2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID 2021)

Forecasting the impact of COVID-19 on GDP based on Adaboost

Ding Jiangying[1],
China university of geosciences,
Wuhan, China,
Xiaoxiannvdjy@163.com,

Huaijin Shi[1]
The Chinese University of Hong Kong
Shenzhen, China
huaijinshi@outlook.com

Yichen Zhang[2]
Tongji University
Shanghai, China
1954267@tongji.edu.cn

Zhang Huiying[*]
YangZhou University
YangZhou, China
18252715718@163.com

2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID) | 978-1-6654-1537-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/AIID51893.2021.9456518

***Abstract*—Covid-19 Pandemic has a unique impact on the economy as other infectious diseases. Epidemics affect people's daily consumption activities, for example, by causing them to shop less, travel less, consume less and invest less. The reduction of a large number of economic activities leads to the suppression of social demand and the reduction of consumption level, which further affects the GDP of various countries around the world. It is necessary to investigate and analyze the impact of the epidemic on GDP in order to control and analyze the economic situation under the impact of the epidemic. In this paper, we take the impact of COVID-19 on the GDP of each country as a regression problem, and propose to forecast GDP through feature engineering combined with Aaboost model. The model was tested on more than 50,000 data records from more than 200 countries provided by the Kaggle platform to prove the validity. The experiment shows that Adaboost has stronger robustness compared with other methods, such as random forest, SVR. Adaboost improves the MSE of random forest by 2.39 and SVR by 0.38.**

***Index Terms*—Adaboost, Covid-19 Pandemic, GDP**

I. INTRODUCTION

Although the COVID-19 Pandemic does not directly damage people's property, it does affect people's confidence in increasing their future wealth and their expectations of future consumption.

On the one hand, some small and micro enterprises have withdrawn from the market due to the disruption of the capital chain caused by the outbreak, which has weakened the total supply. On the other hand, the epidemic has left most enterprises in a state of shutdown and production reduction. The United States, Japan, South Korea and other countries have successively implemented restrictions on travel and entry, which has a direct impact on the tourism and aviation industry in the relevant areas, and leads to the contraction of catering, accommodation, shopping, transportation, finance and other related services. The COVID-19 outbreak has had a particular impact on trade and services, and has further slowed global trade growth. Countries with higher trade dependence, such as Germany, South Korea and Mexico, face more severe challenges. Many countries are actively taking prevention and control measures, restricting or banning group activities, in response to the severe situation of the rapid spread of the global epidemic. This reduced the risk of crowd gathering but slowed the growth of the real economy. At the same time, the implementation of border controls and strict travel restrictions have severely restricted the development of local retail, wholesale, logistics and other industries, which has further exacerbated the downward pressure on the economy. A significant increase in instability and uncertainties in the world economy will inhibit the growth of investment and productivity, and exacerbate the global economic recession. Against the trend, emerging service industries are rising, such as online express delivery, fresh electricity, online education, remote consultation, remote office, etc. New business forms, new models and new products have also been launched. To sum up, the impact of the epidemic on the economies of various countries is comprehensive and needs to be taken into consideration in many aspects.

As the global economy is affected, it is necessary for researchers to analyze the overall impact of the COVID-19 Pandemic on the economy through technical means, so as to assist decision-making organs of various countries to a certain extent. Therefore, this paper collates and analyzes 50000 pieces of relevant economic data from more than 200 countries. The proposed model predicts the economic impact of the COVID-19 Pandemic through feature engineering and the Adaboost model. Experimental results show that our method is superior to many existing machine learning methods.

In conclusion, the main contributions of this paper are as follows:

- A regression model based on Adaboost is proposed to forecast GDP.
- Feature engineering method is used to improve the quality of data and further improve the performance of regression model.

The rest of this article is arranged as follows. In Section 2, the relevant research on the economic impact of COVID-19 is briefly introduced. The details of the proposed model are given in Section 3. In Section 4, we presented the quantitative results and analysis of the experiment that predicts the overall impact

978-0-7381-1083-7/21/$31.00 ©2021 IEEE

of the COVID-19 Pandemic on the economy. Finally, the conclusion is expounded in Section 5..

