Predictive Analysis Based on Prophet Model: Evidence from the Number of Epidemic Infections

Jiachen Pan[1, *], Zitong Xu[2]

[1]Master of Engineering, University of Toronto, Toronto, Canada

[2]Department of Economy, Nankai University, Tianjin, China

* Corresponding Author Email: Jiachen.pan@mail.utoronto.ca

Abstract. This study makes a prediction of infection numbers in the future and investigates the correlation between the new cases and the stock price of companies in three different fields. In the first part of the prediction, we use three methods to make predictions of the future infected number, including the ETS, Auto-Regressive Moving Average (ARIMA), and Prophet model. In the next part, we use the GARCH model to discuss the correlation between infection numbers and stock prices. Considering the heterogeneity, we choose three different companies' stocks to represent three industries respectively. They are Apple, Amazon and Pfizer. The result indicates that the stock price of Apple is negatively correlated to the infection number, the relationship between the infection number and the price of Amazon is inconspicuous, while the price of Pfizer is positively correlated with the development of the pandemic. Thus, we draw the conclusion that the impact of the epidemic is strong to manufacturing, is not obvious to the Internet industry, and boosts the development of the medical corporation.

Keywords: Prophet, ARIMA, ETS, GARCH, Forecasting, Correlation.

1. Introduction

In the past three years from 2020 to 2023, the affection of COVID-19 to the global economy is a hot topic in recent years. The spread of the pandemic, which can be measured by the increased number of infections, caused a serious impact on the global economy. Meanwhile, the effect of the pandemic on different fields and industries is obviously heterogeneous. In the first part, one of our targets is to predict the infection number in the world. The total case number of the world go through a lot of sudden fluctuation, such as the different policies adopted by different government, with the mutation of the virus, such as Omicron. Thus, more accurate model should be introduced in order to take these abnormal events into consideration. In the second part, we are trying to figure out how the rise of the epidemic is related to the development of different industries. The former can be drawn with the new case number or the total case number, while the latter is usually measured by the stock price. We choose three industries: technological manufacturing, the Internet industry, and the medical industry. The company which may reflect the whole industry in respectively Apple, Amazon, and Pfizer. In this research, we are going to use the prophet method to forecast. Prophet is a new method, capable of predict not only fitted values but also errors. The data comes from the open-source of WHO. We will process the data in several different ways. After using those different models, we compare the indicators of those methods, including MSE, RMSE, MAE, and so on. Besides, we will use the machine learning method to deal with data that have gaps. Eventually, we reach the conclusion that the prophet has the best forecast results.

However, the method also has some shortcoming, which lies not only in the method itself but also in reality. In the future, we will make perfections of our results.

2. Literature Review

During the COVID-19 pandemic, the Prophet model, as a time series forecasting tool, has been widely used in predicting the development trend of the epidemic. Some focuses on the Global view, the research by G. Battineni et al. focuses on the four most affected countries in the world, using the

Fb-Prophet machine learning model for prediction. This study highlights the potential of Prophet models for processing large-scale, complex datasets [1]. Also, some scholars do the research about the specific region. The study by Vatsal Tulshyan et al. focuses on forecasts for the next 30 days in India, while Iqra Sardar et al. explore the situation in SAARC countries [2,3]. These studies show the applicability of the Prophet model in different geographical and cultural contexts. Some studies also compared the Prophet model with other forecasting methods such as ARIMA and Random Forest. For example, research by Iqra Sardar et al found that, in some cases, Prophet models may be less accurate than others. This finding reminds us that, while Prophet models are easy to use and understand, caution is still required in choosing the model that best fits a particular situation [3]. The study by Safar M. Alghamdi et al. uses fuzzy time series and Gaussian mixture models combined with Prophet models to demonstrate the potential of interdisciplinary approaches in predicting complex phenomena [4]. Although the Prophet model has performed well in many studies, we must also admit its limitations. For example, the Prophet model mainly relies on historical data and may not fully capture the impact of emergencies or government interventions on the epidemic. In addition, the accuracy of the model may be limited by the quality and availability of data.

