

A study of the GeGDP problem base on the LightGBM regression

Yang Peilin

College of Computer Science and Technology, Ocean University of China

Abstract. This study uses AHP hierarchy analysis to select three indicators as the measurement standard of green GDP: GDP, resource consumption and reduction cost and environmental degradation cost, and analyzes the weight of green GDP indicators and the importance of the three indicators. With GeGDP as the main indicator of a country's economic health, a Lasso regression model is established to analyze the predicted global climate mitigation impacts. Then, this paper establishes LightGBM regression model to predict the future GeGDP and GDP of the United States, and uses r square and other indicators to test the accuracy of the model, and makes a human correlation analysis of GeGDP and GDP before and after the prediction, to analyze the degree of correlation between the two. Finally, taking the United States as an example, we substituted relevant data into the LightGBM regression model developed above, predicted its value, and conducted one-way analysis of variance on it to determine the degree of change before and after. This study uses AHP hierarchy analysis to select three indicators as the measurement standard of green GDP: GDP, resource consumption and reduction cost and environmental degradation cost, and analyzes the weight of green GDP indicators and the importance of the three indicators. With GeGDP as the main indicator of a country's economic health, a Lasso regression model is established to analyze the predicted global climate mitigation impacts. Then, this paper establishes LightGBM regression model to predict the future GeGDP and GDP of the United States, and uses r square and other indicators to test the accuracy of the model, and makes a human correlation analysis of GeGDP and GDP before and after the prediction, to analyze the degree of correlation between the two. Finally, taking the United States as an example, we substituted relevant data into the LightGBM regression model developed above, predicted its value, and conducted one-way analysis of variance on it to determine the degree of change before and after.

Keywords: GeGDP, AHP analysis, Lasso regression, LightGBM regression.

1. Introduction

1.1 Research Background

Gross Domestic Product (GDP) is an important indicator of a country's economic health, calculating the monetary value of the final goods and services produced by the country over a given period of time. Although GDP is often quoted as the main indicator of economic health, it fails to take into account the depletion of natural resources and the negative impact on the environment and therefore may not be a good measure of a country's true economic health. This has led to the creation of a new indicator, the Green GDP (GeGDP), which is an economic indicator that takes into account environmental and sustainability factors, including the consumption of natural resources, loss of ecosystems and environmental pollution. Compared to traditional GDP, GeGDP is a more comprehensive and accurate indicator of a country's true economic health.

1.2 Study purpose and significance

In this study, AHP hierarchical analysis will be used to construct the measurement standard of green GDP, and GeGDP will be taken as the main indicator of a country's economic health, and a regression model will be established to analyze the predicted global climate mitigation impact. The first aim is to develop a more stable and concise model, using the GeGDP as a primary indicator of a country's economic health, to estimate the expected global impact on climate mitigation. The second purpose was to determine whether the models showed that such a shift would be worthwhile on a

global scale, comparing the potential advantages and disadvantages of replacing GeGDP with climate mitigation impacts.

1.3 Innovation point

Few scholars have been able to go on to subdivide these factors into different types and study interactions between factors under different types of interaction combinations or single factors. This experiment uses Lasso regression algorithm, which can process multicollinearity data, such as ridge regression, which is a biased estimation.

By using a histogram-based algorithm, the LightGBM algorithm can be trained on large data sets in a relatively short time.

2. Literature review

The net economic welfare index proposed by Jamps Tobin became the theoretical basis for many scholars to discuss green GDP. Robert Lupetto's measure of net domestic product, which adjusts for the net change in natural resources, excludes elements that cannot be measured in money, such as air pollution.

Konin put forward "National accounting Matrix including environmental accounts" (NAMEA for short), NAMEA matrix uses the traditional input-output model to connect with environmental data, and uses real material quantity to calculate environmental data and further transform it into environmental indicators that can be converted accordingly, so as to calculate green GDP.

