Analysis of the Medical Industry in the Stock Market under the Influence of the Pandemic in China

Ziyue Mao[1,a,*,†], Ruixiang Wu[2,b,†], Zhao Dong[3,c,&], Xian Zhang[4,d,&]

[1]Hamden Hall Country Day School, Hamden, CT, 06517, United States
[2]International School, Beijing University of Post and Telecommunications, Beijing, 102206, China
[3]Applied Physics, Huaiyin Institute of Technology, Huaian, 223003, China
[4]Department of Mathematics and Computer Science, Guangdong Technion- Israel Institute of Technology, Shantou, 515063, China
a. ziyuemaoleo@gmail.com, b. 2022213758@bupt.cn, c. dongzhao0921@outlook.com, d. zhang09111@gtiit.edu.cn
**corresponding author*
[†]Ziyue Mao and Ruixiang Wu contributed equally to this work and should be considered co-first authors.
[&]Zhao Dong and Xian Zhang contributed equally to this work and should be considered co-second authors.

Abstract: This paper is working to analyze the influence of COVID-19 on the medical industry in China's stock market. Because the pandemic has influenced our society and development impressively, using statistical methods to analyze its influence is a necessary method to estimate the risk of investment. Some common statistic analyzing methods, the ARIMA model and GARCH model, will be used in this paper, and the data of the stock price of a medical manufacturing company Zhangzhou Pientzhng Phrmctcl Co Ltd during the COVID-19 is collected and introduced in the models. Meanwhile, machine learning is also implemented in the prediction and analysis of the data. The results of these two models both demonstrate the high volatility in the stock market under the influence of pandemic, which is meaningful for further research about the risks of future investment in medical industry of China's stock market.

Keywords: ARIMA model, GARCH model, Medical industry, Machine learning

1. Introduction

Since its emergence, the medical industry has been the backbone of a region and even a country, especially in the 21st century of China. The medical industry market has no doubt been promising and available for China in the past 20 years, however, because of the emergency of COVID 19, there emerged a lot of medical companies, which created a sudden flourishing in the stock market of the medical industry [1]. Moreover, the increase in people's medical care is linked to their participation in the stock market, and the reason for this is mainly because of the progressing policies that ensure families decrease their spending on medical insurance, and as a result, allow people to be more confident on participation in invest in stock market [2].

As the medical industry became more and more significant for people and the stock market in China, the pandemic COVID-19 offered more opportunities to this market. Allocated by the government, people received about 10 billion CNY (about 1.5 billion dollars) in public healthcare service, and many medical companies were encouraged to invent new medicines, such as vaccines, with the assistance of new policies [3]. With more and more people focusing on the medical industry in the stock market, some companies have undergone a huge increase in their market capitalization, accompanied by their high liquidity [4]. However, these features reflected the competitive environment of China's stock market in the medical industry, which led to the volatile stock prices [4]. Therefore, it is necessary to anticipate some investment risks, by using numerical models (based on historical data) to predict the stocks' prices to help understand the market's instability.

The ARIMA (Autoregressive Integrated Moving Average) model, known in the forecasting numerical models and has been introduced in many stock market simulation cases, has also been analyzed and tested a lot for its accuracy [5] It displayed effective results in different types of time-series data, including linear trend, seasonal, and periodic data. Moreover, the ARIMA model can provide forecasts of future values and can explain the confidence intervals of the forecasts [5] However, a single ARIMA model cannot meet the higher forecasting accuracy requirements, because it can only handle smaller forecasting periods, and the contribution of forecast errors to the results due to different decompositions of the time series has not yet been analyzed [6,7].

The GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model, introduced by Tim Bollerslev in 1986, can also provide forecasts of future volatility, explain the confidence intervals of the forecasts, and accommodate different types of volatility patterns; additionally, it includes autocorrelation and nonlinear relationships in volatility [8]. This is particularly useful in finance where volatility clustering is often observed, i.e., high volatility days tend to be followed by high volatility days and low volatility days tend to be followed by low volatility days. Nevertheless, there are also some drawbacks of the GARCH model: it can only capture changes in variance and cannot handle more complex patterns of volatility, such as long-term dependence and nonlinear relationships [9].

Thus, this paper is working to analyze the medical industry in the stock market by both the ARIMA model and the GARCH model, by comparing the accuracy of both models under the influence of COVID-19 to help forecast the stock market for the medical industry.

2. Data Selection

By using the ARIMA model and the ARCH model to analyze the medical industry in the stock market, this paper chooses a stock in the RMB Ordinary Shares(A-Shares), Zhangzhou Pientzhng Phrmctcl Co Ltd (Abbreviated as ZPPCL). This company owns one of the largest market values in the medical industry aspect of A-Shares, and it also maintains great liquidity. The simulation could reflect certain regularity features of the company, so this can help to understand the trend of the medical industry in the stock market. The two models introduced the close stock price and 149 data points, which is trade data, of these two stocks from Sep. 1st to Jul. 31st in 2023, and then simulated the stock value for each day in the next month. After that, it would be able to compare and analyze the prediction value and the actual stock price of these two companies in the stock market in August.

3. ARIMA Model

3.1. Definition of the ARIMA model

As introduced in the last part, the ARIMA model is one of the effective time-series forecasting models. To understand the model, we can split the model into three parts: Autoregression (AR) is a model that shows a changing variable that is regressed on its own lagged or prior value; Integration (I): indicates the difference between the original observations to smooth the time series; and Moving Average (MA): Combines the dependence between observations and the residuals of a moving average model applied to lagged observations [10]. Besides, for the ARIMA model, p, d, and q, are three major parameters that

- p: the number of lagged observations in the model; also known as the lag order [10].
- d: the number of times the original observations differ; also known as the variance [10].
- q: the size of the moving average window; also known as the moving average order [10].
- When combining each part from the ARIMA model, it comes out the Equation (1):

$$\nabla^d y_t = c + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{1}$$

In Equation (1), ∇^d represents the differencing steps applied d times to the data. The coefficients from $\emptyset_1$ to $\emptyset_p$ are the coefficients of the AR model, and the coefficients from θ_1 to θ_q are the coefficients of the MA model. c is a constant, and ϵ_t is the error term at time t.

3.2. Simulation Method

3.2.1. Augmented Dickey-Fuller Test

The ARIMA model is a statistical model used to predict future values based on the autocorrelation present in historical data. It assumes that the future trend will follow the same pattern as the historical trend and requires that the time series is stationary [11]. Non-stationarity can lead to forecast errors and unstable parameter estimates, thus reducing the reliability of the forecast results. Therefore, it is important to determine whether the time series is stable or not, where the ADF test matters in the experiment that helps to determine whether the historical data is stable or not. There are two hypotheses for the ADF Test: One is the Null Hypothesis, in which the time series has a unit root, and thus it is non-stationary. The contrary one is the Alternative Hypothesis that the time series does not have a unit root, implying it is stationary.

The test statistic in the ADF test compares the estimated coefficient of the lagged series in a regression model to its standard error. If the coefficient is significantly different from zero, the time series is likely stationary. While using the ADF test, it can check the data after using first-order differencing to decide whether is necessary to do second-order or higher-order differencing.

3.2.2. Akaike Information Criterion Method

The AIC is a measure used to compare and select models. In the context of time series forecasting and many other statistical modeling problems, the AIC provides a trade-off between the goodness-of-fit of the model and the complexity of the model. The AIC penalizes the addition of extra parameters to prevent overfitting. Lower AIC values indicate a better-fitting model, but the real utility of AIC is in comparing different models. When deciding between multiple models, the one with the lowest AIC value is generally the best.

$$AIC = 2k - 2ln(L) \tag{2}$$

In Equation (2), k is the number of parameters in the model and L is the likelihood of the model. Selecting the best parameters for an ARIMA model is a common practice, the model loops over potential values of p, d, and q, which is essentially a grid search. For each combination, calculate the AIC. Finally, the output will provide you with the combination of p, d, and q that produced the lowest AIC, which should be the best combination for modeling the series.

