Research on the investment ratio allocation of gold and bitcoin based on price prediction model and Markowitz model

Run Zhu[1], Ziyang Jin[1], Yukun Zhou [2], Yuze Fang[2], Tianyu Hong[2, *]

[1]College of Economics and Management, China Agricultural University, Beijing, China, 100083

[2]College of Engineering, China Agricultural University, Beijing, China, 100083

*Corresponding author: vj162774@163.com

Abstract. With the development of modern society and economy, people are more willing to trade volatile assets to maximize their returns. People often want to use the historical data they have to predict the future price trend of investment products. This article uses the price data of gold and bitcoin as an example to make predictions. Gold has significant seasonality, so this paper combines the Seasonality and Trend decomposition using Loess (STL), and the time series model ARIMA to build the STL-ARIMA model for forecasting, and the RMSE of STL-ARIMA (5,1,5) was determined to be 17.0634. Bitcoin is affected by volatility aggregation and momentum effects. The LSTM model combined with GRACH (1,1) and the momentum effects equation achieves 91% accuracy in forecasting results.

Keywords: STL-ARIMA; LSTM; Volatility aggregation; Momentum effect.

1. Introduction

With the continuous development of the social economy, people have various ways to use or distribute their assets. Over the past decade, big data networks and VR(virtual reality) have brought great changes to the way of economic development, thus various investment products continue to emerge. People are no longer limited to deposited cash in the bank and buy related bonds. Investing in the stock market, buying gold, or putting cash into the emerging virtual currency market are also very popular and sought after. But emerging investment products tend to be more risky, and correspondingly, but the return rate is also greater[1].

In the current trading investment market, market traders often buy and sell volatile assets because these relatively unstable products are more likely to yield large instantaneous gains. This paper want to predict the price of investment products for traders by using gold and bitcoin as examples. Since the price of gold varies seasonally, we need to consider the impact of different dates.[2]

Therefore, this paper focused on building a price forecasting model to predict the price trend of gold and bitcoin. Meanwhile, some optimization of the model were done to fit more situations.

2. Model assumptions and notation

2.1 Assumptions

1. The transactions of gold and bitcoin are received in real time without delay.

2. Gold trading has a fixed start and end time. It is considered that the given gold price is the price at the end of the transaction, that is, the price at the beginning of the transaction is the price at the end of the previous day.

3. Because Bitcoin does not have a time to start trading and end trading in practice, assuming a one-hour "market closure", 11:30 is regarded as the closing time of trading, and trading starts at 12:30. One hour allows us to provide the model is calculated for investors to analyze and make decisions.

2.2 Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Symbol	Desription
AR	Auto Regression
MA	Moving Average
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
p	The number of autoregressive terms
q	Sliding average number of items
d	Number of differences
Tv	Trend component
Sv	Seasonal component
Rv	Remainder component
Z	Integer

3. Model construction and solving

3.1 The explanation of the STL-ARIMA model

The price of gold does not change much over time, and the future price can be regarded as the result of the influence of the recent price, so the time series model can be used for forecasting. Series stationarity is a prerequisite for time series analysis, and the price fluctuation of gold is small, so time series can be used for analysis. A time series is a series of data points indexed (or listed or plotted) in chronological order. Modeling with a time series enables predictions about the future by using known data from the past. Common time series forecasting methods include moving average forecasting method, exponential smoothing forecasting method, seasonal exponential forecasting method and ARMA model. The autoregressive moving average model ARMA(p, q) is one of the most important models in time series. It is mainly composed of two parts: AR stands for the p-order autoregressive process, and MA stands for the q-order moving average process[3].

In the time series, the ARIMA model is based on the ARMA model with a difference operation. The difference is a method used by the time series to eliminate periodic factors. It mainly performs linear subtraction on the data of equal periodic intervals. Since the price of gold has seasonal and cyclical regular changes, we choose the time series ARIMA model, and make the first prediction by eliminating the cyclical influence, and then take the cyclical factors into account, so that we can get the total prediction curve. The ARIMA model emphasizes stationarity. The so-called stationarity requires that the predicted fitting curve obtained through the time series of gold price samples every day can still continue in accordance with the existing form for a certain period of time in the future, with certain regularity[4].

