Report of Final Project

WENFEI YAN
114280105

Introduction

In this project, I predict the value of bitcoin and gold in a time period of five years, from 09/11/2016 to 09/10/2021. A common choice for prediction of time series is Autoregressive Integrated Moving Average algorithm (ARIMA). So I developed the model to apply ARIMA algorithm on these to datasets. The only given materials are two .csv files of value of bitcoin and gold during the five years, whose resource is London Bullion Market Association.

I was not allowed to use data from any other recourse, and I can only make my prediction based on values before the intended date rather than all the data provided. I calculated the first order difference, draw the autocorrelation and partial autocorrelation function of bitcoin and gold to analysis the stationary and other characters of dataset, then select the required parameters, order of AR model p, order of difference d and order of MA model q, by calculate the AIC and BIC values of the dataset. Applied the ARIMA algorithm on both datasets and finished the prediction. And compared the actually value and predict value in the last step. The programming language used in this problem is python and I am only member of the team.

Background

A market trader keeps buying and selling two volatile assets frequently, gold and bitcoin. In order to maximize the total return, they need to predict the value of the two assets of dates or months in the future, which is necessary to make decision on the kind and amount of asset he should buy or sell. The trader holds $1000 dollars and no gold and bitcoin initially, he needs to maximize the return during the five years.

Techniques and Tools

As python is the programming language used in this problem. Several libraries are used during the process. Including numpy, pandas and itertools to process the data, seaborn, matplotlib for displaying and visualization of data, statsmodels to finish the implementation of ARIMA model.

Autocorrelation Function (ACF): "The autocorrelation function (ACF) defines how data points in a time series are related, on average, to the preceding data points" (Box, Jenkins, & Reinsel, 1994).

Partial Autocorrelation Function (PACF): "The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags" (Paul, Andrew)

Akaiku Information Criterion and Bayesian Information Criterion (AIC & BIC): "A general way of comparing single-level models (models that do not include random effects or latent variables) is the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), also known as the Schwarz Criterion. The AIC and BIC balance the level of fit (quantified in terms of the log-likelihood) with model complexity (a penalty for using the sample data to estimate the model parameters)" (Anthony J.).

Autoregressive model (AR model): autoregressive model is a linear predictive modeling technique. It predicts sample based on previous samples by using the AR parameters as coefficients. The number of samples used for prediction determines the order of the model. The major advantage of AR model is it only use former value in the same dataset, which makes it suitable for information limited situation like this problem. Also, it is relatively simple as a linearly model.

"The AR model implementation has its shortcomings including limited robustness against nonstationarity, additive noise and aggregation (time-window averaging). Many of these have been recognized in econometric literature early on and proposed solutions consist in using preprocessing strategies such as detrending or differencing, and particularly in more sophisticated time series models, such as autoregressive moving average (ARMA) models" (Arthur W Toga).

Moving Average model (MA model): "The moving average (MA) model captures serial autocorrelation in a time series y_t by expressing the conditional mean of y_t as a function of past innovations, $\varepsilon t-1$, $\varepsilon t-2$, …, $\varepsilon t-q$. An MA model that depends on q past innovations is called an MA model of degree q, denoted by MA(q)."(Mathwork).

ARIMA model: it combines both AR and MA model to make prediction in a straightforward way and efficiently process the error.

First I noticed that dataset of gold still missing several values of dates. Gold market does not open on weekend and holiday. The solution is set the price of gold on the missing dates be equal to the former day. Which can be considered like gold market still opened the whole year, but in some days, there is not trade of gold so the price of gold stays still. Second, dates in column "date" is not in the same format, so we need to standardize it in order to sort the order of the values. Third, calculate the first order difference of both datasets. Forth, plot the figure of ACF and PACF of both datasets, calculation of ACF and PACF is already in the library statsmodels. In ACF figure, when the correlation located in the shadow area, which is the 95% confidence interval, that corresponding lag value is the order of moving average model q, by the same way I can also get the order of autoregressive model p. Fifth, check the selected p and q by compute the AIC and BIC value of the model using selected p and q, after computing AIC and BIC for all p and q value, a matrix will be generated, the smaller the AIC and BIC value is, the better the model is. If AIC and BIC are pointing to the same pair of parameters, we can use them in the ARIMA model and finish the prediction. If not, further selection will be needed. Sixth, using the selected parameters in ARIMA model and store the prediction value in a new .csv file. Seventh, draw the figure of both actual value and prediction value.

Result

For gold, its basic dataset information can be shown as followed.

