MAFS 5130 Project Report

Group14 : Jiang Xiaoyue, Shen Ruilin, Jia Yaoyao

1 Abstract

In recent years, China has increasingly focused on promoting new energy vehicles, and the implementation of many policies has also supported the development of the new energy vehicle industry. The export scale of new energy vehicles manufactured in China has been increasing year by year. As a leading enterprise in the production of new energy vehicles, BYD's stock price also reflects the impact of policy promotion exerted on the development of the enterprise. This article focuses on the stock prices, which is beneficial to understanding the development of the new energy vehicle industry and analyzing the impact of policies on the overall market. Therefore, this topic has a practical significance for us to investigate.

We obtain the stock data of BYD (002594. SZ) from 2012 to 2018 via Wind, calculate the return and percentage log-return of stock prices, and construct models to fit the time series data. In this article, we use multiple models to investigate the dataset, such as ARIMA model, GARCH model, ARIMA-GARCH model and LSTM model.

2 Data Processing

We plot the line chart of the dataset which contains the open prices of BYD stock from 2012 to 2018 as the figure shows below.

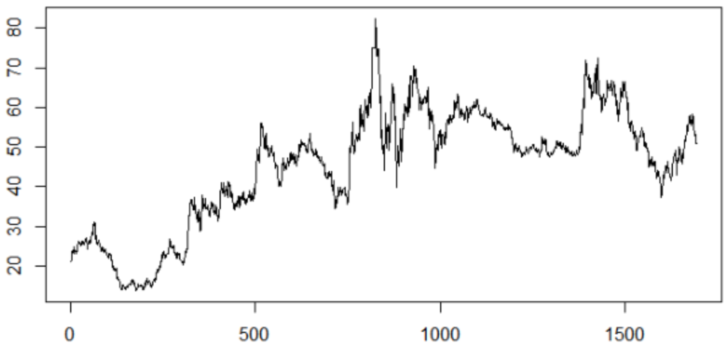


Figure 1: The stock price of BYD

Besides, we use the following formulas to calculate the percentage log-returns, during which the number one hundred is to play a role in amplifying numerical values. It can be found that the fluctuation of returns is great from the figures.

$$R_t = 100 * (logP_t - logP_{t-1})$$

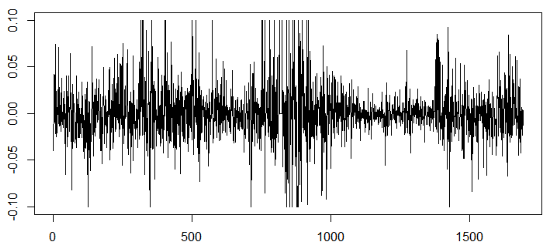


Figure 2: The plot of returns

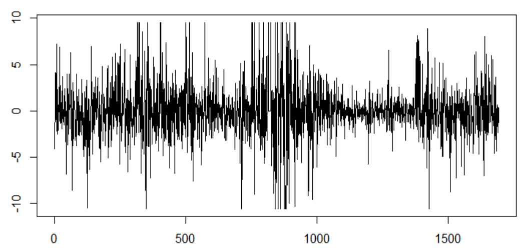


Figure 3: The plot of percentage log-returns

In order to examine whether the dataset is stationary or not, we do the ADF test on the log-return series. From the result, we can not reject the null hypothesis that the time series is stationary.

```
> adf.test(lr)

        Augmented Dickey-Fuller Test

data:  lr
Dickey-Fuller = -12.026, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary
```