The ARIMA(p,d,q) model is expressed as:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t, d\in Z, d>0. \tag{1}$$

3.2 Steps build

3.2.1 Data cleaning and processing

This paper used the mean value to fill up the data, where i is the number of the school, j is the number of the indicator.

$$NULL_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} = \bar{x}_j (i = 1, 2, ..., n; j = 1, 2, ..., p) \tag{2}$$

Then using the equation below to standardize the data.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{\sqrt{Var(x_j)}} \tag{3}$$

$$Var(x_j) = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \tag{4}$$

Finally this paper sorted the daily price of gold and obtained the time series data of gold.

3.2.2 Data decomposition

In the process of time series analysis, we take into account the periodic seasonal characteristics of gold, use EViews to process the practical series data, and then use the STL algorithm to calculate. STL (Seasonal-Trend decomposition procedure based on Loess) is a common algorithm in time sequence. Based on LOESS, the data at a certain moment is decomposed into trend component, seasonal component and remainder component), the formula is as follows:

$$Y=Tv+Sv+Rv \quad v=1,\ldots,N \tag{5}$$

After having decomposed the different components, fitting each component with a time series model can help improve the accuracy of the predictions. As Figure 1 shown.

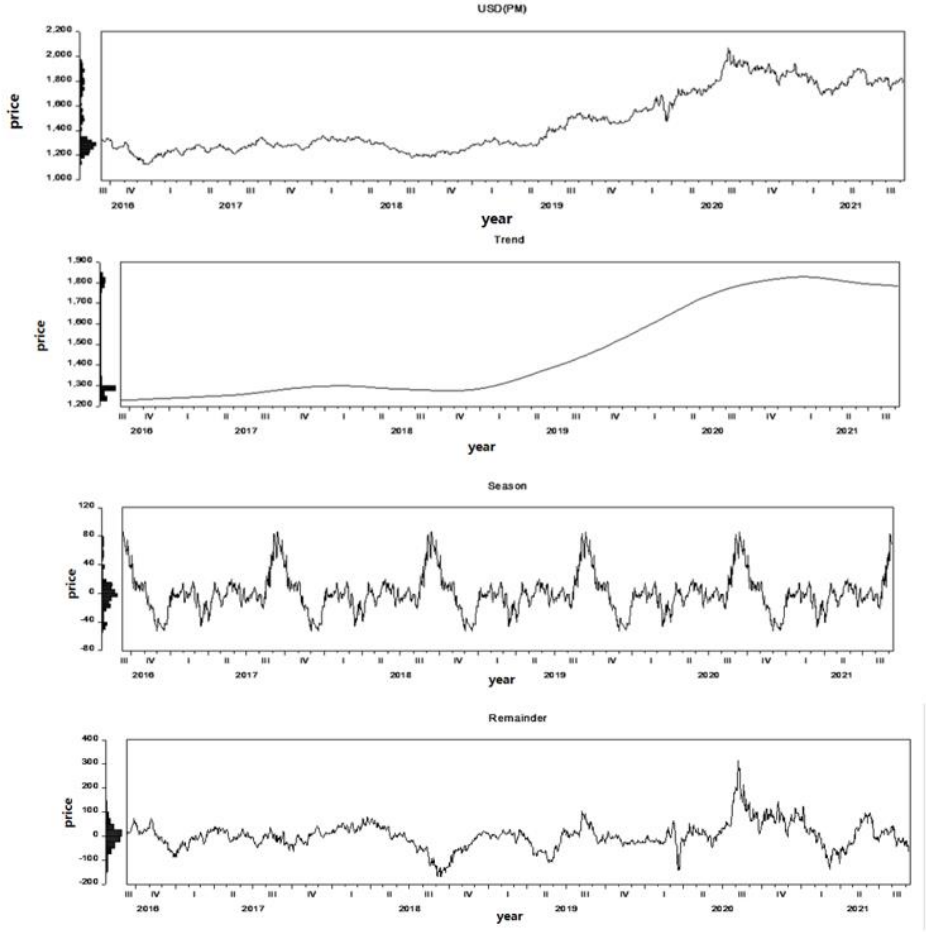


Fig. 1 Separate Data

3.2.3 Observation

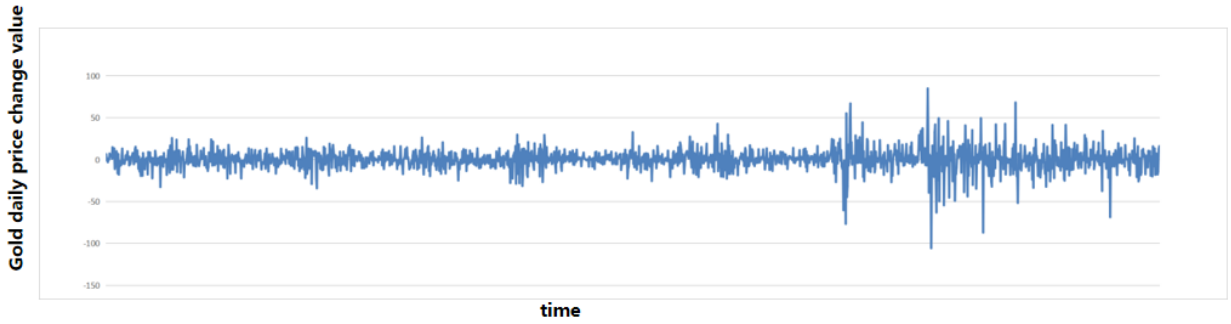


