An Elastic Net Based Algorithm for China Agriculture GDP Prediction

Zihan Qiu(✉)

Xiamen University, Xiamen, China
qiu.zihan.xmu@qq.com

Abstract. Gross domestic product (GDP) refers to the final result of production activities of all resident units in a country or region within a certain period of time. There are a variety of GDP forecasting methods, which can be classified into three types: Time Series Analysis, Regression Analysis and VAR Model. In our paper, we utilize the agricultural yields data to predict the agriculture GDP, that can be seen as a regression model. We adopt Elastic net linear regression using the penalties from both the lasso and ridge techniques to regularize regression models. We evaluate our result using the metrics of Mean Absolute Error (MAE). The lower MAE, the better performance the model will owns. From the result, Elastic Net method owns the lowest MAE score 2.34. In contrast, the other methods like Linear Regression, Lasso, Ridge and VAR's MAE are 3.25, 4.25, 3.06, 4.45 respectively.

Keywords: GDP prediction · Regression Analysis · Elastic Net · Mean Absolute Error

1 Introduction

Gross domestic product (GDP) [1] refers to the final result of production activities of all resident units in a country or region within a certain period of time. This indicator summarizes the output results of all activities of the national economy in a very concise statistical figure, providing the most comprehensive scale for evaluating and measuring the economic performance of national economic conditions, economic growth trends and social wealth, which is also the most important economic indicator that affects economic life and even social life. Its analysis and prediction have important theoretical and practical significance. The GDP of a country is composed of the GDP of each province. The study of the GDP of each province plays an important role in the study of the GDP and the economy of each province and even the whole country.

China is a country with a large population and agricultural production. According to the National Bureau of Statistics of China [2], in 2020, the added value of agriculture and related industries nationwide will be 16.69 trillion yuan (RMB), accounting for 16.47% of the gross domestic product (GDP). According to crop yield, the reasonable use of machine learning methods to predict per capita GDP plays an important role in economic development research.

F. Balli et al. (Eds.): ESFCT 2022, AEBMR 226, pp. 843–849, 2022.
https://doi.org/10.2991/978-94-6463-052-7_96

If we take multiple agricultural yields as input and the value of GDP as output, we can view this task as a linear regression problem. The dataset is provided by Kaggle platform. In our paper, we utilized an Elastic Net for regression. Related work is described in Sect. 2, and we introduce our methodology and experiment in Sects. 3 and 4.

2 Related Work

At present, there are a variety of GDP forecasting methods, which can be classified into two types: Time Series Analysis and Regression Analysis Model.

2.1 Time Series Analysis

A time series is a set of data columns that are observed by a phenomenon at different times, in chronological order. Time series prediction [3, 4] is the root of a thing using its past characteristic values to predict its future trend. There are regularities and irregularities in the movement of time series data. The size of each observation in a time series is a combination of the various factors that affect the change at the same time.

Time series forecasting can be used for short-term, medium-term, and long-term forecasts. Time series forecasting methods can be divided into two broad categories: one is deterministic time series models, and the other is random time series models. Deterministic time series prediction method refers to the time series with a definite time function $y - f(x)$ to fit, with different function forms to represent different changes in the sequence, with different function superposition to represent the superposition of different changes in the sequence. This method can be divided into trend prediction method, decomposition analysis method, smooth forecasting method and so on. The stochastic time series analysis method analyzes the correlation between the eigenvalues of the sequence at different times. It reveals the internal correlation structure between them, and uses this correlation structure to make predictions.

Vector autoregression model (VAR) [6] is a statistical model utilized to capture the relationship between multiple quantities as they change over step. Actually, it is a kind of stochastic process model. VAR models can generalize the single-variable (univariate) autoregressive model by allowing for multivariate time series.

2.2 Regression Analysis

Regression analysis [7, 8] and prediction method is to establish a regression equation between variables and variables on the basis of analyzing the correlation between independent variables and dependent variables, and use the established regression equation as a prediction model. The change of dependent variable is predicted according to the quantitative change of the independent variable, and the change relationship is generally a correlation relationship. Regression analysis prediction method is mainly to analyze and predict variables with causal relationship. The model established by regression analysis can only be applied to the actual prediction after passing various tests and the prediction error is small. According to the number of independent variables in the model, regression models can be divided into univariate regression models and multiple regression models; or according to whether the model is linear, it can be divided into linear regression models and nonlinear regression models.