Figure 4: The result of ADF test

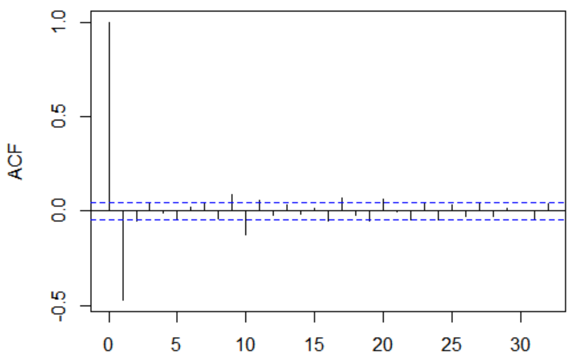


Figure 5: ACF of percentage log-returns

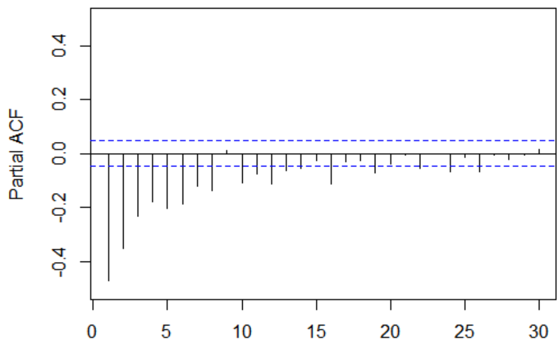


Figure 6: PACF of percentage log-returns

We use the percentage log-returns of stock to find the serial correlation by Ljung-Box test. Since the p-value of the test is less than the significance level 0.01, we can think that there exists the series correlation in the dataset.

```
> Box.test(lr,lag=12,type="Ljung")

        Box-Ljung test

data:  lr
X-squared = 31.566, df = 12, p-value = 0.001614
```

Figure 7: The result of Ljung-Box test

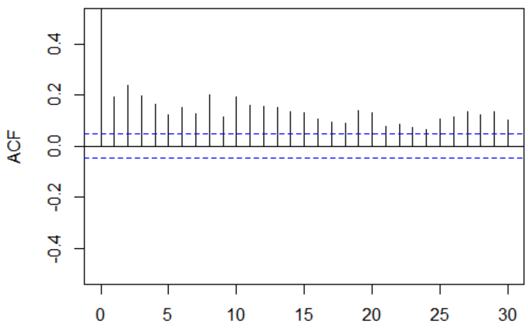


Figure 8: ACF of log-return square

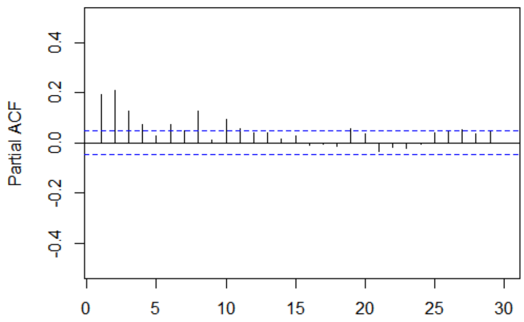


Figure 9: PACF of log-return square

In addition, to find whether there is ARCH effect in the time series, we do the ArchTest on the square of log-returns. Since p-value is less than 0.01, we can reject the null hypothesis. Therefore, the series of log-returns have the ARCH effect and we may consider using GARCH models to fit the dataset.

```
> ArchTest(lr, lags=12)

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  lr
Chi-squared = 229.01, df = 12, p-value < 2.2e-16
```

Figure 10: The result of ARCH effect test

3 Model Fitting

Since we have seven years of daily frequency data, we choose data from 2012 to 2017 as our training set, data of 2018 as our testing set.

3.1 ARIMA model

The result is that ARIMA model does not fit our dataset. But we still put our attempt and analysis process here.

Firstly, we want to use auto arima to obtain the most suitable ARIMA model. However, since we have many negative data, we cannot take the result directly. We still need to try different parameter combinations manually.

3.1.1 AR(p) model

For the single AR model, at first, we found that AR(19) model with all coefficients significant has the lowest AIC, but it is still very high.

```
> ar19=arima(log_BYD,order=c(19,0,0)) # aic = 4924.28
> ar19

Call:
arima(x = log_BYD, order = c(19, 0, 0))

Coefficients:
          ar1      ar2      ar3      ar4     ar5      ar6     ar7      ar8      ar9
      -0.9011  -0.8874  -0.8346  -0.7850  -0.782  -0.7124  -0.591  -0.5472  -0.4393
s.e.   0.0262   0.0352   0.0421   0.0471   0.051   0.0546   0.057   0.0582   0.0587
         ar10     ar11     ar12     ar13     ar14     ar15     ar16     ar17     ar18
      -0.5108  -0.4525  -0.3991  -0.3274  -0.2787  -0.2477  -0.2371  -0.1138  -0.1020
s.e.   0.0583   0.0587   0.0581   0.0570   0.0545   0.0510   0.0472   0.0422   0.0353
         ar19  intercept
      -0.0806    -0.0001
s.e.   0.0263     0.0033

sigma^2 estimated as 1.687:  log likelihood = -2441.14,  aic = 4924.28
```

Figure 11: The result of AR(19) model

The Ljung-Box test also shows that the p-value is too small, which means that there is autocorrelation in the residual sequence of the model, and AR(19) model is not adequate.

```
> Box.test(ar19$residuals,lag=24,type='Ljung') #lag必须比参数数量大

	Box-Ljung test

data:  ar19$residuals
X-squared = 25.598, df = 24, p-value = 0.3739

> pv=1-pchisq(25.598,5) #24-19=5
> pv
[1] 0.0001067668
```

Figure 12: The result of Ljung-Box test of AR(19) model

Hence, we can conclude that AR model is not correct.