3.3. Analysis of the ARIMA model’s simulation

3.3.1. Augmented Dickey-Fuller Test Analysis

After the simulation of the ADF Test, there are some significant standards of statistical value in the method that can determine whether the stock data of the company is stationary. P-value, which checks the possibility of the observed data under the condition of the null hypothesis. In the simulation, the p-value equals to 0.1175, which is larger than 0.05. When the p-value is larger than 0.05, the experiment could reject the null hypothesis, which means the data is stationary. On the opposite, if it is less than 0.05, the data is likely to be non-stationary. In this case, the p-value demonstrates that the stock data of the company is not stationary.

Critical value, for the boundary values at different significance levels, the ADF test compares the test’s statistic value to these critical values to determine whether the null hypothesis could be rejected or not. The test standard that the data performed and being calculated by the models is -2.491123.

Table 1: Comparison of critical value and test standard for the ARIMA model

Type of Critical Value	Values	Compare with the Test Standard
Critical value 1%	-3.457664	Less than -2.491123
Critical value 5%	-2.873559	Less than -2.491123
Critical value 10%	-2.573175	Less than -2.491123

For each type of critical value, they are all less than the test standard, so this result can not reject the null hypothesis either, and so the data is non-stationary in this case.

After these judgments of the stability of the data, the non-stationary state requires the data to be differenced to achieve a stationary state. The original data and the differenced data are measured in the figure 1 below:

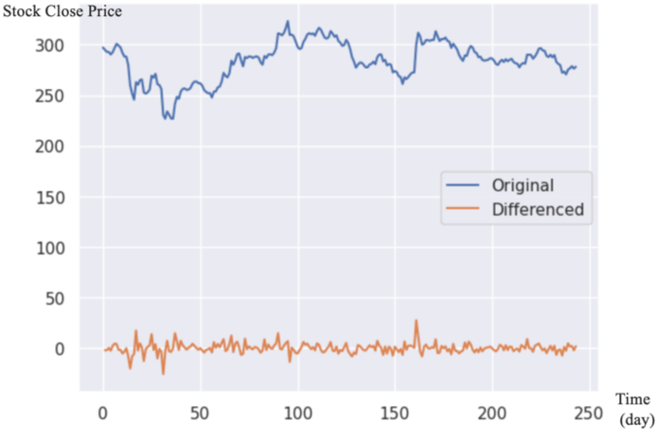

Figure 1: Original and differenced data for the stock price

In the figure, it’s obvious that the amplitude of the orange line is much less than the blue line, which means the differenced data indeed transfer the data into the stationary state. Also, the original data does not show any features of seasonality or periodicity. In fact, the amplitude of data is not stationary and not uniformed distributed either.

3.3.2. Akaike Information Criterion Method Analysis

To prevent the model from being excessively complex could causing overfitting problems, the experiment sets up $p \in [0,5]$ and $q \in [0,5]$. After running the loop of each combination of p and q with the first differenced data, the best combination that outputs the lowest AIC is $(p, d, q) = (2,1,4)$.