In 1993, the Economic and Social Information and Policy Analysis Branch of the United Nations Statistical Office proposed "System of Environment Economic Accounting" (SEEA). In this system, the framework of integrated environmental and economic accounting, asset classification, environmental cost and benefit assessment, natural resources and environmental valuation methods and so on are discussed.

3. Experimental study

3.1 AHP hierarchical analysis on GeGDP

3.1.1 Experimental procedure

We use the AHP hierarchical analysis method to select three indicators: GDP, resource depletion and reduction costs and environmental degradation costs, give the relative importance values of these three indicators, and analyse the weight of the indicators of green GDP accounted for by these three indicators from this perspective, and analyse the importance of these three indicators.

3.1.2 Results of the AHP model

Table 1 Indicator indices

Indicators	Environmental	Resource	GDP
Environmental downgrade costs	1	0.5	0.333
Resource consumption and cost reduction	2	1	0.5
GDP	3	2	1

The table above shows the judgment matrix constructed.

Table 2 Results of AHP hierarchical analysis

Item	Eigenvectors	Weighting	Maximum	CI
Environmental downgrade costs	0.55	16.342		
Resource consumption and cost reduction	1	29.696	3.009	0.005
GDP	1.817	53.961		

The results of the weight calculation of the hierarchical analysis (square root method) show that the weight of Environmental downgrade costs is 16.342%, the weight of Resource consumption and cost reduction is 29.696% and the weight of GDP is 53.961%.

Table 3 Consistency test results

Maximum characteristic root	CI value	RI value	CR value	Consistency test results
3.009	0.005	0.525	0.009	By

The results of the hierarchical analysis showed that the maximum characteristic root was 3.009 and the corresponding RI value was found to be 0.525 according to the RI table, so $CR=CI/RI=0.009<0.1$, which passed the one-off test.

Ultimately, we obtain the formula as

$$GeGDP = GDP - \text{Resource Depletion Costs} - \text{Environmental Degradation Costs}$$

3.2 Lasso regression GeGDP model

3.2.1 Experimental procedure

We collected annual GDP data, CO2 emissions data, temperature, environmental index data, and resource data from the World Bank and UNEP data sites for five different countries for the period 2000-2020 to create Lasso regressions to analyse the expected global impact on climate mitigation by prediction. the Lasso method is an alternative least squares compression estimation method.

3.2.2. Results of the social value orientation experiment

The results of the Lasso regression show that the standardised coefficients based on the variable intercept term, Carbon dioxide emissions, Temperature, Environmental Index, Resources, the variable intercept term, Carbon dioxide emissions, Temperature, Environmental Index, Resources were retained and no variables were removed.

Normalised equation for the model:

$$y = -35.608 + 0.002 \times \text{Carbon dioxide emissions} \\ + 0.34 \times \text{Temperature} + 0.48 \times \text{Environmental Index} \\ + 0.012 \times \text{Resources}.$$

Unstandardised equation for the model:

$$y = -30.488 + 0.002 \times \text{Carbon dioxide emissions} \\ + 0.257 \times \text{Temperature} + 0.427 \times \text{Environmental Index} \\ + 0.029 \times \text{Resources}.$$