Fig. 2 Gold Price Fluctuation

As can be seen from the Figure 2, the price fluctuation of gold was small. The data can be converted into a stationary time series by performing a simple first-order difference operation on the data.

3.2.4 Calculate ACF and PACF

After the second step of processing, a stationary time series was obtained, and then both its autocorrelation coefficient ACF and partial autocorrelation coefficient PACF for the stationary time series respectively was calculated.

The autocorrelation function refers to the linear relationship between the sequence value X_t at any time t (t=1,2,3…n) and its own lag k value X_{t-1}. The autocorrelation coefficient represents the correlation coefficient value between time series with an interval of K, and the formula is as follows:

$$\rho_k = \frac{corr(X_{t+k}, X_t)}{\gamma_0} = \frac{\gamma_k}{\gamma_0}, k \in Z. \tag{5}$$

k is the order of the interval.

The partial autocorrelation is the partial correlation coefficient after deducting the influence of ρ_1 to ρ_{p-1}, which is ϕ_p, and the formula is as follows:

$$pacf(k, p) = \frac{\sum_{j=1}^{p}(\phi_j \rho_{j-k}) - \sum_{j=1}^{p-1}(\phi_j \rho_{j-k})}{\rho_{p-k}} = \phi_p \tag{6}$$

This model visualized the autocorrelation plot with the lag value on the x-axis and the autocorrelation coefficient on the y-axis, as can be seen in Figure 3.

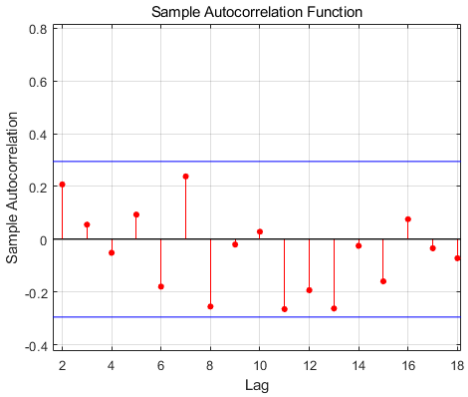


Fig. 3 ACF

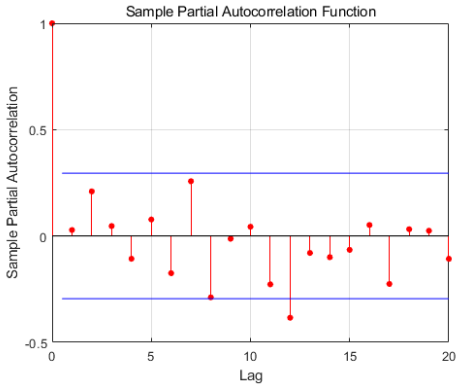


Fig. 4 PACF

A partial autocorrelation diagram was drawn as shown in Figure 4.

