ISSN:2790-1661 Volume-5-(2023)

Analysis of international stock data based on python

Ping Du[1,a],Yanfen Shen[2,b]

[1]Building Information Department Guangdong Construction Polytechnic Guangzhou, China

[2]Building Information Department Guangdong Construction Polytechnic Guangzhou, China

[a]cassiedu@foxmail.com,[b]26438895@qq.com

Abstract. With the continuous development of computers, the gradual opening of the financial market, and China's increasing investment in scientific and technological innovation, artificial intelligence and big data analysis and other technologies have begun to be used in various financial industries such as the stock and bond market. This paper analyzes the correlation of stock indexes in Istanbul and other countries through python data analysis, and finds that the stock markets in various countries have relatively large correlation. The regression prediction analysis of Istanbul stock data is carried out, and three models of linear regression, ridge regression and gradient regression are constructed for prediction, and the performance comparison is made.

Keywords: Stock; correlation; linear regression; Ridge regression; Gradient regression.

Introduction

As a barometer of the economy, the stock market reflects the current situation of the national and world economy. Its powerful ability of financing funds and allocating resources plays a vital role in the construction and development of the economy. Big data and artificial intelligence technology are increasingly used in stock analysis and prediction. As an important part of the world economy, the stock market has become a very important and meaningful thing to diagnose and analyze the international stock market, dig the linkage law of the stock market among various countries, and predict the future trend.

The famous American futurist Alvin Toffler put forward the concept of big data in his book The Third Wave in 1980 [1]. In 2011, Science published a special issue called "Dealing With Data"; In 2012, the US announced its "Big Data Research and Development Plan", which aims to enhance the ability to extract information from massive amounts of data. At present, the securities industry is booming, which is characterized by a large amount of data, fast change rate of data, and diverse data types, but there are a lot of redundancy and noise, and a lot of data needs to be processed such as data cleaning, refining and data fusion. Big data analysis techniques such as data conversion protocol, visualization technology and knowledge computing are used to analyze securities data. On the other hand, machine learning prediction model is combined to forecast and analyze stock price data with time sequence [2]. Fully integrate the stock and securities industry with big data technology to better guide the stock and securities market and promote economic development.

1. Data sources

This paper analyzes 536 days of international stock data, including "Istanbul Index daily return", "S&P Index daily return", "German Composite Index daily return", "FTSE daily return", "Nikkei daily return", "Sao Paulo Stock Index daily return", "MSCI Europe Index daily return", "Istanbul Index daily return". "Msci Emerging Markets Index Daily Return" A number of countries' stock indexes. The data is shown in Figure 1. In the early stage, the data was processed to remove the weight and missing value, and the data was standardized.

	Istanbul daily income	Standard & Poor's daily income	Germany's comprehensive daily income	British rich time income	Nikkei daily income	Daily income of Sao Paulo securities	Morgan stanley capital international's European daily income	Daily income of emerging markets in morgan stanley capital international
0	0.038376	-0.004679	0.002193	0.003894	0.000000	0.031190	0.001698	0.028542
1	0.031813	0.007787	0.008455	0.012866	0.004162	0.018920	0.011341	0.008773
2	-0.026353	-0.030489	-0.017833	-0.028735	0.017293	-0.035899	-0.017073	-0.020015
3	-0.08476	0.003391	-0.011726	-0.000455	-0.040051	0.028283	-0.005561	-0.019424
4	0.009658	-0.021533	-0.019873	-0.012710	-0.004474	-0.009764	-0.010989	-0.007802
……								
531	0.013400	0.006238	0.001925	0.007952	0.005717	0.018371	0.006975	0.003039
532	0.015977	0.003071	-0.001186	0.000345	0.002620	0.001686	-0.000581	0.001039
533	-0.001653	0.001923	0.002872	-0.000723	0.000568	0.000628	0.000572	0.006938
534	-0.013706	-0.020742	-0.014239	-0.011275	0.001358	-0.011942	-0.012615	-0.009058
535	-0.019422	0.000000	-0.00473	-0.002997	-0.017920	-0.012252	-0.005465	-0.014297

Figure 1 Daily return data of multi-country stock indexes

2. Correlation analysis

The correlation analysis of the data of each country reveals the linkage law between international stock markets. Thus brings a kind of inspiration for financial analysis and stock investment. Correlation coefficient quantitatively describes the degree of correlation between and two variables, that is, the greater the absolute value of correlation coefficient, the greater the degree of correlation, and vice versa [3]. If the correlation coefficient is 0, then the two variables are completely unrelated.

