Analysis on the influencing factors of multiple indicators in the United States based on multiple linear regression model

Jiahui Bu¹,*, Yixuan Tian², Yifan Zong¹

¹School of Computer, School of Software, School of Cyberspace Security, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
²School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

*Corresponding author: bujiahui2018@njupt.edu.cn

Abstract. The US election will have a certain impact on the US economy. In this paper, different index systems are established to analyze the influence of different factors and American economy. After establishing the index system, this paper will use machine learning regression algorithm to fit the data in recent decades. We choose a variety of models, including ordinary linear regression model, Bayesian ridge regression model, elastic network regression model, support vector machine regression model and gradient enhanced regression model. In this paper, a prediction model based on machine learning is established without considering the epidemic situation. First of all, preprocess the data; secondly, determine the index system; then, based on the regression algorithm of machine learning, compare and select the optimal model for data fitting and prediction; finally, use linear regression to predict the indicators of the next four years.

Keywords: Machine learning, Regression algorithm, Comparative test, Economic influencing factors.

1. Introduction
The US presidential election is held every four years. 2020 is the year of the US presidential election. Republican candidate Donald Trump and Democratic candidate Joe Biden will compete for president. Candidates of the two parties have different political stances and policy agendas in finance and trade, economic and financial governance, and other key development areas [1] (such as new crown pneumonia prevention measures, infrastructure, taxation, environmental protection, medical insurance, employment, trade, immigration, education, etc.)

This paper selects (GDP) as an index to measure the economic situation of the United States, analyses the relationship between GDP and other fields, determines the main factors affecting GDP, and explores different presidents' advocacy of quantitative adjustment of field data to observe the possible impact on the American economy.
2. Model and data preprocessing

2.1. Data process
Data preprocessing should include four aspects: missing value processing, outlier processing, elimination of constant variables, elimination of collinearity features, but the data set used in this question does not require elimination of constant variables, so data preprocessing will be performed from the following three aspects.

(1) **Missing value processing**: Analysing the data set found that most features have a small percentage of missing values, and only a few features have a high percentage of missing values. This article deals with missing values as follows: features with a missing value ratio of more than 20% are directly eliminated, and other missing values Numerical features with fewer values are filled with median or average values [2].

(2) **Outlier processing**: Some data in the data set obviously deviate from the rest of the observations of the sample it belongs to, and such data needs to be eliminated.

(3) **Eliminate collinearity**: Collinear features refer to features that are highly related to each other. In the field of machine learning, high variance and low model interpretability lead to a decline in generalization ability on the test set. This article uses SPSS to eliminate the collinearity feature.

2.2. Index selection based on the incidence matrix
When measuring the U.S. economy, this article constructs an indicator system to explore the possible impact of the election of different candidates on the U.S. economy. When establishing the indicator system, this article contains as many sample indicators as possible, so as to better classify the indicators and find out the importance of each indicator.

This paper uses the Pearson correlation coefficient to measure the linear relationship between the distance variables. The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the variables:

\[
\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}
\]  
(1)

The above formula defines the overall correlation coefficient, and Greek lowercase letters are commonly used as representative symbols. Estimated sample covariance and standard deviation, correlation coefficients obtained samples (Sample Pearson coefficient), common English lowercase r Representative:

\[
r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}
\]  
(2)

The value range of \( r \) is [-1,1]. The greater the absolute value of the correlation coefficient, the higher the correlation degree, that is, the closer the correlation coefficient is to 1 or -1, the higher the correlation degree between the two, the closer the correlation coefficient is to 0. The lower the correlation between the two.

\[
r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{X_i - \hat{X}}{\sigma_X} \right) \left( \frac{Y_i - \hat{Y}}{\sigma_Y} \right)
\]  
(3)
Of Pearson statistical test product moment correlation coefficient was calculated t statistic, calculated as:

$$t = \frac{r - 0}{S_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

(4)

The Pearson correlation coefficient is introduced here to describe the correlation of each index, and the correlation matrix between each index is calculated:

$$D = \begin{bmatrix}
    r_{1-1} & \ldots & r_{1-j} & \ldots & r_{1-50} \\
    \vdots & & \vdots & & \vdots \\
    r_{j-1} & \ldots & r_{j-j} & \ldots & r_{j-50} \\
    \vdots & & \vdots & & \vdots \\
    r_{50-1} & \ldots & r_{50-j} & \ldots & r_{50-50}
\end{bmatrix}$$

(5)

The matrix $r_{i-j}$ indicates the degree of relevance of each indicator. Among them, i and j are numbers arranged in order from left to right.

3. Regression model based on machine learning

In this paper, regression analysis is used to establish the relationship model [3] between the dependent variable GDP and the independent variables, namely various indicators $(X_1, X_2, \ldots, X_n)$ to predict the economic situation of the US in the next four years. We selected a variety of regression algorithms to predict the data respectively, compared and selected the algorithm with the best result.