Next, this paper determined the number of strata by analyzing the autocorrelation and partial autocorrelation diagrams. First, judging which model it is through the two graphs of ACF and PACF. Then making decision by using the following Table 2.

Table 2. Judgment parameter

Model	AR(p)	MA(q)	ARMA(p,q)
ACF	Decay to zero	Q-order post-censoring	The decay tends to zero after the q-order
PACF	P-order post-censoring	Decay to zero	The decay tends to zero after the p-order

The maximum possible values of p and q can be obtained by using the data of five years in history. Due to the large amount of data, we use the exhaustive method to select, and we get that the level p of our model is 5 and the order q is 5.

3.2.5 Model testing and normal distribution

Therefore, after the above calculation and analysis, the ARIMA model this paper used is ARIMA(5,1,5).

For autocorrelation, by using the Durbin-Watson test, or D-W test for short, which is currently the most commonly used method for testing autocorrelation. When the DW value is significantly close to 0 or 4, there is autocorrelation, and when it is close to 2, there is no first-order autocorrelation. After inspection, the DW value of our model is 1.0525, and it can be concluded that the autocorrelation is good[5].

After that, it is verified by observing whether the data conforms to the normal distribution and using the QQ plot trained in MATLAB. The QQ plot can visually verify whether a set of data comes from a certain distribution, or whether a set of data comes from the same distributed. One dataset corresponds to the x-axis, the other corresponds to the y-axis, and a 45-degree reference line is made. If the data of the two datasets come from the same distribution, the points will fall near the reference line. As far as this question is concerned, the expected value of our prediction of the price of gold is the x-axis, and the actual price observation value is the y-axis. The two are close to the 45-degree reference line, indicating that the model's prediction effect is good. Sample Data versus Standard Normal is shown in Figure 5.

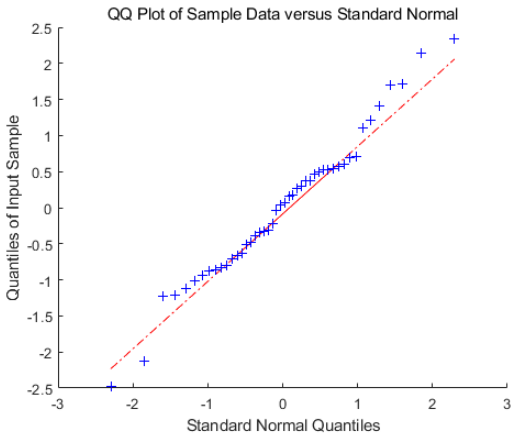


Fig. 5 Sample Data versus Standard Normal

3.3 Predicted outcome

After the above series of model establishment and calculation process, the price trend of gold can be predicted. Since ARIMA fits the data after relevant preprocessing, its predicted value needs to be restored through relevant inverse transformation. The root squared error (RMSE) is used to evaluate how well the model fits within the sample. RMSE is a commonly used measure of the difference between measured values. When using this criterion to discriminate, it is necessary to exclude the influence of "non-predictive" data[6].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{7}$$