The correlation coefficient is usually represented by a letter and measures the linear relationship between two variables. The announcement is as follows:

Where, is the covariance of and, the variance of PI, and the variance of PI. After the correlation analysis of each variable data, the following chart is obtained.

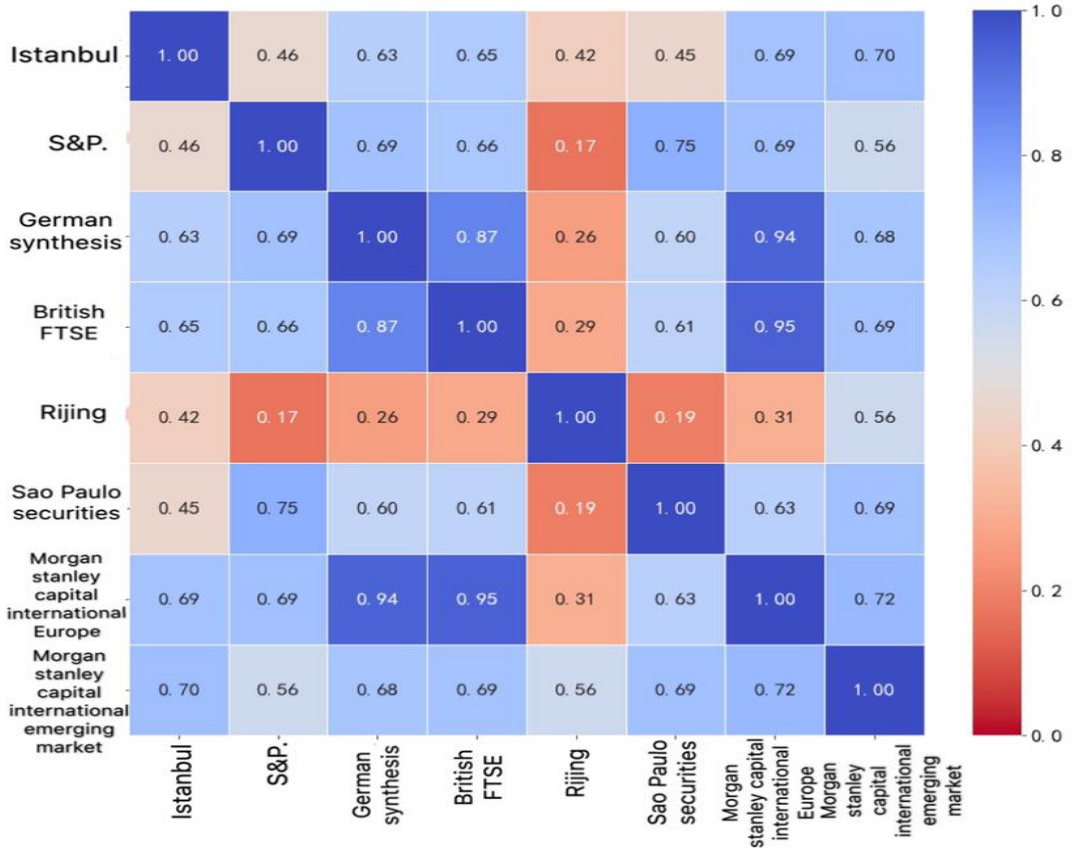

Figure 2 Correlation analysis of daily returns of stock indexes in different countries

It can be seen from Figure 2 that the correlation coefficient of "daily return of stock index" between different countries is basically above 0.6, or even above 0.9. But we also see an exception in Japan, where the correlation between stocks and other countries is low. Taking the relationship between the daily returns of the index of Establ and other countries as an example, we can see that the correlation coefficient between it and Germany is 0.63, and that between it and the UK is 0.65, indicating a relatively large correlation. The correlation with the US, Japan and Brazil is relatively low, between 0.4 and 0.5.