3.1. Bayesian Ridge regression model

If there is a correlation between the features in the model, this will increase the complexity of the model, and the interpretation ability of the entire model has not improved. At this time, we have to perform feature selection [5]. During selection, generally three ways: subset selection, dimensionality reduction, contraction mode (Shrinkage Method), also known as regularization (Regularization), including ridge regression a lasso return.

Ridge Regression (Ridge Regression) adds a regular term on the basis of the squared error:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\omega_jx_{ij}\right)^2 + \lambda\sum_{j=0}^{p}\omega_j^2, \lambda > 0$$

(6)

By determining $\lambda$ the value, a balance can be achieved between variance and bias: as $\lambda$ the increase, the model variance decreases and the deviation increases.

For the $\omega$ derivative, the result is:

$$2X^T(Y - XW) - 2\lambda W$$

(7)

Let it be 0, the value that can be obtained $\omega$:
\[
\hat{\omega} = \left( X^T X + \lambda I \right)^{-1} X^T Y
\]  

(8)

3.2. Elastic Net Regression Model
Flexible Network (Elastic-Net) is a linear model, which was used in the objective function while L1, L2 penalty term [4]. This combination not only learns a sparse model (similar to Lasso), but also maintains the regular properties of ridge regression. The objective function is:

\[
\min_\omega \frac{1}{2n} \| X\omega - Y \|^2 + \alpha \rho \| \omega \|_1 + \frac{\alpha (1-\rho)}{2} \| \omega \|_2
\]  

(9)

In the elastic network, we use parameters to \( l_1 \_ \text{ratio} \) control the convex combination of l1 and l2. When there are multiple related features in the model, Lasso tends to choose one of them immediately, while Elastic-Net chooses all of them. You can think of Elastic-Net as a compromise between the Lasso and Ridge models.

3.3. Support vector machine regression model
SVM objective regression model is to have each point in the training set to try to fit a linear model [6]:

\[
y_i = \omega \cdot \phi(x_i) + b
\]  

(10)

SVM defines a constant deviation \( \varepsilon \textgreater 0 \). The loss is calculated only when the absolute value of the difference between the model output and the true output is greater than \( \varepsilon \), which is equivalent to taking the center and constructing an interval band with a width of \( 2\varepsilon \). If the sample falls within the interval band, the prediction is considered correct.

The loss function measurement of the SVM regression model is:

\[
\text{error}(x_i, y_i) = \begin{cases} 
0 & \text{if } |y_i - \omega \cdot \phi(x_i) - b| \leq \varepsilon \\
|y_i - \omega \cdot \phi(x_i) - b| - \varepsilon & \text{if } |y_i - \omega \cdot \phi(x_i) - b| > \varepsilon 
\end{cases}
\]  

(11)

The objective function of the SVM regression model is:

\[
\min_{\omega, b} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{m} \max(0, |y_i - \omega \cdot \phi(x_i) - b| - \varepsilon, 0)
\]  

(12)

The regression model \((x_i, y_i)\) introduces a slack variable to each sample, and both sides of the absolute value need a slack variable, which is defined as \( \xi_i, \bar{\xi}_i \), then the loss function measurement of the SVM regression model becomes after adding the slack variable:

\[
\min_{\omega, b} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{m} (\xi_i + \bar{\xi}_i)
\]

s.t. \( -\varepsilon - \bar{\xi}_i \leq y_i - \omega \cdot \phi(x_i) - b \leq \varepsilon + \xi_i \)

\( \xi_i \geq 0, \bar{\xi}_i \geq 0, i = 1, 2, \ldots, m \)

(13)
3.4. Gradient enhancement regression model

The earliest implementation of the Boosting method is actually Adaptive Boosting or AdaBoost [7]. The gradient booster is an extension of AdaBoost under the statistical framework. The basic models used by AdBoost and the gradient booster are decision trees. And the decision tree in the gradient enhancement machine only needs to be a regression tree, that is, a prediction of continuous values. The gradient enhancement machine algorithm is helpful to reduce the variance and bias of the model, so the effect is usually more obvious.

In the application of gradient enhancement machine, gradient descent is to find the best direction in the loss equation space. In the implementation of the gradient enhancement machine, it is assumed that the prediction function after m iterations is \( F_m(x) \), and the corresponding loss function is \( L(y, F_m(x)) \). In order to achieve that the fastest decrease loss function object, it should be configured in a direction of the gradient of the loss function decline \( m + 1 \) iterations submodel function.