3.1.2 MA(q) model

The following shows the result of MA(1) model.

```
> ma1=arima(log_BYD,order=c(0,0,1)) #aic = 4875.73
> ma1

Call:
arima(x = log_BYD, order = c(0, 0, 1))

Coefficients:
          ma1  intercept
      -1.0000      0e+00
s.e.   0.0022      1e-04

sigma^2 estimated as 1.667:  log likelihood = -2434.86,  aic = 4875.73
```

Figure 13: The result of MA(1) model

The absolute value coefficient of ma1 is equal to 1, which is not correct. And when we take q larger than 1, the coefficients become insignificant. Hence, we can conclude that the MA model is not adequate.

3.1.3 ARMA(p,q) model

When we try different combinations of (p,q), we find that the AIC values are around 5000, which is still very high, and the p-values of Ljung-Box test are very small, which shows that ARMA model is not adequate. The following gives two examples of ARMA model which seem feasible before the Ljung-Box test.

```
> arma2_3=arima(log_BYD,order=c(2,0,3)) #aic = 4878.96
> arma2_3

Call:
arima(x = log_BYD, order = c(2, 0, 3))

Coefficients:
          ar1      ar2      ma1     ma2      ma3  intercept
      -0.5925  -0.9888  -0.4031  0.3856  -0.9826      0e+00
s.e.   0.0141   0.0131   0.0170  0.0170   0.0174      1e-04

sigma^2 estimated as 1.661:  log likelihood = -2432.48,  aic = 4878.96
> Box.test(arma2_3$residuals,lag=12,type='Ljung')

        Box-Ljung test

data:  arma2_3$residuals
X-squared = 36.923, df = 12, p-value = 0.00023

> pv=1-pchisq(36.923,7) #12-5=7
> pv
[1] 4.850844e-06
```

Figure 14: The result of ARMA(2,3) model

```
> arma3_4=arima(log_BYD,order=c(3,0,4),fixed=c(NA,NA,NA,NA,0,NA,NA,NA)) #aic = 4869.6
> arma3_4

Call:
arima(x = log_BYD, order = c(3, 0, 4), fixed = c(NA, NA, NA, NA, 0, NA, NA,
    NA))

Coefficients:
          ar1      ar2      ar3      ma1  ma2     ma3      ma4  intercept
      -0.1743  -0.2085  -0.8230  -0.7825    0  0.6546  -0.8722      0e+00
s.e.   0.0691   0.0598   0.0653   0.0588    0  0.0726   0.0569      1e-04

sigma^2 estimated as 1.648:  log likelihood = -2426.8,  aic = 4869.6
> Box.test(arma3_4$residuals,lag=12,type='Ljung')

        Box-Ljung test

data:  arma3_4$residuals
X-squared = 22.449, df = 12, p-value = 0.03278

> pv=1-pchisq(22.449,5) #12-7=5
> pv
[1] 0.0004300652
```

Figure 15: The result of ARMA(3,4) model

Our conclusion is that the ARMA model is not fit.

3.1.4 ARIMA(p,d,q) model

In our test, we found that only when q=0, the model might be fit. Hence we fixed q=0, and try different combinations of (p,d). Here follows two examples of ARIMA model before the Ljung-Box test.

```
> arima1=arima(log_BYD,order=c(2,1,0)) #aic = 6197.88
> arima1

Call:
arima(x = log_BYD, order = c(2, 1, 0))

Coefficients:
          ar1      ar2
      -0.9583  -0.4973
s.e.   0.0228   0.0228

sigma^2 estimated as 4.173:  log likelihood = -3095.94,  aic = 6197.88
```

Figure 15: The result of ARIMA(2,1,0) model

```
> arima2=arima(log_BYD,order=c(1,2,0)) #aic = 7984.35
> arima2

Call:
arima(x = log_BYD, order = c(1, 2, 0))

Coefficients:
          ar1
      -0.7307
s.e.   0.0179

sigma^2 estimated as 14.37:  log likelihood = -3990.18,  aic = 7984.35
```

Figure 15: The result of ARIMA(1,2,0) model

As we can see, if d becomes larger, AIC will be large. Even when d=1, AIC is larger than 5000, which means the ARIMA model is even worse than AR model. Hence, we can conclude that ARIMA model is not adequate for our dataset.