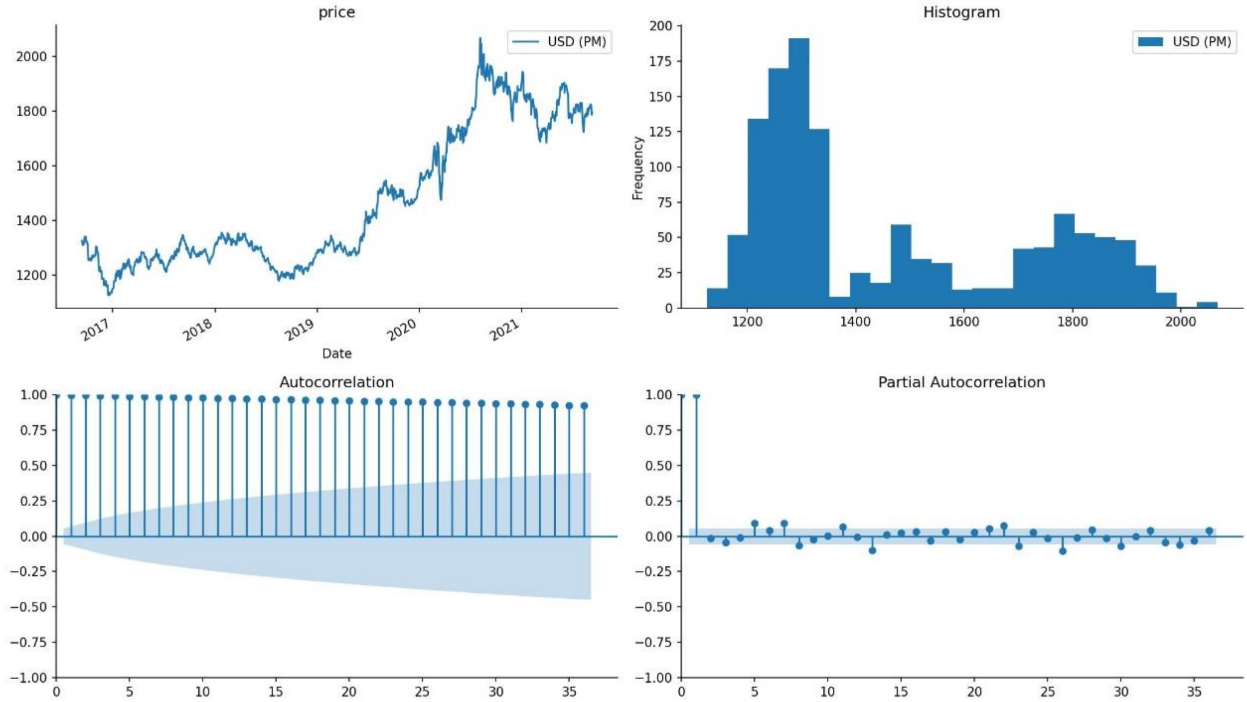


Figure1: Basic analysis of gold

ACF figure shows that the correlation between values of samples with in 20 units of time period lag is all high enough to say a strong relation. PACF indicates only the value of sample at former one lag of time has a strong enough relation. So we can conclude that the price of gold on a certain date is strongly dependent on price of gold of yesterday, and because of such a correlation, price of gold is highly related with its price up to twenty days ago.

By analyze the BIC value of each pair of p and q, it shows that p=1 and q=0 is the best choice of parameters for ARIMA model. On the other side, AIC value matrix indicate that p=1 q=4 is the best choice. Considered the ACF and PACF figure, ACF figure did not show strongly decreasing of autocorrelation after lag four. So we reject the outcome of AIC value and take p=1 and q=0 as the final decision.

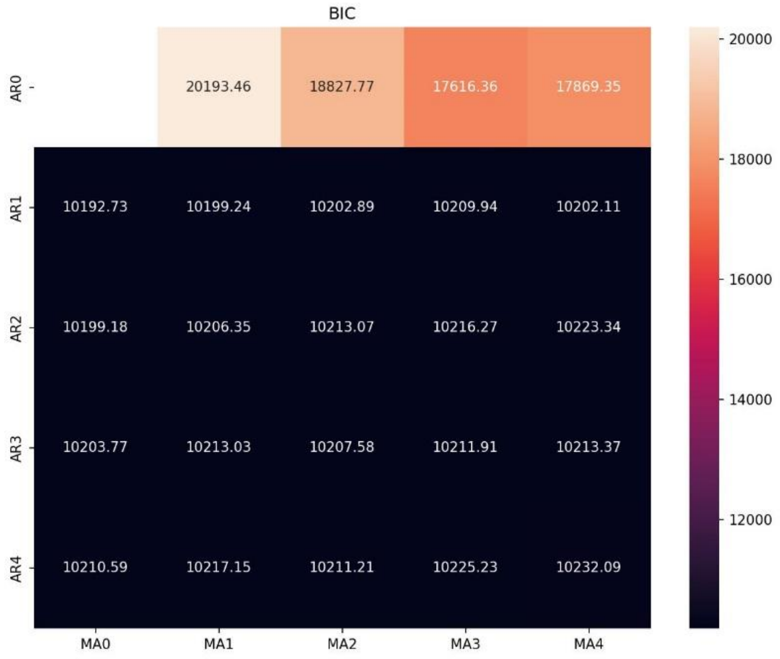


Figure2: Matrix of BIC value of gold

After applied ARIMA model, the prediction value and actual value can be shown as following:

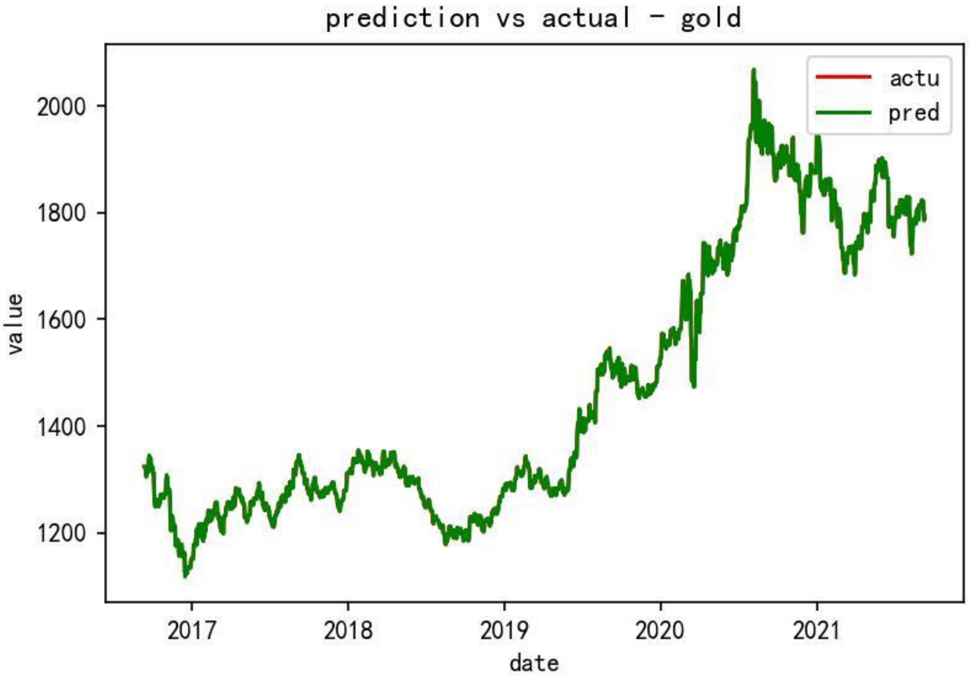
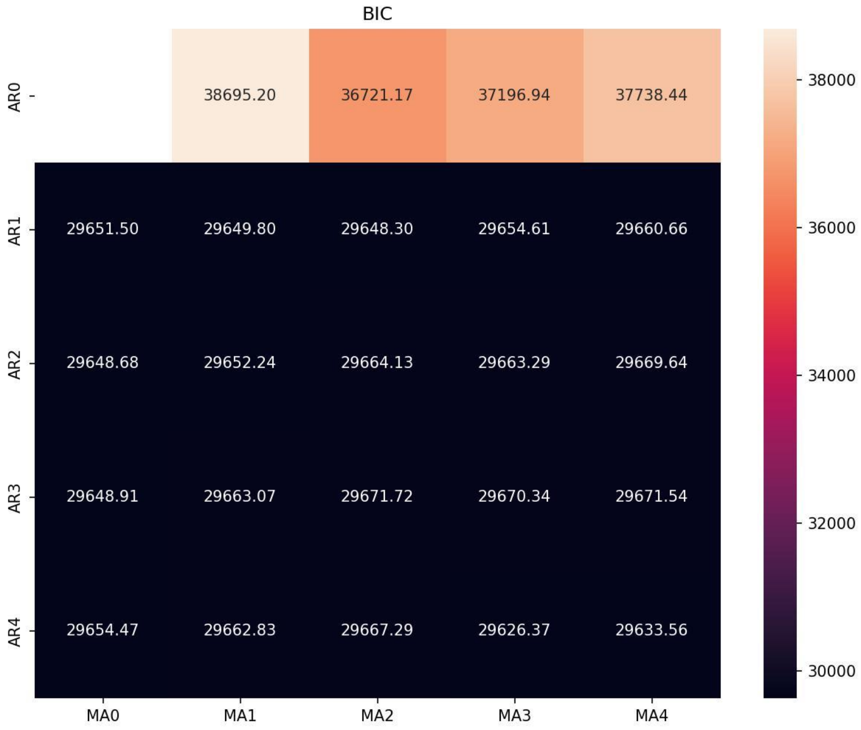