Autocorrelation and partial autocorrelation are fundamental concepts in time series analysis, especially when building time series forecasting models like ARIMA. Autocorrelation, also known as serial correlation or lagged correlation, measures the relationship between a variable's current value and its past values. ACF is a way to measure the linear relationship between an observation at time t and the observations at previous times (lags). It provides correlations for every lag (e.g., lag 1, lag 2, etc.). The autocorrelation function plot (often called the correlogram) displays the autocorrelation values for different lags. While autocorrelation measures the correlation between a variable and its lags, partial autocorrelation measures the correlation between a variable and its lag, controlling for the correlation with all shorter lags. The figure 2 below shows the ACF and PACF of the data:

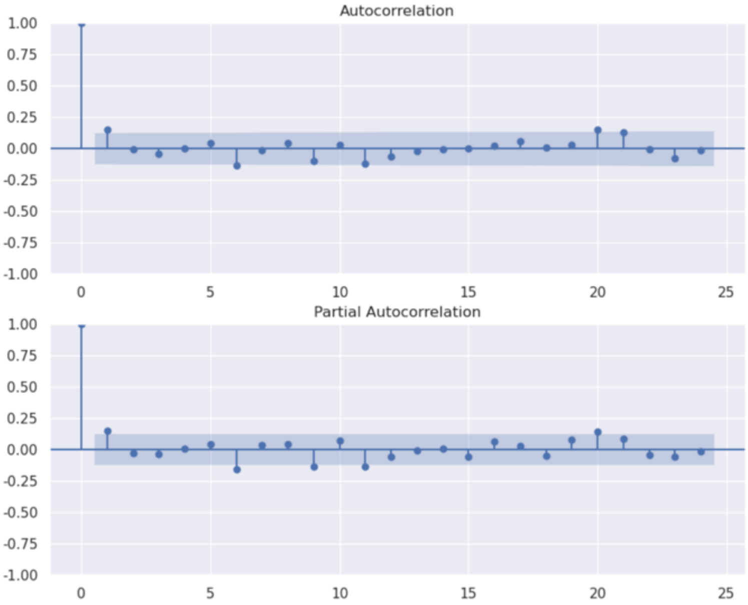

Figure 2: ACF and PACF of the ARIMA model

In the figure above, ϕ is the value of the y-axis, and k is the x-axis, which represents the lag order. Since $|\phi| < 1$ (smoothness), the absolute value of the autocorrelation coefficient decreases exponentially as the lag length k grows. When $|\phi| > 0$, the autocorrelation coefficients are all greater than 0. When $|\phi| < 0$, the autocorrelation coefficients are positively and negatively staggered. When $|\phi|$ is near 1, the exponential decreases slowly, and when $|\phi|$ is far away from 1, the exponential decreases rapidly. Therefore, when $|\phi|$ is near 1, the strong correlation will last for many periods. If $\phi > 0$, the series will be relatively smooth (and may even look like it is trending), and if $\phi < 0$, the series will be jagged.

3.3.3. Prediction with machine learning of ARIMA model

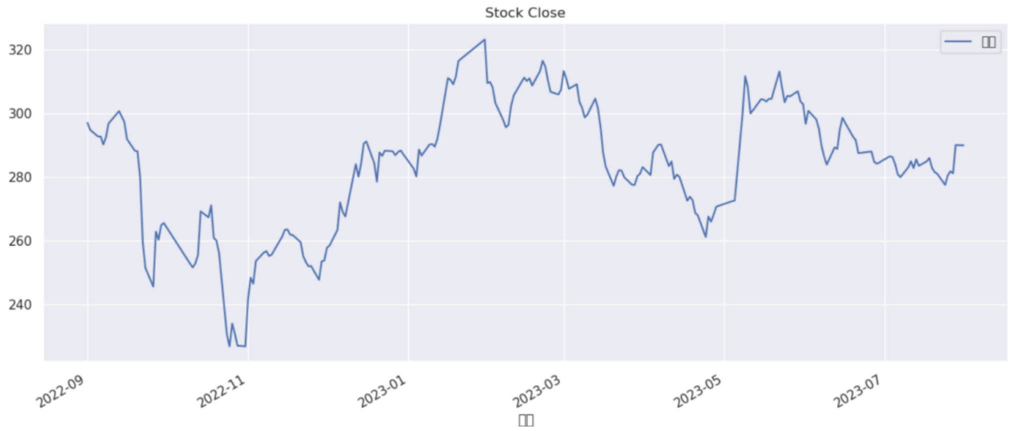

Figure 3: Stock Close Price from 09/2022 to 07/2023 of ZPPCL

Machine learning is a subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to perform a task without using explicit instructions [12]. Instead, they rely on patterns and inference. In other words, it's about creating algorithms that allow computers to learn from and make decisions based on data. The training set of this model introduced the data from 09/01/2022 to 07/31/2023 to predict the data for the next month and compare it with the actual data (from 08/01/2023 to 09/01/2023).