3.2 GARCH(m,s) model

The main idea of the GARCH model is to replace the conditional variance of *t* return with the *t-1* measurable function of the information set. And we have the representation formulas of GARCH models as follows.

$$a_t = \sigma_t \varepsilon_t$$

$$\sigma_t{}^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}{}^2 + \sum_{j=1}^{s} \beta_i \sigma_{t-j}{}^2$$

According to the following plots of EACF, we try to establish the GARCH models and compute AIC values of models with different variables. We also use the t-test result to show that the mean value is significantly equal to 0.

```
> eacf(abs(lr))
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x o o o o o o x x x o  o  o  o
2 x o o o x o o o o x o  o  o  o
3 x x o o o o o o o x o  o  o  o
4 x x x x o o o o o x o  o  o  o
5 x x o x x o o o o x o  o  o  o
6 x x o x x o o o o x o  o  o  o
7 x x x x x x x o o x o  o  o  o
```

```
> eacf(lr^2)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x o o o o o x x x o  o  o  o
2 x o o o x o o o o x o  o  o  o
3 x x o o o o o o o x o  o  o  o
4 x x x x o o o o o o o  o  o  o
5 x x x x o o o o o x o  o  o  o
6 x x x x x x o o o o o  o  o  o
7 x x x x x x x x x x o  o  o  o
```

Figure16: The result of EACF

```
> t.test(lr)

        One Sample t-test

data:  lr
t = 0.69937, df = 1693, p-value = 0.4844
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.08914527  0.18794980
sample estimates:
 mean of x
0.04940226
```

Figure17: The result of t-test

From the following table, the GARCH(1,1) model has the lowest AIC value so we use GARCH(1,1) to do the research.

GARCH(m,s)		
m	s	AIC
1	1	4.733213
1	2	4.733342
2	1	4.734434
1	3	4.734462
3	1	4.735067
2	2	4.734484
2	3	4.735443
3	2	4.734686
3	3	4.733855

Table 1: AIC of different GARCH models

```
> Res=residuals(m3,standardize=T)
> Box.test(Res,10,type="Ljung")

        Box-Ljung test

data:  Res
X-squared = 13.558, df = 10, p-value = 0.1941
```

```
> Box.test(Res,20,type="Ljung")

        Box-Ljung test

data:  Res
X-squared = 18.6, df = 20, p-value = 0.5479
```

```
> Box.test(Res^2,10,type="Ljung")

        Box-Ljung test

data:  Res^2
X-squared = 7.5911, df = 10, p-value = 0.6687
```

```
> Box.test(Res^2,20,type="Ljung")

        Box-Ljung test

data:  Res^2
X-squared = 12.518, df = 20, p-value = 0.8971
```

Figure18: Residuals analysis of GARCH(1,1)

We do the Ljung-Box test on the residuals and the residuals analysis shows that there is no serial correlation or ARCH effect and the GARCH(1,1) model is adequate. Besides, the parameters of GARCH(1,1) model are obtained below.

$$a_t = \sigma_t \varepsilon_t, {\sigma_t}^2 = 0.0558 + 0.07474 {a_{t-1}}^2 + 0.922 {\sigma_{t-1}}^2$$

```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
omega    0.05580     0.01835    3.041  0.00236 **
alpha1   0.07474     0.01220    6.125 9.09e-10 ***
beta1    0.92200     0.01166   79.106  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log Likelihood:
 -4006.031    normalized:  -2.364836
```

Figure19: Parameters of GARCH(1,1) model

3.3 ARIMA(p,d,q)-GARCH(m,s) model

The formulary definition of ARMA(p,q)-GARCH(m,s) is shown below:

$$r_t = \sum_{i=1}^{p} \phi_i r_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j}$$

$$a_t = \sigma_t \eta_t$$

$${\sigma_t}^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i {a_{t-i}}^2 + \sum_{i=1}^{s} \beta_i {\sigma_{t-j}}^2$$

We continue to use the above method to calculate the AIC values for different ARIMA-GARCH models. The result shows that ARMA(2,2)-GARCH(1,1) model has the smallest AIC. Therefore, we choose the ARMA(2,2)-GARCH(1,1) model to fit the dataset.