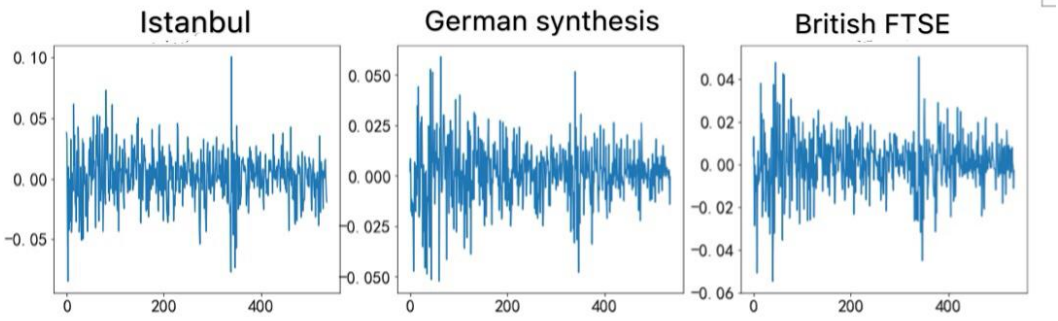

Figure 3 Data line plots of three countries

Figure 3 is the line chart of daily index returns of "Istanbul", "Germany" and "America". It can be seen that the trend of rise and fall of the three countries is very similar, but the similarity between the UK and Germany is higher. The correlation coefficient has been calculated before as 0.87, which indicates that the stock markets of these two countries have a great influence on each other.

3. Linear Regression

How to accurately describe and forecast the volatility of stock return? This has always been one of the hot issues discussed in the field of finance. To grasp the characteristics and trends of stock return volatility is of great theoretical and practical significance for investors to measure, avoid and

manage stock market risks. Therefore, for a long time, many scholars have used various forecasting models to empirically analyze and forecast the volatility of stock return, hoping to get useful enlightenment and follow the rules [4].

Regression is a statistical analysis method to study the dependence of dependent variables on independent variables. The purpose is to estimate or predict the mean value of dependent variables by a given value. Regression studies the relationship between dependent and independent variables, and can be used to discover causal relationships between variables, as well as to make predictions.

Linear regression can combine many variables to make an optimal forecast, while separating the effects of each variable. In linear regression, when there is more than one factor affecting the independent variable, if there is one factor, it can usually be expressed as the following linear relation [5] :

We want to analyze the impact of "Istanbul" data on the world stock market, so it should be used as the dependent variable and the data of other countries as the independent variable. That is, 70% of the data (375 pieces) should be used to train the linear regression model, and 30% of the data (161 pieces) should be used to verify the model. The comparison between the 30% predicted data and the real data is shown in Figure 4.

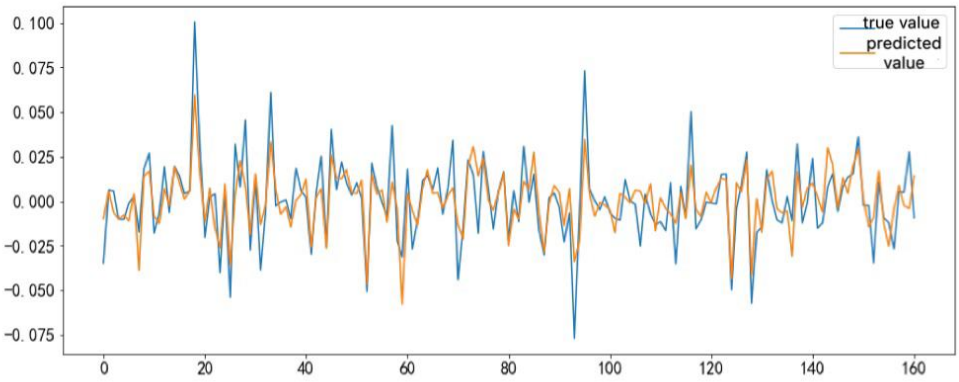

Figure 4 Comparison of predicted data and real data

The size of the predicted data and the real data is compared, and the result is shown in Figure 5. Therefore, the predicted data is 50.93% if it is greater than or equal to the real data, and 50.93% if it is less than the real data.