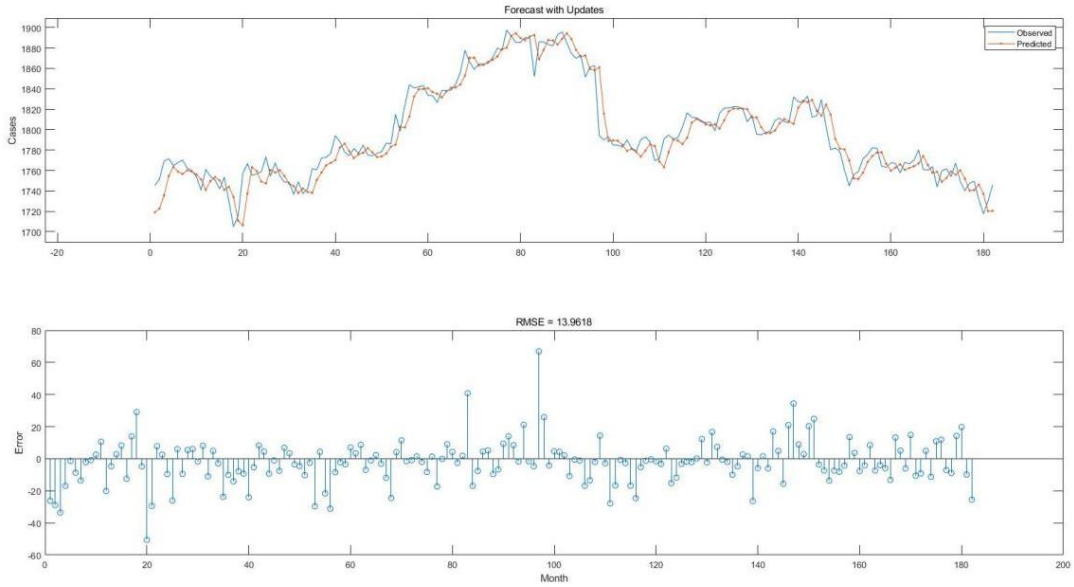


Fig. 6 Forecast results and errors

Observe the fitting effect of the above Figure 6, and the RMSE is 17.0634. It can be seen that the ratio of the difference between the predicted value and the actual value to the actual value is much less than 1, indicating that the model's fitting is good and it can be used to predict the future price trend reasonably and scientifically.

3.4 Improving the LSTM model

This model will use different models for gold and bitcoin to forecast separately. The price of gold is affected by seasonality and periodicity, while bitcoin is not. It will be affected by seasonal changes, and its price stability is poor and there are more uncertain factors. It is more suitable for machine learning with a large amount of calculation to make predictions, such as LSTM, RNN, etc. Therefore, according to the different characteristics of gold and Bitcoin, this paper will select different prediction models for decision-making and judgment, and adopt the improved LSTM model based on volatility aggregation analysis and momentum effect[7].

When researching a problem, LSTM is mainly used to make predictions about future developments, provided that data over a period of time in the past is known. Based on this, this paper tries to use past data to predict the price of a certain day in the past, and then makes a judgment based on the predicted price. The commonly used LSTM model is no longer applicable and it has certain limitations. This paper improves and supplements the LSTM model, and introduces the theoretical knowledge of economics to supplement the volatility aggregation and momentum effects.

3.5 Volatility aggregation

The phenomenon of volatility aggregation was first proposed by Mandlebrot in 1963. He found that high volatility of asset prices always tends to the appearance of continuous aggregation. Volatility refers to a statistical indicator that a financial asset or market rises or falls sharply over a period of time. Commonly used indicators to describe volatility are the variance or standard deviation of the price of the asset or the rate of return. The volatility of many financial time series will show a regular feature, that is, the volatility of assets will change with time, and similar fluctuations are often easy to appear in groups. One large price fluctuation will be followed by another large fluctuations, this situation is called volatility aggregation.

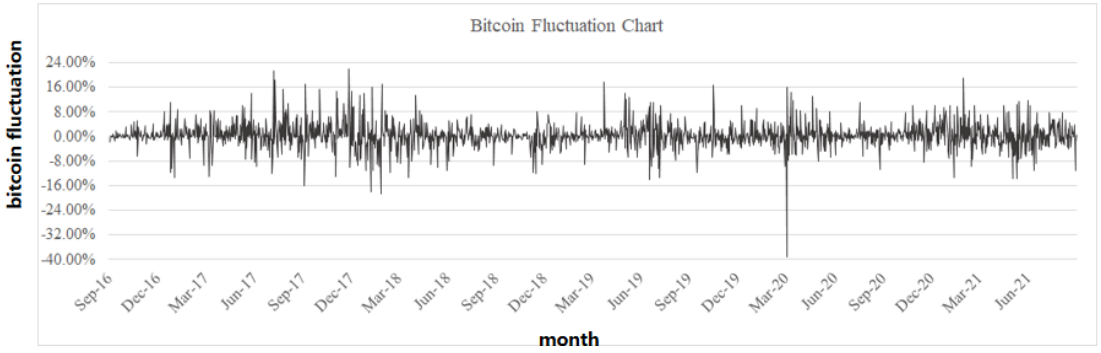


Fig. 7 Bitcoin Fluctuation Chart

Bitcoin Fluctuation Chart is shown in Figure 7**.** This paper uses the GARCH model for volatility analysis and prediction, which can better describe the volatility aggregation phenomenon of Bitcoin time series. It enhances the LSTM model's ability to identify economic phenomena, and can use less data to train the model. The specific form is as follows.