ARIMA(p,d,q)-GARCH(m,s)					
p	d	q	m	s	AIC
1	0	1	1	1	4.734309
2	0	2	1	1	4.733104
0	1	0	1	1	5.383260
1	1	0	1	1	5.118440
0	1	1	1	1	4.746703
1	1	1	1	1	4.747611

Table 2: AIC of different ARIMA-GARCH models

We do the Ljung-Box test on the residuals and the residuals analysis shows that there is no serial correlation or ARCH effect and the ARMA(2,2)-GARCH(1,1) model is adequate. Besides, the parameters of ARMA(2,2)-GARCH(1,1) model are obtained below.

```
Ljung-Box Test     R     Q(10)  11.29865  0.334729
Ljung-Box Test     R     Q(15)  12.12803  0.6693131
Ljung-Box Test     R     Q(20)  16.46204  0.6875785
Ljung-Box Test     R^2   Q(10)  7.586023  0.6692016
Ljung-Box Test     R^2   Q(15)  8.62459   0.8963146
Ljung-Box Test     R^2   Q(20)  12.62053  0.8930656
```

Figure20: Residuals analysis of ARMA(2,2)-GARCH(1,1)

$$r_t = 0.77389 r_{t-1} - 0.81021 r_{t-2} + a_t - 0.76525 a_{t-1} + 0.82148 a_{t-2}$$
$$a_t = \sigma_t \eta_t, \quad {\sigma_t}^2 = 0.05701 + 0.07518 {a_{t-1}}^2 + 0.92146 {\sigma_{t-1}}^2$$

```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)    
ar1      0.77389     0.06048   12.795  < 2e-16 ***
ar2     -0.81021     0.11656   -6.951 3.63e-12 ***
ma1     -0.76525     0.05918  -12.931  < 2e-16 ***
ma2      0.82148     0.11492    7.148 8.79e-13 ***
omega    0.05701     0.01881    3.030  0.00244 ** 
alpha1   0.07518     0.01240    6.065 1.32e-09 ***
beta1    0.92146     0.01190   77.444  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure21: Parameters of ARMA(2,2)-GARCH(1,1) model

3.4 LSTM model

This is the basic information of our LSTM model:

Model Structure:

```
LSTMModel(
  (linear_1): Linear(in_features=1, out_features=32, bias=True)
  (relu): ReLU()
  (lstm): LSTM(32, 32, num_layers=2, batch_first=True)
  (dropout): Dropout(p=0.2, inplace=False)
  (linear_2): Linear(in_features=64, out_features=1, bias=True)
)
```

Model Setting:

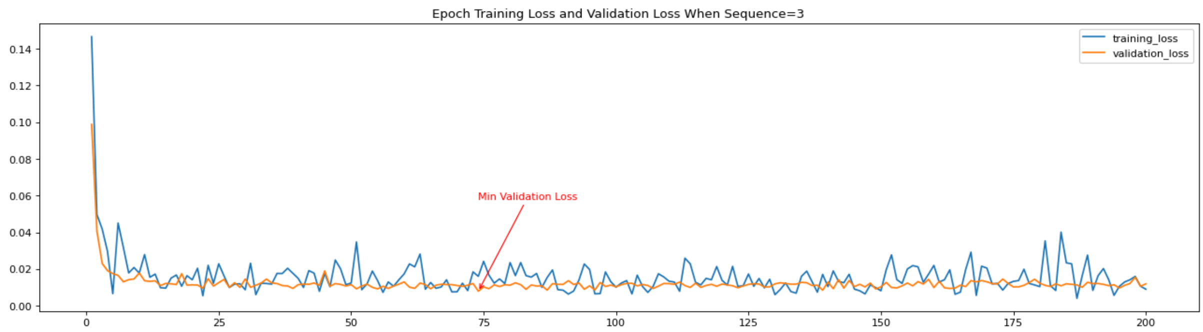
Hyper parameter	Setting
Criterion	**MSELoss**

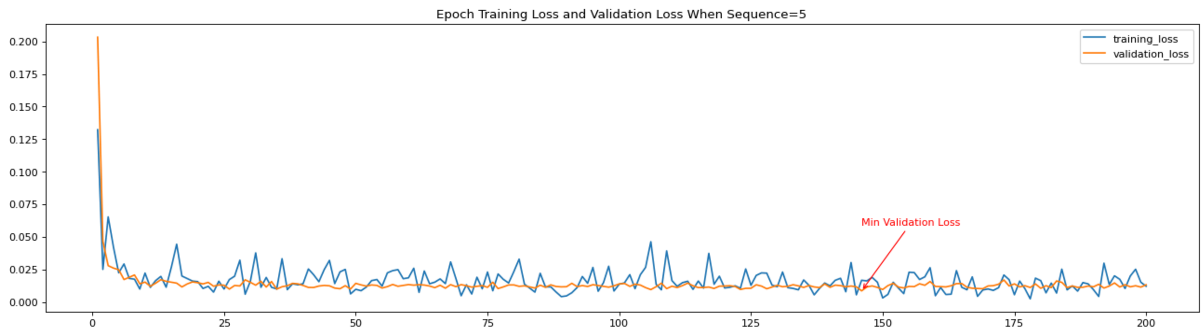
Optimizer	**Adam**
Dropout	**0.2**
Learning Rate	**0.001**
Epoch	**200**
Sequence (window_size)	**3,5,10,15,20**

Table 3: LSTM model setting

Firstly use 2012-2016 data as train set, 2017 data as validation set, then use 2013-2017 data as train set, 2018 data as test set.

Here are some figures about training loss and validation loss during model training:





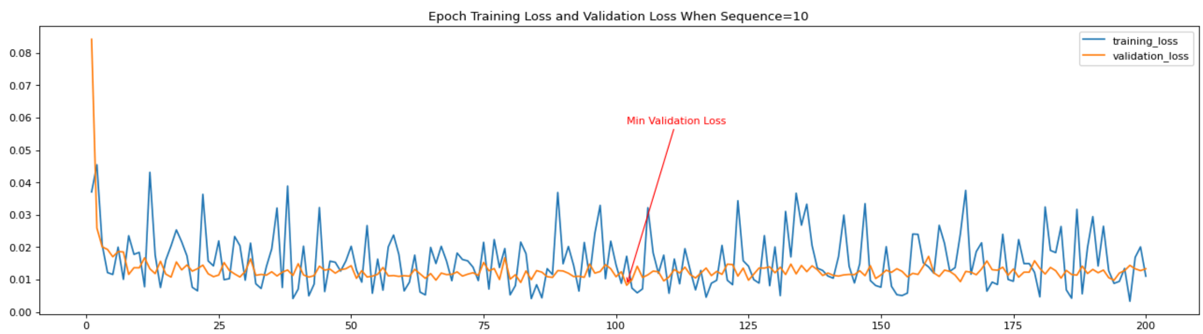


Figure22:Training Loss and Validation Loss Figures

And following is the Minimum Validation Loss with different parameter “sequence”.

Sequence	**Minimum Validation Loss**
3	**0.007991445**
5	**0.008481636**
10	**0.008229232**
15	**0.009599054**
20	**0.009699056**