II. RELATED WORK

With the outbreak of the COVID-19 Pandemic, the economic impact of the COVID-19 Pandemic has attracted increasing attention from researchers [1-3].

Fernandes et al. [4] proposed that global economic recession is inevitable and the function of global supply chain has already been destroyed. Millions of people may lose their jobs in the future, and GDP will inevitably decline. The bidirectional Long Short-Term Memory (LSTM) model was used for the early prediction of economic conditions and positive cases, and a novel economic promotion scheme was proposed to promote economic activities by Vekaria et al. in [5]. Albu et al. [6] estimated the impact of the COVID-19 Pandemic on economic growth by simulating the daily dynamics of the COVID-19 Pandemic worldwide, as well as in the European Union and Romania. A multi-layer artificial neural network model proposed by Jena et al. [7], predicted the quarterly GDP data of eight countries (namely the US, Mexico, Germany, Italy, Spain, France, India and Japan) for the period of April-June 2020. Noy et al. [8] used data from 2014 to 2018 and conceptual disaster risk models to calculate and measure the exposure, vulnerability and resilience of the local economy to the impact of the epidemic. Sa'adah et al. [9] used two deep learning technology methods (LSTM and RNN) to forecast Indonesia's GDP. The model proposed by Kim et al. [10] uses the LSTM algorithm to predict the economic impact of popular trends and shows high performance in time series forecasting, and its performance results show validity in describing the inflation rate.

III. METHOD

In this section, the data used, feature engineering, and algorithm details are described..

A. Feature Engineering

The data set used for the experiment in this paper is from the data set of the impact of the COVID-19 Pandemic on the global economy published by the Kaggle platform [11]. The dataset contains a variety of features. Specific features include: Gross Domestic Production pre-captcha (GDP), Human Development Index (HDI), Stringency Index (STI), Total number of Cases (Total Cases, TC), Total Deaths Reported (TD), COUNTR, Population (POP). Record dates (DATE) and country/area codes (COUNTRY) are also provided. The changes of TD, TC, and STI of the dataset over time in 2020 are shown in Figure 1.The statistics of HDI, POP and GDP are shown in Figure 2.

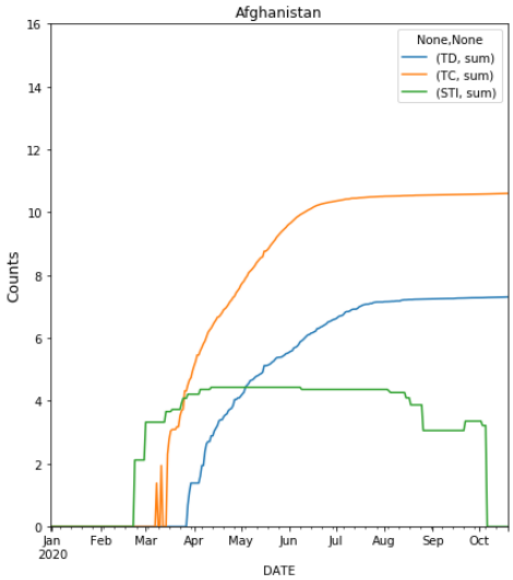

Figure 1. The changes of TD, TC, and STI over time.

In data processing, the first step is to judge whether the features in each column contain missing values. According to the features with missing values, the number of missing data and the total data of the features are counted. If the number of missing data exceeds half of the total number, the features in the column are directly deleted, so as to avoid the large deviation caused by the filling values to the fitting of the algorithm. And if not, the median of the column features is populated to ensure that the data does not generate large distribution deviations. In addition, we also tried to fill 0 and fill mode respectively, but the final fitting effect was not as good as the filling median. For HDI, normalization is adopted to map the features to [0, 1], which can be formulated as follows:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Normalization makes the optimization process of the optimal solution smooth and easier to correctly converge to the optimal solution. And as for COUNTR, we use One-Hot coding, which extends the values of discrete features to Euclidean space. After coding, one value of the discrete feature corresponds to a point in Euclidean space. The method of one-hot coding discrete features makes the calculation of distance between features more reasonable. Besides, for countries with a small sample size, we use over-sampling to ensure that the categories are balanced. This will ensure that the sample size of each country stays at the same order of magnitude, so that countries with large sample sizes do not influence the model too much.