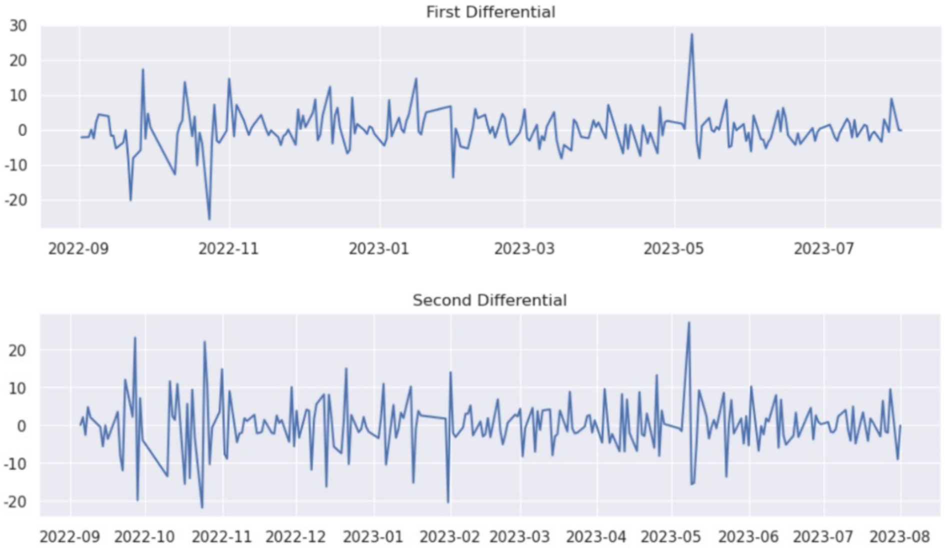

Figure 4: First- and Second-order difference for the original stock close price data

The figure 4 above is the first-order difference and second-order difference. The first-order difference is using each data point in the data to subtract the previous one to obtain a new data set. Second-order difference is performing a difference operation on a sequence of first-order differences, then, subtracting the previous data point from each data point in the sequence of first-order differences. By implementing these two operations, the data set can become more stationary and be able to come out with a better simulation result. Before using difference, the ADF method is always introduced to determine how many orders the data set needs to achieve a stationary state. Because the second-order differenced data is more uniform than the first-order differenced data, it is better to use the second-order data for simulation.

After conducting the AIC method and determining that the (p, d, q) equals $(2,1,4)$, the simulation produced the result and compared the prediction with the actual value.