Table 4: Minimum Validation Loss with Different Sequence

4 Forecasting

4.1 GARCH(1,1) model forecasting

The follows shows the prediction of GARCH(1,1) model with confidence intervals.

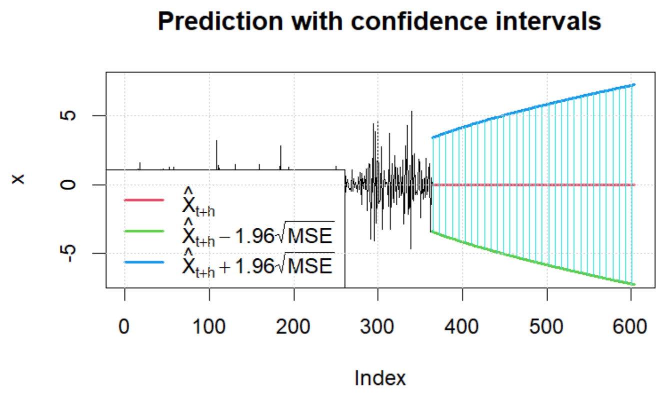


Figure23: Prediction of percentage log return of GARCH(1,1) model

The above prediction is about percentage log return. If we should predict the open price, we need to pay attention that the GARCH model has noise term. If we choose to simulate the noise term, the results of each simulation will be very different. So here just give one example of the prediction of the price. And for the ARMA(2,2)-GARCH(1,1) model, we will not simulate the noise term to get the price, because we think that it is not so meaningful.

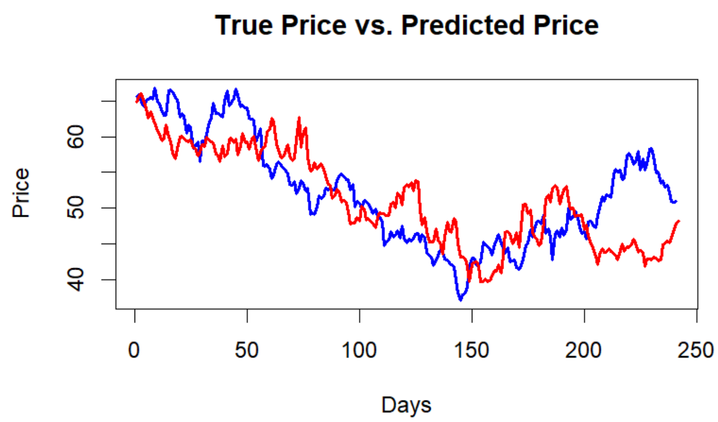


Figure24: Prediction of open price of GARCH(1,1) model

The blue line is the true price, and the red line is the predicted price.

4.2 ARMA(2,2)-GARCH(1,1) model forecasting

The follows shows the prediction of ARMA(2,2)-GARCH(1,1) model with confidence intervals.