B. *Adaboost*

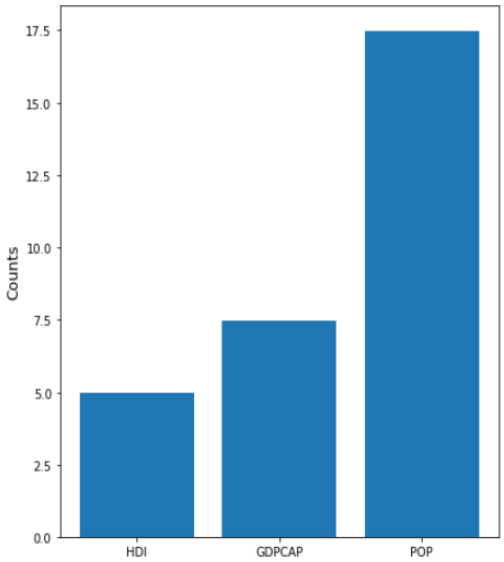

Figure 2. The statistics of HDI、POP and GDP.

The working mechanism of Adaboost is briefly described as follows. Firstly, a weak learner is trained from the training set with the initial weight, and the weights of the training samples are updated according to the badcases output by the weak learner, which makes the weights of the badcases output by the weak learner become higher, so that the later learners pay more attention to learning the features of these badcases. The training is iterated in the above way until the number of weak learners reaches the pre-specified number. Ultimately, all weak learners are integrated through Boosting strategy to get the final strong learners.

Suppose the sample of training set is shown as Eq. (2):

$$X = \{(x, y_1), (x_2, y_2), \dots (x_m, y_m)\} \tag{2}$$

The number of iterations of the weak learner is K, and the output is the final strong learner f(x). The weight of the initialized sample set is formulated as follows:

$$W = (w_{11}, w_{12}, \dots w_{1m}); w_{1i} = \frac{1}{m}; \tag{3}$$

In the k-th ite, the sample set with weight W was used as the training data to obtain the weak learner Mk. The loss L_{ki} in the training set is calculated to find the badcase output by M_k. The loss function uses the MSE, and can be formulated as follows:

$$L_{ki} = \frac{(y_i - M_k(x_i))^2}{E_k^2} \tag{4}$$

Where y_i is the groundtruth, $M_k(x_i)$ is the prediction result output by the weak learner, and E_k^2 is the maximum error. The regression loss rate L_k can be obtained by loss L_{ki} which is formulated as follows:

$$L_k = \sum_{i=1}^{m} w_{ki} L_{ki} \tag{5}$$

The regression loss rate Lk is used to calculate the coefficient of the weak learner and to update the weight distribution of the training set. After K-ite training, the output of multiple weak classifiers $M_k(x_i)$ can be integrated to get the final output result of the strong learner.

IV. EXPERIMENTS

In this section, we will introduce the details of the methods involved in the experiment and the corresponding summary and analysis of the experiment. In order to ensure the objectivity of the experiment, the dataset is divided into training set and test set, and the ratio of training set and test set is 5:1. The algorithms for comparison were random forest and SVR. The weak classifier used by Adaboost is the CART regression tree. And the reason why CART regression tree is adopted is that the size of this data set is small, and the training samples will be overfitted by the common weak learner. In this way, the learner is too large and complex for the test sample, which may produce a high classification error rate and lead to poor generalization performance of the learner. The CART regression tree can be pruned by algorithm, that is, removing some nodes to solve the over-fitting problem, resulting in better generalization. The number of iterations of the base learner is set to 50, and the learning rate is set to 1. The random seed is set to 2021.