Univariate linear regression only involves a dependent variable y and an independent variable x◦A linear function of x is used to predict y, that is, $y = a + bx$, where a and b are called regression coefficients. The corresponding straight line is called the regression line. When making predictions with a univariate linear regression model, a, b must first be estimated. The least squares method is generally used.

The multiple linear regression model is similar to the univariate linear regression model. The multiple linear regression model has one dependent variable, but two or more independent variables. The modeling process is all the same.

The per capita GDP prediction model in this paper uses a multiple regression model, which takes the output of multiple agricultural products as the input variable and the per capita GDP as the output. Based on this, a regression analysis equation is established, and this regression equation is used to predict the Agriculture GDP in a certain year.

2.3 Our Contribution

We create a regression model to predict the Agriculture GDP, and adopt the Elastic Net to advance the performance of model, then do compared experiments and choose MAE as our matric.

3 Methodology

Elastic net [9] uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

Elastic Net regression is a combination of Ridge regression and Lasso regression. All three methods solve the problem of overfitting in regression.

The ultimate goal of linear regression is to minimize the following loss function, and the basic idea is the least squares method:

$$\underset{\theta}{min}\frac{1}{2m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2$$

Specifically, ridge regression is a regularization method (l2 regularization) [10] that adds the sum of squares to the loss function, namely:

$$\underset{\theta}{min}\frac{1}{2m}\left[\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\|\theta\|_2^2\right]$$

Lasso regression is a regularization method (l1 regularity) [11] that adds the absolute value sum to the loss function, namely:

$$\underset{\theta}{min}\frac{1}{2m}\left[\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\|\theta\|\right]$$

The Elastic Net regression combines two regularization methods. Hence, in many cases, it can achieve a better result for

$$\min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}\left(\theta^{T}x^{(i)}-y^{(i)}\right)^{2}+\lambda_{1}\sum_{j=1}^{n}\|\theta\|+\lambda_{2}\sum_{j=1}^{n}\|\theta\|_{2}^{2}\right]$$

4 Experiments

4.1 Experimental Data

Our dataset is provided by Kaggle, this dataset contains the yield production of 98 agricultural products on years from 1961 to 2007 and GDP on these years. It contains 'apples', 'watermelons', 'arecanuts', 'asparagus', 'bananas', 'barley', 'beans_dry', 'beans_green', 'berries_nes', 'grapes'. We split data from1961 to 2001 as train dataset, and data of next five years (from 2002 to 2007) as evaluation dataset. The task is to predict GDP using agricultural products. Figure 1 shows the three fruits production vs year. Figure 2 shows Agriculture GDP vs year.

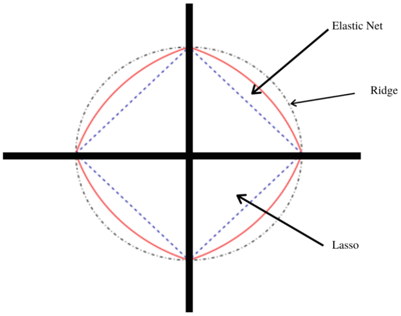

Fig. 1. Penalty of three methods

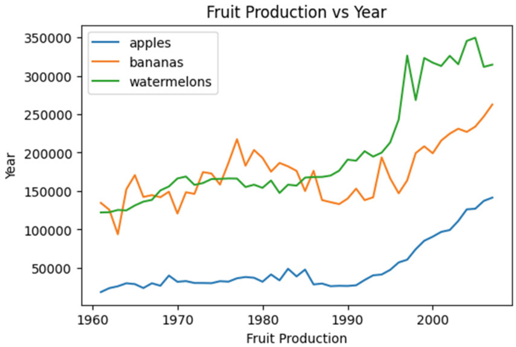

Fig. 2. Fruit Production vs Year

4.2 Experimental Setting

Firstly, the feature engineering part is done before training the model. For the year feature, we subtract it by the minimal year which is 1961. For all the features, the standard scaler with a mean value of 0 and std of 1 is used. This preprocessing can reduce the variance of the data distribution and hence lead to a better result.

We do compared experiments among Linear Regression, Var, Lasso Regression, Ridge Regression and Elastic Net. The model is imported directly from sklearn which is a python library dedicated to machine learning modules.