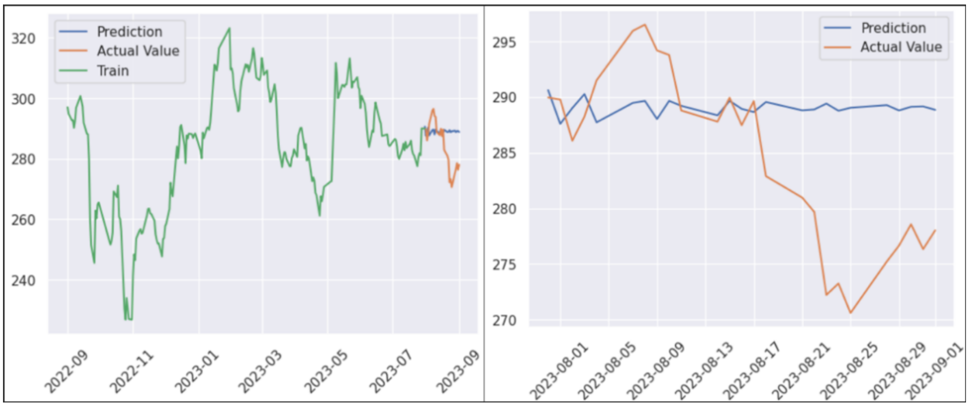

Figure 5: Prediction value and Actual value

The figure 5 shows that the prediction is not satisfactory, because it does not fit with the actual value well. However, the result has a certain reference value. The value of prediction reflects the average of the actual value to some degree.

4. GARCH Model

4.1. Definition of the GARCH model

In many financial time series, even though returns might be unpredictable (i.e., have a mean that doesn't change over time), the variance or volatility of returns can change. Periods of increased volatility tend to be followed by more periods of increased volatility, and similarly for decreased volatility. The GARCH models aim to capture this clustering of volatility. The mathematical form of a GARCH (p, q) model is:

$$r_t = \mu + \epsilon_t \tag{3}$$

In Equation (3), r_t is the return at time t, and μ is the mean of the return series. ϵ_t is the innovation or at time t, which is modeled as:

$$\epsilon_t = \sigma_t z_t \tag{4}$$

z_t is a white noise error term mean and unit variance, and σ_t is the conditional standard deviation of ϵ_t. The GARCH (p, q) variance equation is:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_i \sigma_{t-j}^2 \tag{5}$$

In Equation (4), p is the order of the GARCH terms (lagged variances). q is the order of the ARCH terms (lagged squared innovations). α_0 is a constant term. α_i are coefficients for the ARCH terms. β_j are coefficients for the GARCH terms. The intuition behind the model is that the variance of the series at any given point in time is a function of past squared returns (ARCH terms) and past variances (GARCH terms), and this enables the model to capture volatility clustering.

4.2. Analysis of the GARCH model's simulation

4.2.1. Augmented Dickey-Fuller Test Analysis

Similar to the ADF test in the ARIMA model, the GARCH model also introduced the same method to judge whether the data is stationary or not. However, the data that is analyzed is different, it introduced the percent of the change between the current date's close price and the previous one. According to the result, the test standard of this data set is -5.502986. The p-value is 2.0514×10^{-6}, which is way less than 0.05, so this data set could be stationary.

Table 2: Comparison of critical value and test standard for the GARCH model

Type of Critical Value	Values	Compare with Test Standard
Critical value 1%	-3.458855	Greater than -5.502986
Critical value 5%	-2.874080	Greater than -5.502986
Critical value 10%	-2.573453	Greater than -5.502986

For each type of critical value, they are all greater than the test standard, so this result can indeed reject the null hypothesis. Thus, the data is stationary in this case.

After that, it built a simple ARCH model, but the result shows that the simple ARCH model is not enough for our requirements. Therefore, it requires optimization of the model, and it comes out p and q values. To prevent complex calculation cost, the p and q value is defined in the interval of [1,5], and the best combination in this interval is (p, q) equals to $(3, 3)$.

4.2.2. Hedgehog plot

Hedgehog Plot, a visualization tool that helps understand conditional heteroskedasticity in GARCH models, is measured in the figure 6 below:

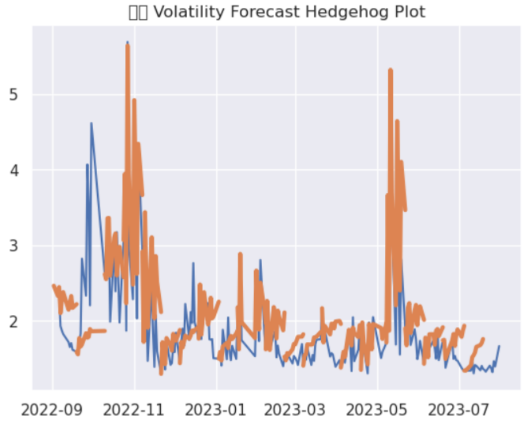

Figure 6: Hedgehog plot for the volatility of the GARCH model

In a hedgehog plot, there are usually two lines, one showing the estimate of the conditional variance and the other showing the standard deviation of the conditional variance. The changes in these two lines show how volatility changes in the time series. Typically, you will see that the conditional variance estimate is higher in some periods and lower in others, reflecting volatility and non-stationary in the time series. In the figure, it reflects that from Sept. to Nov. in 2022 and from May to Jun. in 2023, the volatility is relatively high.

4.2.3. Prediction with machine learning of the GARCH model

After selecting the optimized model and analyzing the volatility, it is practical to use the model to implement prediction. The prediction and actual value is shown in Figure 7 below:

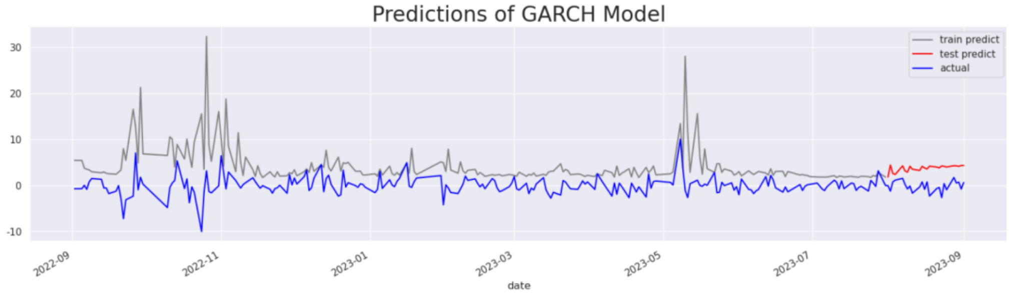

Figure 7: Prediction and actual value

In **Figure 7**., the prediction value and actual value have some similarities, but the absolute difference is relatively high. Thus, it is still necessary to refine the model.

This chart plots a graph containing line and scatter plots to visualize return data versus VaR (value at risk). The different colored scatters represent different scenarios: black represents a VaR that exceeds 5%, red represents a VaR that exceeds 1%, and purple represents not exceeding either risk level. This helps to visualize and understand the relationship between return and risk. What the model founds:

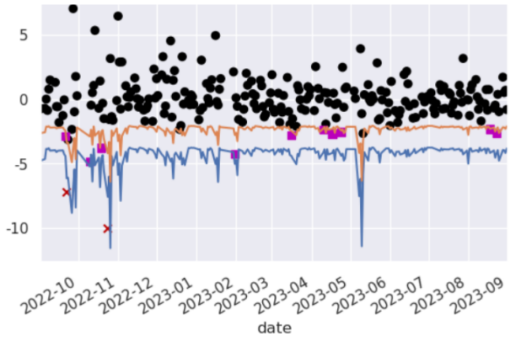

Figure 8: Parametric VaR for risk of trading stock

According to the figure 8, it found that most of the data points is “black”, which means the risk of this model is high. Thus, it is not safe to trade in this situation.

5. Conclusion

In the analysis of the company ZPPCL, two types of models are introduced in the stock price simulation. The data that is selected from Sept. 2022 to Jul. 2023 is non-stationary of its stock close price, but the volatility is stationary. For the simulation and analysis of the ARIMA model, it found that the comparison of prediction and actual value is not satisfactory. Same in the GARCH model, the simulation also does not output an effective predicting model. Through both methods, it reflects that both models cannot effectively predict the ZPPCL in the stock market under the influence of COVID-19. During the research, there were some problems: firstly, there is only one stock included, which means the training data is relatively not enough; secondly, in our training set, the model over-trained the data, and meanwhile, in our simulation of the ARIMA model and GARCH model, the interval of p and q is limited under 5 (the actual best choice could be larger than 5). Therefore, some ways can improve the simulation: 1. Selecting more stocks and lengthen the time of selecting data; 2. Choose more complex models other than the ARIMA and GARCH model (using the method of deep learning to develop new models); 3. Analyzing the industry rotation to analyze some other factor's contribution.