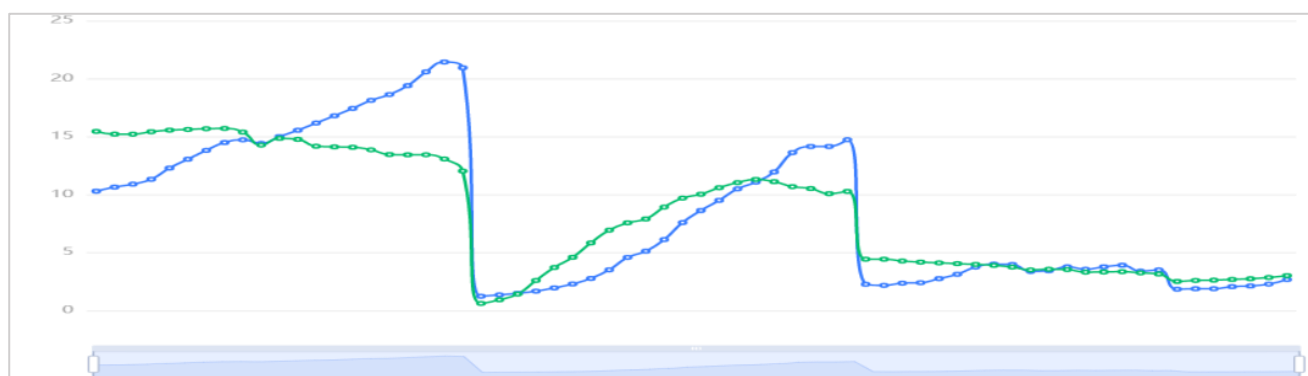


Figure 4 Graph of model results

The graph above shows the raw data plot, model fitted values, and model predicted values for this model.

3.3 LightGBPGBM regression model

3.3.1 Experimental procedure

We first used the data collected to calculate the GeGDP indicators corresponding to different countries in different periods, build a LightGBM regression model to forecast the future GeGDP and GDP of the US, and also use indicators such as R-squared to test the accuracy of our model, and conduct Person correlation analysis on GeGDP and GDP before and after the forecast to analyse their the degree of correlation between the two.

After that, we collected the US GDP, GeGDP, economic status status, and future ability to provide for future generations for the years 2000-2020, used the data above, substituted them into the LightGBM regression model developed above, predicted their values, and performed a one-way ANOVA on them to determine the degree of change before and after.

3.3.2 Results of the LightGBPGBM regression

Table 4 Table of correlation coefficients

	GDP	Carbon s	Temperatur e	Environmental	Resources	GeGDP
GDP	1 (0.000***)	0.536 (0.000***)	0.073 (0.562)	0.309 (0.012**)	0.592 (0.000***)	-0.275 (0.025**)
Carbon dioxide	0.536 (0.000***)	1 (0.000***)	0.336 (0.006***)	-0.513 (0.000***)	0.8 (0.000***)	0.196 (0.115)
Temperature	0.073 (0.562)	0.336 (0.006***)	1 (0.000***)	-0.522 (0.000***)	0.144 (0.249)	0.516 (0.000***)
Environmenta l Index	0.309 (0.012**)	-0.513 (0.000***)	-0.522 (0.000***)	1 (0.000***)	-0.205 (0.098*)	- 0.775(0.000***)
Resources	0.592 (0.000***)	0.8 (0.000***)	0.144 (0.249)	-0.205 (0.098*)	1 (0.000***)	-0.306(0.013**)
GeGDP	-0.275 (0.025**)	0.196 (0.115)	0.516 (0.000***)	- 0.775(0.000***)	- 0.306(0.013**)	1 (0.000***)

Note: ***, ** and * represent 1%, 5% and 10% significance levels respectively

The table above shows a table of the results of the parameters tested for the model, including the correlation coefficient, and the significant P-value. The test is first conducted to determine whether there is a statistically significant relationship between XY and whether the p-value presents significance ($p < 0.05$). If it shows significance, then there is a correlation between the two variables, and if not, then there is no correlation between the two variables. Finally, analyse the positive and negative direction of the correlation coefficient and the degree of correlation.