Compared to the GARCH(1,1) model, we can see that the confidence intervals of ARMA(2,2) - GARCH(1,1) model is narrower in the long term, hence ARMA(2,2) - GARCH(1,1) model may better in prediction.

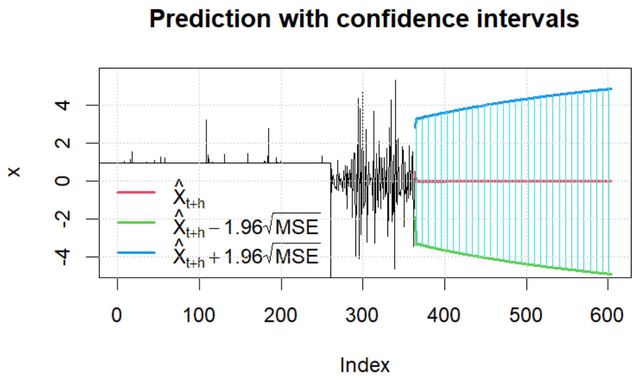
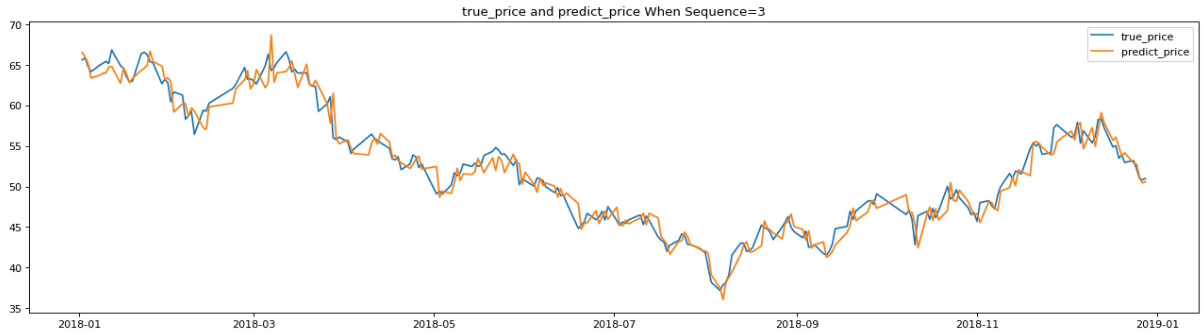
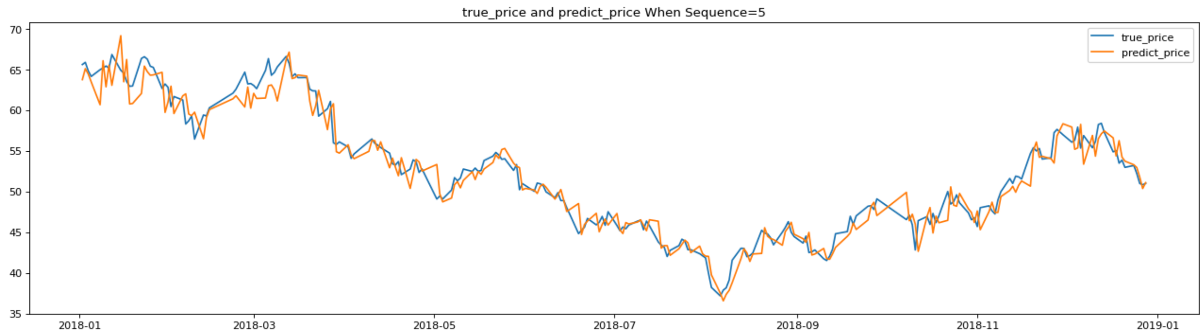


Figure25: Prediction of percentage log return of ARMA(2,2)-GARCH(1,1) model

4.3 LSTM model forecasting





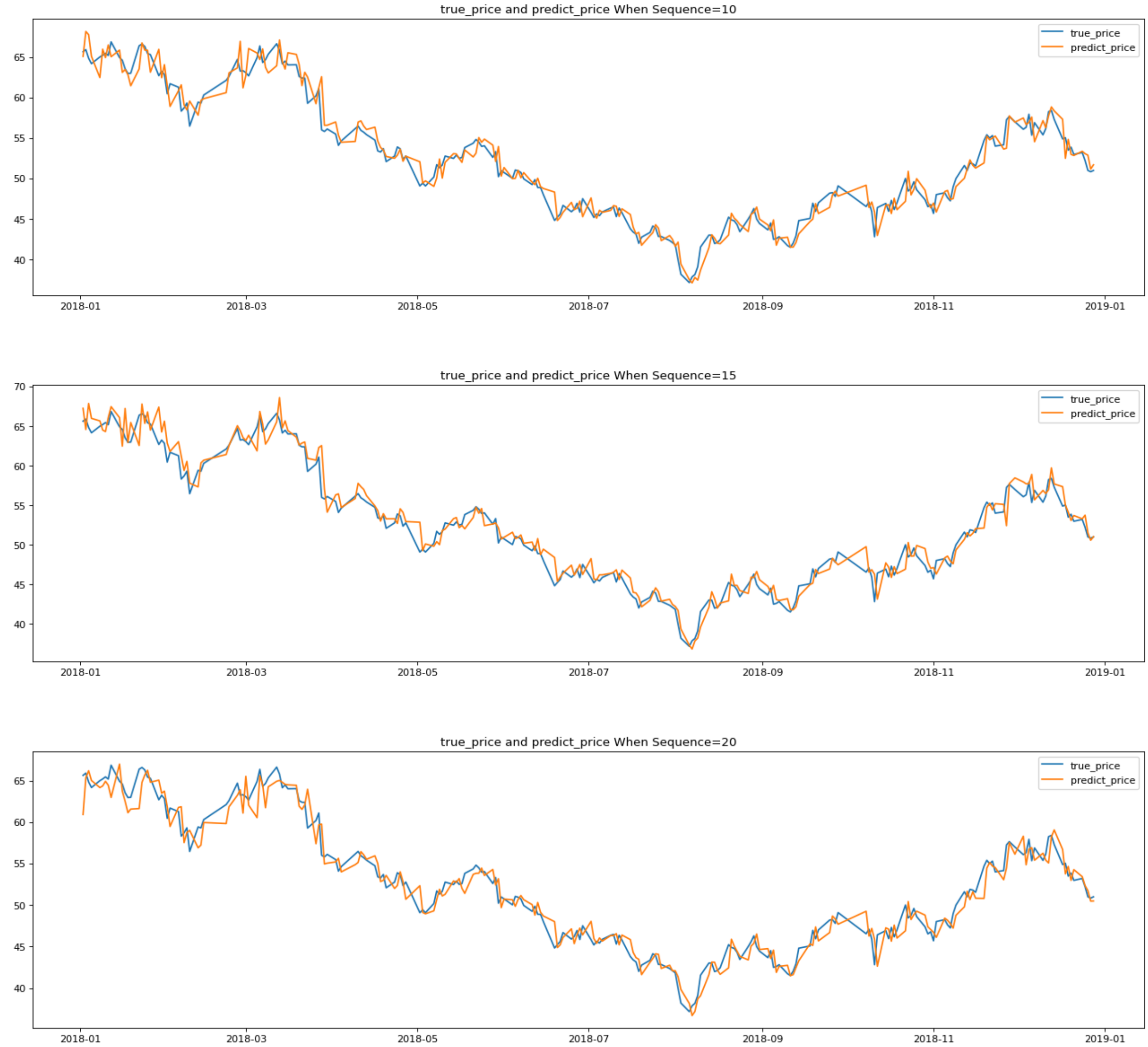


Figure26: Real Price and Prediction Comparison using LSTM

Following is comparisons of models with different sequence:

Sequence	**MSE**	**RMSE**	**MAE**	**R Square**
3	**1.983**	**1.408**	**1.082**	**0.9652**
5	**2.745**	**1.657**	**1.245**	**0.9519**
10	**2.057**	**1.434**	**1.087**	**0.9639**
15	**2.096**	**1.448**	**1.079**	**0.9633**
20	**2.177**	**1.476**	**1.141**	**0.9618**

Table 5: Comparison between LSTM models with different sequence

The result approves that the model with least validation loss may have better performance.

5 Appendix

All code and data can be found at: https://github.com/jiayaoyao666/MAFS5130-Project