At least, the results of the simulation, it proves that the simple case of these two time-series forecasting models could not simulate a fitting prediction of this stock. In this case, it represents that the factors that affect the stock close price could be some other factors, including political factors. China's desegregation of outbreaks in 2023 may also be a key factor in forecast accuracy. Anyway, under the influence of COVID-19, the medical industry in China's stock market has high volatility and not easily be able to predict. Thus, the risk of trading the stock is also relatively high.

Acknowledgment

Ziyue Mao and Ruixiang Wu contributed equally to this work and should be considered co-first authors. Zhao Dong and Xian Zhang contributed equally to this work and should be considered co-second authors.

References

[1] International Trade Administration (2023, April 7th) China-Country Commercial Guide. Healthcare. ITA https://www.trade.gov/country-commercial-guides/china-healthcare

[2] Shi, G. F., Li, M., Shen, T. T. and Ma, Y (2021) The Impact of Medical Insurance on Household Stock Market Participation: Evidence From China Household Finance Survey. Front. Public Health 9:710896. doi: 10.3389/fpubh.2021.710896. https://www.frontiersin.org/articles/10.3389/fpubh.2021.710896/full

[3] The Ministry of Finance and the National Health Commission of the People's Republic of China. (2020, January 28). China issues 9.95b yuan additional funds for public health, epidemic control. MOF&NHC http://en.nhc.gov.cn/2020-01/28/c_76023.htm (accessed March 7, 2020)

[4] Wang, Z. X., Dong, Y. L. and Liu, A. L. (2022) How does China's stock market react to supply chain disruptions from COVID-19? ELSEVIER https://www.sciencedirect.com/science/article/pii/S1057521922001326

[5] Shu, Y., Yu, M., Yang, O. W. W., & Liu, J. (2003, January). Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models. ResearchGate. https://www.researchgate.net/publication/30952666_Wireless_Traffic_Modeling_and_Prediction_Using_Seasonal_ARIMA_Models

[6] Ji, S., Yu, H., &amp; Zhang, Z. (2016, December 23). Research on sales forecasting based on Arima and ... - ACM Digital Library. ACM Digital Library. https://dl.acm.org/doi/abs/10.1145/3028842.3028883

[7] Aasim, Singh, S. N. and Mohapatra, A. (2019) Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting. EconPapers. doi: 10.1016/j.renene.2019.01.031

[8] Vo, N., &amp; Ślepaczuk, R. (2022, January 20). Applying hybrid Arima-SGARCH in algorithmic investment strategies on S&amp; P500 index. MDPI. https://www.mdpi.com/1099-4300/24/2/158

[9] Gong, Y., Li, Z. P., and Peng, L. (2010) Empirical likelihood intervals for conditional Value-at-Risk in ARCH/GARCH models. Wiley InterScience. doi:10.1111/j.1467-9892.2009. 00644.x

[10] Hayes, A. (2022) Autoregressive Integrated Moving Average (ARIMA) Prediction Model. Investopedia https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical,values%20based%20on%20past%20values

[11] Shahriar, S. A., Kayes, I. Hasan, K., Hasan, M., Islam, R., Awang, N. R., Hamzah, Z., Rak, A. E. and Salam, M. A. (2020) Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for Atmospheric PM2.5 Forecasting in Bangladesh. MDPI. https://www.mdpi.com/2073-4433/12/1/100

[12] Jordan, M. I. and Mitchell, T. M. (2015) Machine learning: Trends, perspectives, and prospects. SCIENCE