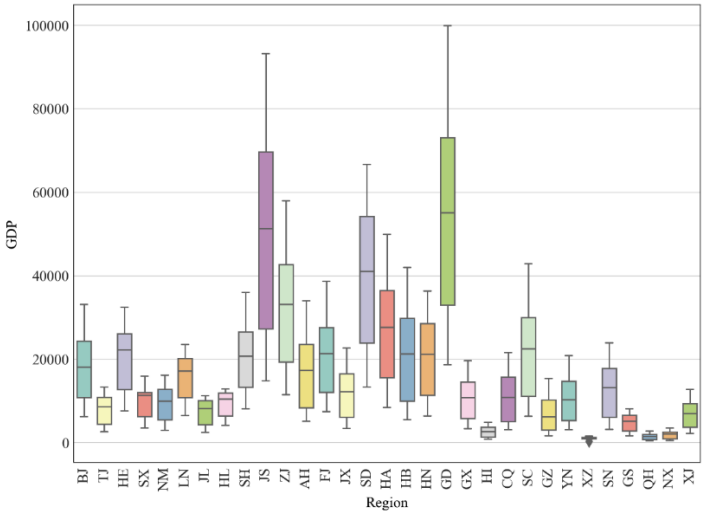

Fig. 1. Regional GDP range boxplot

The data was split to training dataset and testing dataset, and LR(linear regression) model and XGR(XGBoost regression) model were built separately, to predict the next year's GDP value, and we want to elect one method to make a balance for the performance of the LR and XGR. As the entropy method can evaluate the weights of the indicators, and trade off the performance, to make a more precise prediction. And the prediction result shown below as described in Table 2.

Table 2. prediction models comparison

Actual Data	LR	XGR	Entropy
1173.0	478.50	1267.770020	871.95
2020.5	1910.81	1887.880005	1899.38
2264.1	2384.03	2282.639893	2333.49
2327.7	2376.06	2374.969971	2375.52
2748.0	2807.93	2741.669922	2774.90
...	...	...	...
42326.6	42596.91	42191.921875	42395.03
53717.8	53937.18	60429.781250	57173.66
59349.4	59272.58	62856.898438	61059.33
63012.1	64054.48	59329.351562	61699.06
70540.5	71494.04	73912.398438	72699.56

Unit: thousand billion RMB

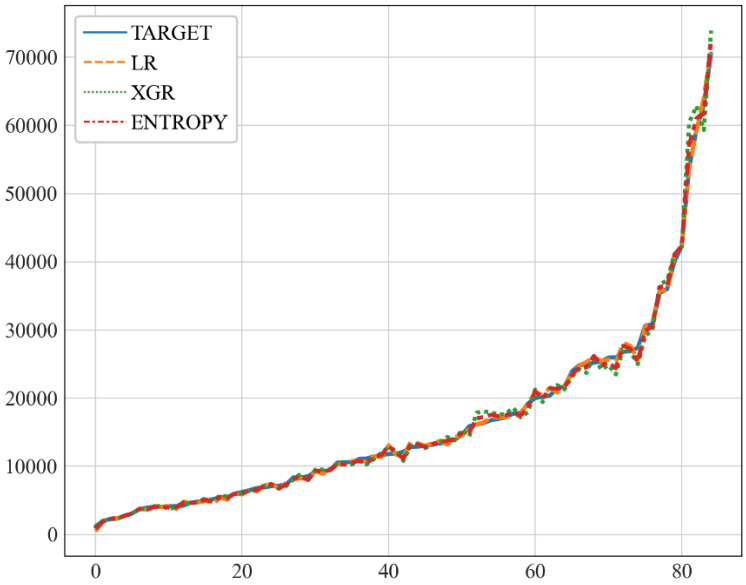

Fig. 2. Predicted results of all algorithms.