$$a_t = \sigma_t \varepsilon_t \tag{8}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_0 a_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{9}$$

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with mean 0 and variance 1. P is the order of the GARCH model and q is the order of the ARCH model. and $\forall i >$ 0, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, ensuring that the conditional variance is non-negative.

$$\sum_{i=1}^{\max(p,q)}(\alpha_i+\beta_j)<1 \tag{10}$$

Make sure the process is smooth. If both α_i and β_j are significantly greater than zero, it can be shown that there is indeed a positive correlation between the conditional variance and the previous period, that is, the previous volatility will have an impact on the future volatility, and the phenomenon of volatility aggregation can be verified[8].

Taking test set price data of Bitcoin as an example, the final model GARCH(1,1) is obtained after testing, as shown in the Table 3.

Analyzing the fitting effect of the GARCH model, both α and β are significantly greater than 0, and the value of $\alpha + \beta$ is very close to 1, indicating that there is a significant fluctuation aggregation phenomenon in the price fluctuation of Bitcoin.

Table 3. Estimated Results

Coefficient	**GARCH(1,1)**	**GARCH-M(1,1)**
Mean Equation		
Constant(μ)	-0.011317	-0.541699
Risk Premium(λ)	-	0.536232
Variance		
Constant(ω)	0.122520	0.105044
ARCH Effect(α)	0.037931	0.040665
GARCH Effect(β)	0.838434	0.852582
α+β	0.876365	0.893247
Log Likelihood	352.4975	-350.9641
AIC	2.876619	2.882300
SIC	2.943451	2.953340
ARCH-LM test for Heteroscedasticity		
ARCH-LM test	0.4487	0.4242
Probability Chi-square(1)	0.4466	0.4221

3.6 Momentum effect

Momentum effect is generally called "inertia effect". Momentum effect was proposed by Jegadeesh and Titman, which means that the return rate of assets has a tendency to continue the original direction of movement, that is, the rate of return in the past period of time. Higher-yielding assets will still earn higher returns in the future than lower-yielding assets in the past, and assets that have lower-yielding in the past will still earn lower returns in the future. In a robust financial market, the momentum effect can help investors adjust their investment direction and investment amount in a timely manner[9].

Zhang Mao(2015) proposed to use the momentum theorem in physical knowledge and its nonlinear cumulative characteristics to construct an autoregressive equation of price series in the form of momentum, so as to model and predict the price series of the Heng Sheng Index[10]. Learning from the idea, this paper uses the 10-day price of the input of the LSTM model as the input of the autoregressive equation and use the least squares method to estimate the parameters λ and Δt_i, and the residual sum of squares.

$$Q=\sum_{T+1}^{N}\left[\left(x_t-\Delta t_1 x_{t-1}\frac{\lambda}{1}-\ldots-\Delta t_1 x_{t-T}\frac{\lambda}{T}\right)^2+\left(y_t-\Delta t_1 y_{t-1}\frac{\lambda}{1}-\ldots-\Delta t_1 y_{t-T}\frac{\lambda}{T}\right)^2\right] \tag{11}$$

Since Q is a quadratic function about λ and Δt_i, its minimum value always exists, and then the partial derivatives are calculated for λ and Δt_i respectively.

Let the partial derivative be equal to 0, and solve the equation system to obtain λ and Δt_i. Due to the interference of some other factors such as noise, we set the time coefficient Δt_i to 0, which can ensure that Δt_i has physical meaning.

The structure of the improved LSTM model is as Figure 8.