4.3 Experimental Results

Figure 3 shows the performance different models on the evaluation dataset. It is clearly that Elastic Net perform better than other models (Fig. 4).

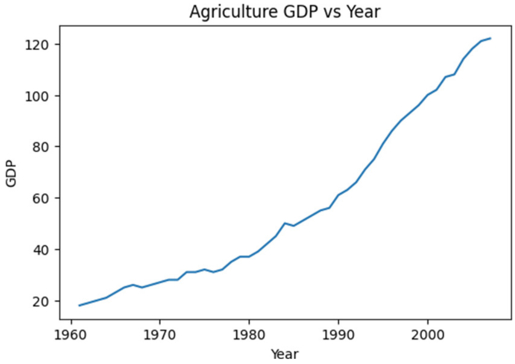

Fig. 3. Agriculture GDP vs Year

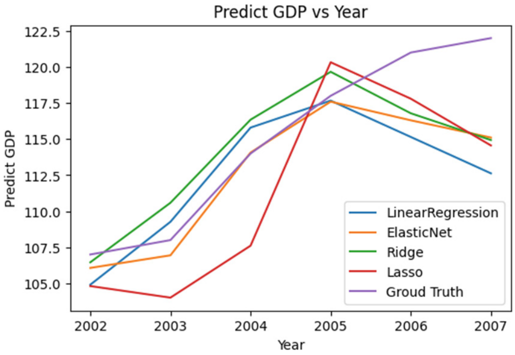

Fig. 4. Predict GDP vs Year

Table 1. Performance of different models

Models	MAE
Linear Regression	3.25
Lasso	4.25
Ridge	3.06
Var	4.45
Elastic Net	2.34

To compare our model's performance, we evaluate our result using the metrics of MAE (Mean Absolute Error).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

When the predicted value is completely consistent with the actual value, the MAE is equal to 0, which is an accurate prediction; the larger the error, the larger the value. The experimental results of competing models and our model are shown in Table 1.

From the compared experiments, Elastic Net method owns the lowest MAE score 2.34. In contrast, the other methods like Linear Regression, Lasso and Ridge, Var's MAE are 3.25, 4.25, 3.06, 4.45 respectively.

5 Conclusion

In our paper, we utilize the agricultural yields data to predict the agriculture GDP, that can be seen as a regression model. We introduce our methodology and experiment in Sect. 3. Section 4 shows experimental results, The lower MAE, the better performance the model will owns. From the result, Elastic Net method owns the lowest MAE score 2.34. In contrast, the other methods Linear Regression, Lasso, Ridge and VAR's MAE are 3.25, 4.25, 3.06, 4.45 respectively.

Acknowledgement. Thanks to the Kaggle platform, which provides online gpu and cpu resources for our research. Thank you for food and agriculture organization of the United Nations, which provides us the agriculture and gdp data in the experiment.

References

1. Ang A, Piazzesi M, Wei M. What does the yield curve tell us about GDP growth?[J]. Journal of econometrics, 2006, 131(1-2): 359-403.
2. Aslam, Bilal, et al. "The nexus of industrialization, GDP per capita and CO2 emission in China.“ Environmental Technology & Innovation 23 (2021): 101674.

3. Sapankevych N I, Sankar R. Time series prediction using support vector machines: a survey[J]. IEEE computational intelligence magazine, 2009, 4(2): 24-38.
4. Weigend A S. Time series prediction: forecasting the future and understanding the past[M]. Routledge, 2018.
5. Geskus R B. Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring[J]. Biometrics, 2011, 67(1): 39-49.
6. Canova F , Ciccarelli M . Panel Vector Autoregressive Models: A Survey[J]. Working Paper, 2013, 31:205–246.
7. Draper N R, Smith H. Applied regression analysis[M]. John Wiley & Sons, 1998.
8. Allen M P. Understanding regression analysis[M]. Springer Science & Business Media, 2004.
9. Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the royal statistical society: series B (statistical methodology), 2005, 67(2): 301-320.
10. McDonald G C. Ridge regression[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2009, 1(1): 93-100.
11. Ranstam J, Cook J A. LASSO regression[J]. Journal of British Surgery, 2018, 105(10): 1348-1348.
12. Xiao X, Duan H, Wen J. A novel car-following inertia gray model and its application in forecasting short-term traffic flow[J]. Applied Mathematical Modelling, 2020, 87: 546-570.