4. Solution of regression analysis model

4.1. Solve the incidence matrix

Use SPSS using Pearson correlation coefficient is a measure of the linear relationship between a given variable pitch, constructed 50 * 50-dimensional correlation matrix. In this question, US GDP is used as the dependent variable, so only the correlation between GDP and other indicators needs to be analyzed.

| Table 1. The final indicator system |
|-----------------------------------|
| **A system of indicators affecting the U.S. economy** |
| First level indicator | Secondary indicators | Environmental protection | Total crude oil production |
| time | time | | |
| Finance and Trade | Budget expenditure | | New employment: education and health services |
| | Budget revenue | | employed population |
| Economic | GDP: current price | | Federal government current tax receipt |
| | Total personal income | | medical insurance |
| | Total personal expenditure | | health care |
| | Minimum hourly wage | | education |
| Domestic Trade | Retail and food service sales | | Education service |
| Foreign trade | Import Price Index: Month-on-year | | transport and water infrastructure expenditure |
| | | | financial |
| | | | Dow Jones Index |
| | | | currency |
| | | | M2 |
4.2. Solve data fitting
Machine learning to fit the data, obtained as the learning curve of each model, respectively, in order Bayesian Ridge, Linear Regression, Elastic Net, the SVR, GBR learning curve:

Figure 1. Learning curve of each model
Figure 2. shows the comparison of the fitting effects of five models:

![Figure 2. Comparison of fitting effects of five models](image)

It can be seen that in addition to the poor SVR fitting effect, other models are well fitted to the true value. In order to get the best model, we further processed these five models as follows.

Table 2 shows the 6-fold cross validation score (cross validation result):

|       | 0     | 1     | 2     | 3     | 4     | 5     |
|-------|-------|-------|-------|-------|-------|-------|
| Bayesian Ridge | 0.966572 | 0.926781 | 0.676108 | 0.885698 | 0.998588 | 0.24563 |
| Linear Regression | 0.682891 | 0.963519 | 0.398161 | 0.805192 | 0.849411 | -0.039133 |
| Elastic Net | 0.398651 | 0.961873 | 0.952109 | 0.840686 | 0.976893 | 0.958356 |
| SVR | -40.537336 | -17.139667 | -13.109858 | -26.804683 | -65.724334 | -248.053957 |
| GBR | -5.146209 | 0.802728 | -0.429329 | 0.409609 | 0.802409 | -6.334683 |

Table 3. Cross validation result

|       | Ev             | Mae            | Mse             | r2             |
|-------|----------------|----------------|-----------------|----------------|
| Bayesian Ridge | 0.999038 | 105.117781 | 1.735209+04 | 0.999038 |
| Linear Regression | 0.999932 | 25.671591 | 1.220240+03 | 0.999932 |
| Elastic Net | 0.999821 | 46.368514 | 3.236754+03 | 0.99821 |
| SVR | 0.001195 | 3612.154207 | 1.807223e+07 | -0.001763 |
| GBR | 1.000000 | 0.112748 | 2.144818e-02 | 1.000000 |

The symbols in the table 3 are explained as follows:

- **ev** is explained_variance_score: explain the variance score of the regression model. Its value range is [0,1]. The closer to 1, the more the independent variable can explain the variance of the dependent variable. The smaller the value, the worse the effect;

- **mae** is mean_absolute_error: Mean Absolute Error (MAE), which is used to evaluate the degree of closeness between the predicted result and the real data set. The smaller the value, the better the fitting effect;
mse is mean_squared_error: Mean squared error (Mean squared error, MSE), this indicator calculates the mean value of the sum of squares of the error between the fitted data and the corresponding sample points of the original data. The smaller the value, the better the fitting effect;

r2is r2_score: coefficient of determination, which means to explain the variance score of the regression model. Its value range is [0,1]. The closer to 1, the more the independent variable can explain the variance of the dependent variable. The smaller the value, the worse the effect.

Based on the above data, it can be found that the GBR fitting effect is the best, so we select the data fitted by the GBR model.

5. Conclusion
This paper explores the economic effects of different regression models based on machine learning methods.

Data preprocessing is the first step in establishing a machine learning model, which can improve the quality of the data, which helps to improve the accuracy and performance of the subsequent learning process. In this paper, we will use reasonable methods to deal with missing values, outliers and multicollinearity. Then, establish an index system that affects the American economy. There are many indicators that affect the economy. We selected multiple models, including ordinary linear regression model, Bayesian Ridge regression model, elastic network regression model, support vector machine regression model and gradient enhancement regression model, compared the test results and selected the best model.

References
[1] Zhang Lixin. Research on feature selection of high-dimensional data and integrated learning based on feature selection [D]. Tsinghua University.
[2] Zhang Zongtang, Wang Sen, Sun Shilin. An ensemble learning algorithm for imbalanced data classification [J]. Journal of Shandong University (Engineering Science Edition), 2019, 49 (4): 32 - 37.
[3] Liu Lu. Research on the formation mechanism of contemporary American financial crisis [D]. 2012.
[4] Michael JJ Douglass. Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edition by Aurélien Géron [J]. Physical and Engineering ences in Medicine: 1 - 2.
[5] Jiao Ruiqiang. The choice of regularization parameters in Bayesian penalty regression [D]. 2017.
[6] Su Baoxia. The basic model of the spread of new coronary pneumonia [J]. 2020. [7] Tang Xiaoyan. The effect of foreign trade on China's economic growth [J]. China Business Review, 2018, 000 (023): 62 - 63.