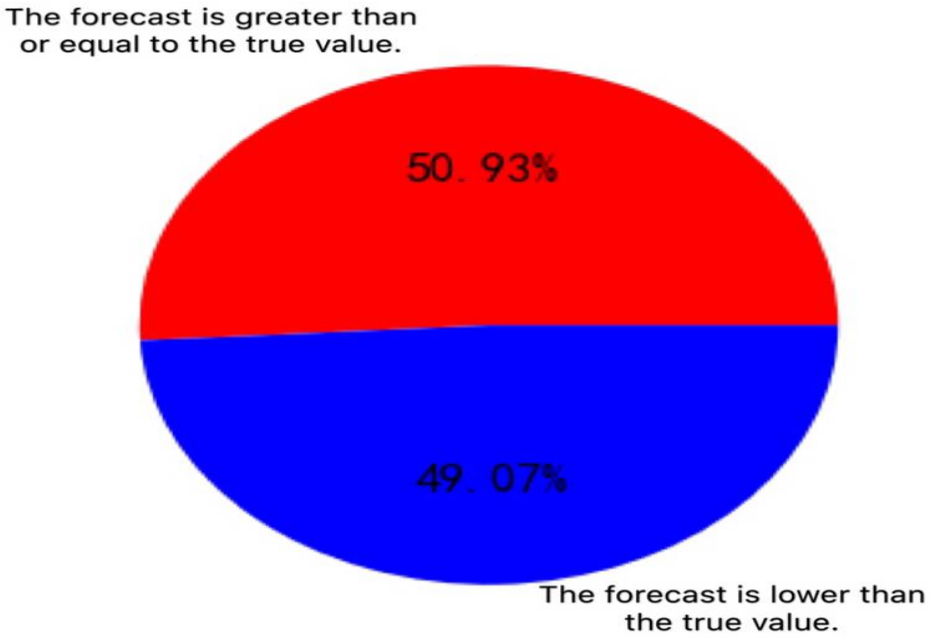

Figure 5 Comparison between predicted data and real data

At the same time, gradient regression and ridge regression were used to analyze and forecast the data, and their performance was compared as shown in the figure. In ridge regression, we adjusted the value of the parameter alpha, and the regression result would increase with the decrease of the alpha value. Here, we set the alpha value to 0.1.

4. Regression model evaluation [6]

The fitting coefficient R2 and the mean square error (MSE) are the main tools to evaluate the model.

4.1 Fitting coefficient R2[7]

R2 is defined as a set of observed values and a set of predicted values for some set of variables. The larger R2 is, the better the model performance will be. R2 is calculated by the following formula:

4.2 Mean Square Error (MSE)

The formula for calculating mean square error is as follows:

In the above two formulas, m is the sample size, is the predicted value and is the observed value.

The performance evaluation indexes of linear regression, Ridge regression and gradient regression models are shown in Figure 6. The linear regression model has the smallest mean absolute error, the linear regression model has the smallest mean square error, the Ridge regression model has the smallest median absolute error, the gradient regression model has the smallest explanatory variance, and the linear regression has the largest R-square value. Overall, the performance of linear regression is the best, followed by ridge regression.

	Average absolute error	mean-squared error	Median absolute error	Explanatory variance value	R square value
Gradient regression	0.011395	0.00242	0.009739	0.547218	0.547218
Linear regression	0.010551	0.000196	0.007933	0.633928	0.633915
ridge regression	0.010590	0.000198	0.007307	0.629610	0.629606

Figure 6 Comparison of the three regression analyses

Conclusion

Big data technology is widely used in the field of stock and securities. It is a hot research topic to use machine learning algorithm to predict market prices [8]. Through the analysis and verification of this paper, the results show that the stock markets of different countries influence each other, and the development trend and fluctuations of each stock market conduct each other, so that the main stock markets have a significant linkage feature. The correlation between countries is analyzed and the correlation coefficient is found to be large. The visualized results show that they have the same long-term trend and influence each other greatly. The stock markets in Germany and Britain move in the same direction. At the same time, this paper constructs several regression prediction models for predicting the trend of the international stock market, and analyzes its performance, which provides a reference for the trend of the international economy.

References

[1] Alvin Toffler. The Third Wave [M]. Beijing: Citic Press, 1980. (in Chinese)

[2] Yanhui Liu. Application of Big Data Technology in Securities Trading [J]. Modern Commerce and Industry, 2018(18):152-153.

[3] Guocheng Cai. Research on Stock Price Prediction based on Support Vector regression [D]. Hangzhou Dianzi University, 2009:27

[4] Wei Shen, Dongxiao Niu. A Comparative Study on Prediction Models of Stock Return Volatility [J]. Shopping Mall Modernization, 2009(4):354-355.

[5] Research on Influencing Factors of college Enrollment Quality Based on Multiple Linear Regression and Lasso Regression [J]. Journal of Ludong University: Natural Science Edition, 2022, 38(4):7.

[6] Qiaoying Wang. Comparison of standard error and Determinability coefficient of regression estimation [J]. Statistics and Decision, 2006(23):1.

[7] Guangyu Yang. Comparative analysis of Five Stock Prediction Methods Based on Stock Correlation [J]. Modern Business, 2022(29):4.

[8] Tianhua Lin, Qianqian Zhang,Xuyang Qi , Xia Zhao. Research on securities Big Data analysis [J]. Computer Technology and Development, 2020, 30(10):8.