Figure 2 depicts the superior performance of the novel combined model, namely the entropy method based on LR and XGR, compared to the individual LR and XGR models. Table 2 presents the calculated Mean Absolute Percentage Error (MAPE) values for LR, XGR, and Entropy. Specifically, LR predictions yielded a MAPE of 3.57, while XGR resulted in a MAPE of 4.22. Notably, the evaluation of weights for the LR and XGR models using the entropy method yielded a metric value of 3.30, which is lower than both predicted models. These results demonstrate that the novel combination model utilizing the entropy method, incorporating LR and XGR, exhibits enhanced precision. Consequently, it is imperative for scholars to devote further attention to the potential of novel combination models.

5 Conclusion

The aim of this study is to develop a predictive model for regional GDP using historical GDP data and other relevant factors. The study proposes a novel combination model that calculates weights for the LR and XGR models, leveraging the trend-predicting strength of LR and the accuracy-predicting strength of XGR. It provides a comprehensive review of prediction methods and factors influencing regional GDP, focusing on model performance. The study analyzes GDP trends for each region, predicts next year's regional GDP using LR and XGR models, and calculates weights using the entropy method. The models are compared using the MAPE metric and actual regional

datasets, indicating superior performance of the combined entropy-based LR and XGR model. Future research should explore additional factors and alternative models for better regional GDP prediction.

References

1. W. M. Morrison, "China's economic rise: History, trends, challenges, and implications for the United States," *Curr. Polit. Econ. North. West. Asia*, vol. 28, no. 2/3, pp. 189–242, 2019.
2. A. R. Kroeber, *China's Economy: What Everyone Needs to Know®*. Oxford University Press, 2020.
3. E. R. Sadik-Zada and W. Loewenstein, "Drivers of CO2-Emissions in Fossil Fuel abundant settings:(Pooled) mean group and nonparametric panel analyses," *Energies*, vol. 13, no. 15, p. 3956, 2020.
4. C. Wu and P. Chen, "Application of support vector machines in debt to GDP ratio forecasting," presented at the 2006 International Conference on Machine Learning and Cybernetics, IEEE, 2006, pp. 3412–3415. doi: 10.1109/ICMLC.2006.258504.
5. Y. Yusof, S. S. Kamaruddin, H. Husni, K. R. Ku-Mahamud, and Z. Mustaffa, "Forecasting model based on LSSVM and ABC for natural resource commodity," *Int. J. Comput. Theory Eng.*, vol. 5, no. 6, p. 906, 2013.
6. G. Long, "GDP prediction by support vector machine trained with genetic algorithm," presented at the 2010 2nd International Conference on Signal Processing Systems, IEEE, 2010, pp. V3-1. doi: 10.1109/ICSPS.2010.5555854.
7. S. Sa'adah and M. S. Wibowo, "Prediction of gross domestic product (GDP) in Indonesia using deep learning algorithm," presented at the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), IEEE, 2020, pp. 32–36. doi: 10.1109/ISRITI51436.2020.9315519.
8. Y. Li and W. Chen, "A Comparative Performance Assessment of Ensemble Learning for Credit Scoring," *Mathematics*, vol. 8, no. 10, p. 1756, Oct. 2020, doi: 10.3390/math8101756.
9. Y. Li and W. Chen, "Entropy method of constructing a combined model for improving loan default prediction: A case study in China," *J. Oper. Res. Soc.*, vol. 72, no. 5, pp. 1099–1109, 2021.
10. Y. Han *et al.*, "Novel economy and carbon emissions prediction model of different countries or regions in the world for energy optimization using improved residual neural network," *Sci. Total Environ.*, vol. 860, p. 160410, 2023, doi: 10.1016/j.scitotenv.2022.160410.
11. Z. Dai, Y. Hu, and G. Zhao, "The suitability of different nighttime light data for GDP estimation at different spatial scales and regional levels," *Sustainability*, vol. 9, no. 2, p. 305, 2017.
12. D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.
13. X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, "Predicting Missing Values in Medical Data Via XGBoost Regression," *J. Healthc. Inform. Res.*, vol. 4, no. 4, pp. 383–394, Dec. 2020, doi: 10.1007/s41666-020-00077-1.
14. "National Bureau of Statistics." www.stats.gov.cn