In the experiment, it is found that large learning rate is easy to lead to unstable training, and small learning rate leads to slow convergence speed.The learning rate and the number of weak learners need to be tradeoff to achieve better performance for Adaboost.

The evaluation used in the experiment is the mean square error, and it can be formulated as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{6}$$

Where y_i is the groundtruth, $\hat{y}_i$ is the prediction, and the mean square error (MSE) is the mean value of the sum of the square of the difference between the groundtruth and the prediction. Table 1 illustrates the performance of SVR, random forest and Adaboost on the impact of COVID-19 on the GDP of each country under the MSE. It can be seen from Table 1 that Adaboost has stronger robustness compared with random forest and SVR. Adaboost improves the MSE of random forest by 2.39 and SVR by 0.38.

Both random forest and Adaboost are based on the fusion of multiple weak classifiers, but Adaboost is significantly better in performance, which is caused

TABLE I. TABLE 1. EVALUATION RESULTS

Model	Mean Square Error
SVR	7.763
Random Forest	9.771
AdaBoost	7.383

by more bias in the data set. Adaboost is based on boosting, while random forest is based on bagging. Bagging resamples the sample, and each sub-sample set obtained by resampling trains a model, and finally takes the average. Since the similarity of the subsample sets and the models used are the same, the models have approximately equal bias and variance, so the bias after bagging is close to that of a single submodel and generally cannot be reduced significantly. From the perspective of optimization, boosting algorithm uses the greedy method to minimize the loss function. In each iteration, the sample is weighted according to the predicted results of the last iteration, so the error will be smaller and smaller as the iteration goes on, and the bias of the model will be reduced continuously. Therefore, the performance of Adaboos is better than that of random forest on data sets with large bias.

V. CONCLUSIONS

In this paper, we analyzed the impact of COVID-19 on GDP based on the dataset provided by Kaggle platform. Methods such as filling in missing values, over-sampling of unbalanced data and one-hot coding were used to process features to complete high-quality output of features. The regression prediction of GDP is carried out by the Adaboost method. The experiment shows that the Adaboost method has different degrees of improvement compared with random forest and SVR under MSE.

ACKNOWLEDGEMENT

We thank Kaggle platform for providing usefual dataset support that allowed us to investigate the experiments.

REFERENCES

[1] Nicola, Maria, et al. "The socio-economic implications of the coronavirus and COVID-19 pandemic: a review." International journal of surgery (2020).

[2] Chakraborty, Indranil, and Prasenjit Maity. "COVID-19 outbreak: Migration, effects on society, global environment and prevention." Science of the Total Environment 728 (2020): 138882.

[3] Saadat, Saeida, Deepak Rawtani, and Chaudhery Mustansar Hussain. "Environmental perspective of COVID-19." Science of the Total Environment (2020): 138870.

[4] Fernandes, Nuno. "Economic effects of coronavirus outbreak (COVID-19) on the world economy." Available at SSRN 3557504 (2020).

[5] Vekaria, Darshan, et al. "ξboost: An AI-based Data Analytics Scheme for COVID-19 Prediction and Economy Boosting." IEEE Internet of Things Journal (2020).

[6] Albu, Lucian Liviu, et al. "Estimates of dynamics of the covid-19 pandemic and of its impact on the economy." Romanian Journal of Economic Forecasting 23.2 (2020): 5-17.

[7] Jena, Pradyot Ranjan, et al. "Impact of COVID-19 on GDP of major economies: Application of the artificial neural network forecaster." Economic Analysis and Policy 69 (2021): 324-339.

[8] Noy, Ilan, et al. "Measuring the Economic Risk of COVID-19." Global Policy 11.4 (2020): 413-423.

[9] Sa'adah, Siti, and Muhammad Satrio Wibowo. "Prediction of Gross Domestic Product (GDP) in Indonesia Using Deep Learning Algorithm." 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, 2020.

[10] Kim, Myung Hwa, et al. "The Prediction of COVID-19 Using LSTM Algorithms." International Journal of Networked and Distributed Computing (2021).

[11] https://www.kaggle.com/shashwatwork/impact-of-covid19-pandemic-on-the-global-economy