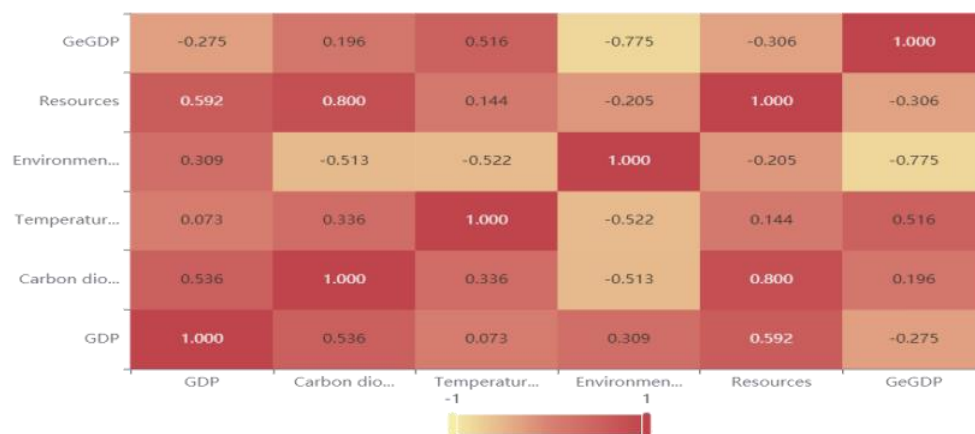


Figure 5 Heat map of correlation coefficient

The diagram above shows the values of the correlation coefficients in the form of a heat. map, with the magnitude of the values indicated mainly by the shade of colour.

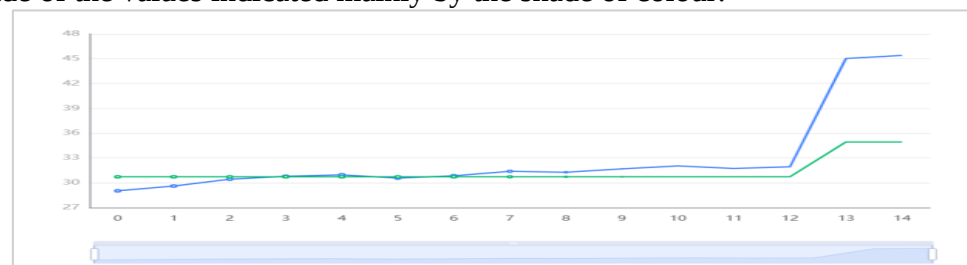


Figure 6 Predicted graph of test data

The graph above shows LightGBM's predictions for the test.

3.3.3 Results of one-way ANOVA

Analysis term: GDP, sample using Shapiro-Wilk test, significance p-value of 0.469, level does not present significance and cannot reject the original hypothesis, therefore the data satisfies normal distribution.

Analysis term: GeGDP, sample using Shapiro-Wilk test, significance p-value of 0.460, level does not present significance and cannot reject the original hypothesis, therefore the data satisfies normal distribution.

Table 5 Chi-square test

	Economic status (standard deviation)			F	P
	Good (n=18)	Unstable (n=2)	Decline (n=1)		
GDP	3.521	4.306	0.000	1.311	0.294
GeGDP	5.733	7.432	0.000	1.508	0.248

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

The results of the chi-square test show that for GDP, the significance p-value is 0.294, which does not present significance at the level that the original hypothesis can be rejected, and therefore the data satisfies the chi-square.

The results of the chi-square test show that for GeGDP, the significance p-value is 0.248, which does not present significance at the level where the original hypothesis can be rejected, and therefore the data satisfies the chi-square.

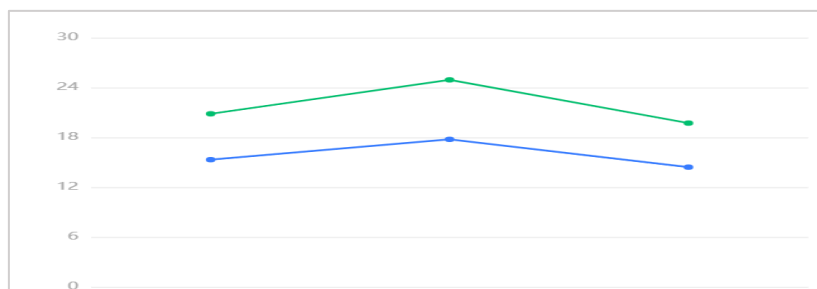


Figure 7 One-way ANOVA comparison chart

The graph above shows the results of the ANOVA for the means, which can be mined for differential relationships by comparing the means.