Figure3: Prediction and actual value of gold

We can say it is already a good enough prediction with less than 5% of error.
As for bitcoin, its BIC value matrix is:



BIC

	MA0	MA1	MA2	MA3	MA4
AR0		38695.20	36721.17	37196.94	37738.44
AR1	29651.50	29649.80	29648.30	29654.61	29660.66
AR2	29648.68	29652.24	29664.13	29663.29	29669.64
AR3	29648.91	29663.07	29671.72	29670.34	29671.54
AR4	29654.47	29662.83	29667.29	29626.37	29633.56

Figure4: Matrix of BIC values of bitcoin

BIC and AIC value gives the same answer p=4 and q=3 is the best choice. So after usage of ARIMA model, the prediction and actual value of bitcoin is shown as following:

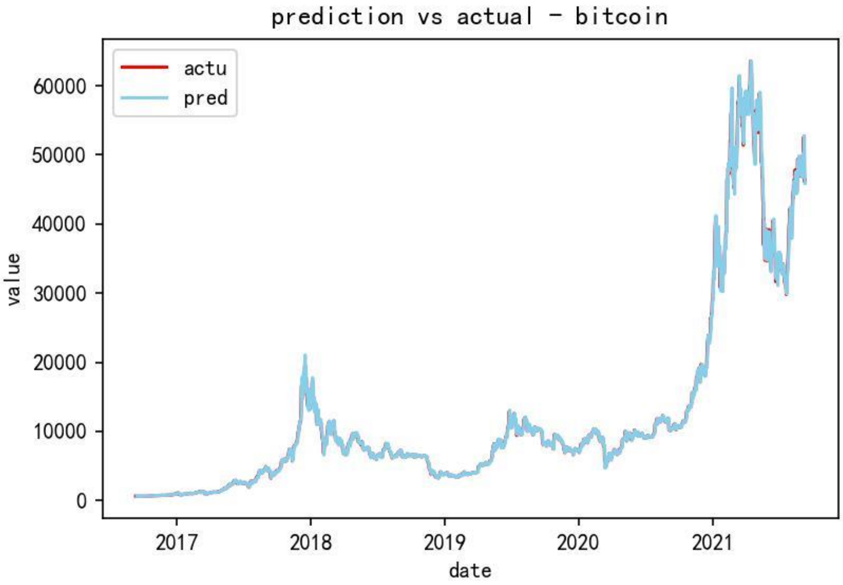


Figure5: Prediction and actual value of gold

But as the bitcoin’s value float significantly depends on time, so there are still about $1000 error in the prediction, it is not obvious on the figure because the absolute value of price of bitcoin is still pretty high. There is more space of improvement of prediction on bitcoin value.

Other Problems

ARIMA model requires its data be a stationary process, including strict stationary process and weak stationary process, strict stationary process means the mean and variance of the sample does not change over time, the white noise is a typical example of strict stationary process. More usual situation is weak stationary process, which means the correlation between terms does not change over time. Bitcoin can hardly be said as a stationary process and that makes the prediction hard to process. Function SARIMAX() in statsmodels has a parameter called “enforce_stationarity=True/False” by setting this parameter be false, the function can still process data which is not a stationary process and return a relatively matching prediction. It became the solution of this problem. But it shows that ARIMA model is clearly not the best choice for prediction on bitcoin. Other approaching still needs more research.

Conclusions

ARIMA model return well prediction based on only two provided .csv files. Gold is easier because of its stability on prices, while the value of bitcoin changes significantly over time. The trader should use gold as a way to store their investment money when value of bitcoin keeps decreasing, and bitcoin is the main resource of expanding the revenue.

Reference

"Arima." *Moving Average Model - MATLAB & Simulink*, https://www.mathworks.com/help/econ/moving-average-model.html.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Cowpertwait, Paul S. P, and Andrew V Metcalfe. *Introductory Time Series with R*. Springer-Verlag, 2009. INSERT-MISSING-DATABASE-NAME, INSERT-MISSING-URL. Accessed 5 July 2022.

Culyer, A. J, editor. *Encyclopedia of Health Economics*. First edition., First ed., Elsevier, 2014. INSERT-MISSING-DATABASE-NAME, INSERT-MISSING-URL. Accessed 5 July 2022.

Toga, Arthur W. *Brain Mapping* : An Encyclopedic Reference. Elsevier/Academic Press, 2015.