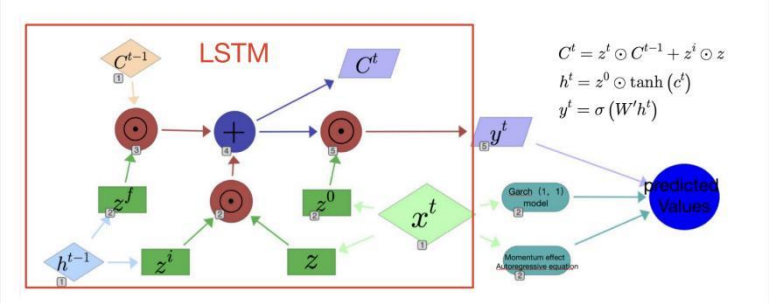


Fig. 8 Optimized LSTM Model

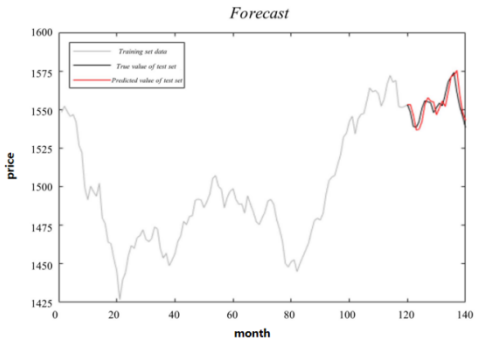


Fig. 9 Bitcoin Price Prediction

The red box in the figure is the traditional LSTM model. Since LSTM is relatively common, it will not be described here. The main improvement here is using the above two models to correct the output of LSTM.

The prediction effect of the final model is shown in the Figure 9.

It can be seen from the figure that the prediction accuracy of this model for this set of data is high, but considering that the price of Bitcoin has changed too much, and to ensure the scientificity of the model, too much historical data cannot be used, so the current 91% accuracy rate can be used, which regarded as the highest accuracy of the model so far.

4. Conclusion

In order to build a quantitative decision-making model (Quantitative Trading Decision Model, QTDM), this paper explores the characteristics of two assets, gold and Bitcoin, mainly about: whether they are affected by seasonality and volatility aggregation analysis, and then create STL-ARIMA model with pre-extracted seasonal factors and LSTM model improved by considering economic principles for price prediction, the overall accuracy is high, but it is still difficult to predict some specific variations. Therefore, this paper uses the predicted values of the above model to calculate the expected return and risk rate of the assets and substitutes them into a multi-objective linear programming model for decision making.

References

[1] Roger S Pulwarty and Theodore S Melis. Climate extremes and [1] Ma Guang, Yang Yashu. Gold price trend, influencing factors and seasonal analysis [J]. Times Finance, 2013(30): 207-208+224.

[2] Wang Chenhuizi, Yang Dong. The value dimension of digital currency from the qualitative perspective of Bitcoin [J]. Learning and Exploration, 2021(06):51-59.

[3] Zhang Jianrong. Quantitative Investment Strategies Based on Herd Effect and Momentum Effect - The Perspective of Small and Medium-sized Board and ChiNext Board [D]. Zhejiang University, 2020. DOI: 10.27461/d.cnki.gzjdx.2020.0019

[4] Zhang Subaru. Using momentum theorem and decay theory to study the autoregressive process of stock prices [J]. Journal of Shaanxi University of Science and Technology (Natural Science Edition), 2015,33(04):178-181+186.

[5] Fang Yiqiu, Lu Zhuang, Ge Junwei. Stock Price Prediction of Combined RMSE Loss LSTM-CNN Model [J/OL]. Computer Engineering and Applications: 1-10 [2022-02-20]. http://kns.cnki.net/kcms/detail/11.2127.TP.20220217.1644.008.html.

[6] Mandelbrot B. New Methods in Statistical Economics[J]. Journal of Political Economy, 1963,71(5):421-440.

[7] Lai Zehua. Research and Improvement of Quantitative Investment Strategy Based on Machine Learning and Volatility Aggregation Analysis [D]. Shanghai University of Finance and Economics, 2020.

[8] Li Cuixia, Chen Yuanyuan. Financial measurement of VaR based on GARCH model based on empirical distribution [J]. Journal of Sun Yat-Sen University (Natural Science Edition), 2021, 60(04): 177-182. DOI: 10.13471/j.cnki. acta.snus.2020.07.31.2020A038.

[9] Huang Chao. Research on some key issues of financial time series mining based on feature analysis [D]. Fudan University, 2005.

[10] Jegadeesh N, Titman S. Returns to buying winners and selling losers: Implications for stock market efficiency[J]. The Journal of Finance, 1993,48(1):65-91.