Table 6 Table of ANOVA results

Variable name	Variable value	Sample size	Average	Standard deviation	F	P
GDP	Good	18	15.319	3.521	0.472	0.632
	Unstable	2	17.765	4.306		
	Recession	1	14.42	0.000		
	Total	21	15.51	3.473		
GeGDP	Good	18	20.84	5.733	0.473	0.631
	Unstable	2	24.925	7.432		
	Recession	1	19.72	0.000		
	Total	21	21.176	5.684		

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

The mean values of good and unstable versus recession on GDP are 15.319*/17.765*/14.420* respectively; the ANOVA results in a p-value of 0.632>0.05 and therefore the statistical results are not significant, indicating that there is no significant difference between the different Economic status on GDP.

The mean values of good and unstable versus recession on GeGDP are 20.840*/24.925*/19.720* respectively; the ANOVA results in a p-value of 0.631>0.05 and therefore the statistical results are not significant, indicating that there is no significant difference between the different Economic status on GeGDP.

Table 7 Quantitative analysis of effects table

Analysis items	Difference between groups	Total deviation	Partial Eta-squared (Partial η^2)	Cohen's f-value
GDP	12.012	241.272	0.05	0.229
GeGDP	32.262	646.213	0.05	0.229

The results of the quantitative analysis of effects show that the Eta square (η^2 value) is 0.05 based on GDP, indicating that 5.0% of the variation in the data is derived from differences between groups. cohen's f value is 0.229, indicating that the degree of variation in the quantification of effects of the data is a small degree of variation.

The results of the quantitative analysis of effects showed an Eta-square (η^2 value) of 0.05 based on GeGDP, indicating that 5.0% of the variation in the data was derived from differences between groups. the Cohen's f value was 0.229, indicating that the degree of variation in the quantification of effects of the data was small.

4. Conclusion

This study uses AHP hierarchy to analyze GDP, resource consumption and reduction costs and environmental degradation costs, and analyzes the importance of indicators. Using GeGDP as a

primary indicator of a country's economic health, a Lasso regression model was established to analyze the predicted global climate mitigation impacts. Then, the LightGBPGBM regression model is established to predict the future GeGDP and GDP of the United States, and the r square and other indicators are used to test the accuracy of the model, and the GeGDP and GDP before and after the forecast are analyzed by human correlation, to analyze the degree of correlation between the two. Finally, relevant data were substituted into the LightGBM regression model developed above to predict its value, and one-way analysis of variance was performed on it to determine the degree of change before and after.

References

- [1] Ma Xiaojun, Sha Jinglan, Niu Xueqi. Design and application of P2P project credit rating model based on LightGBM algorithm[J]. Quantitative Economic and Technical Economics Research, 2018, 35(05): 144-160. DOI: 10.13653/j.cnki.jqte.20180503.001.
- [2] Xie Y, Xiang W, Ji MZ, Peng J, Huang YIhuai. Application analysis of predicting monthly housing rent based on Xgboost and LightGBM algorithms[J]. Computer Applications and Software, 2019, 36(09): 151-155+191.
- [3] Jinglan Sha. A comparative study of P2P online lending default prediction models based on LightGBM and XGBoost algorithms [D]. Northeast University of Finance and Economics, 2017.
- [4] Ke Zhenglin. Application of Lasso and its correlation method in multiple linear regression models [D]. Beijing Jiaotong University, 2011.
- [5] Xiu-Xiu Zhang, Hui Wang, Shuang-Shuang Tian, Nan Qiao, Lina Yan, Tong Wang. LASSO-based independent variable selection in regression analysis of high-dimensional data[J]. China Health Statistics, 2013, 30(06): 922-926.