Overall, the Prophet model provides valuable insights in the prediction of the COVID-19 outbreak, but its limitations and context-specific needs need to be considered when selecting and applying a model. These studies provide valuable lessons for future epidemiological predictions, emphasizing the importance of multi-model, multidisciplinary approaches, and the role of critical thinking in model selection and interpretation.

3. Methodology

The prophet model is a kind of method that improve from the Time Series Decomposition.

3.1. Addictive Method

Time series decomposition is an important technique in time series analysis, which decomposes time series data into several basic components. There are two methods of decomposition: additive model and multiplicative model. The additive method assumes that a time series is the sum of these three components: trend component, seasonal component, and random (or called residual) component which are shown in function (1).

$$Y(t) = Trend(t) + Seasonality(t) + Random(t) \tag{1}$$

3.2. Prophet component

The prophet model is composed of 4 components:

$$Y(t) = G(t) + S(t) + H(t) + \epsilon \tag{2}$$

Similar to the function 2 demonstrated, G(t) denotes the trend function that characterises non-periodic fluctuations in the time series value. S(t) represents periodic fluctuations, such as weekly and yearly seasonality, while H(t) accounts for the impact of holidays that may transpire on irregular schedules spanning one or more days. The error term in this context serves to capture any unexplained or random variations that are not accounted for by the model.

3.2.1 Trending term

In the trending term model *g*, we have two important base functions. One is Logistic regression, and the other is a piece-wise linear function. It changes a little bit in the logistic regression, as we know the core of logistic regression is the sigmoid function.

$$f(t) = \frac{1}{1+e^{-x}} \tag{3}$$

In addition, in the real-time series, the trend of the curve will definitely not remain unchanged. At certain specific times or with some potential periodic curve, the curve will change. Therefore, we can rewrite the sigmoid function to function 4 by adding some time-related term:

$$f(t) = \frac{C}{1+e^{-k(x-m)}} \tag{4}$$

Where C is the curve's maximum asymptotic value, k is the growth rate, and m is the midpoint. When those three values are equal to 1,1, and 0 the sigmoid function is equal to the original one.
At this time, some scholars will study the turning point, which is the so-called change point detection. For example, the red dot line in the figure 1 are the change points of the time series.

Fig. 1 Piece-wise Linear Function with change points

3.2.2 Seasonal term

Seasonal term is based on the equation (5):

$$s(t) = \sum_{i=1}^{N} \left(a_i cos\left(\frac{2\pi it}{P}\right) + b_i sin\left(\frac{2\pi it}{P}\right)\right) \tag{5}$$

Where P is yearly periodic data. This Seasonal term is predicted using parameter a_i and b_i [5,6].

3.2.3 Holiday Term

For the holiday term, prophet have its own holiday file which record every country's holiday . It will setup a k to represent the holiday influence range and add up every holiday.

3.3. GARCH

Apart from prediction, our research also intends to discover the correlation between the infection number and stock prices of some companies. As a classical black swan event, the pandemic acted as a severe shock in the short term. However, not only are the effects heterogeneous, but also some companies are not distinctively affected. Thus, it's reasonable to discuss the correlation of infection people and stock price, which partially reflect the situation and revenue of the companies. One method to test correlation is by using the auto-regressive conditional heteroskedasticity (ARCH) model. Traditional econometrics uses homoskedasticity assumption to deal with the time series problems. The model is created in 1982 by Robert Engle to solve the problem of volatility of the time series. Bollerslev improve the model of generalized auto-regressive conditional heteroskedasticity (GARCH) model. ARCH and its offspring models serves as evaluation volatility of assets, thus can be applied to stock prices. There are five features of volatility. The first is aggregation, meaning the volatility aggregates at certain period at financial market. The fluctuation performs drastically in this term, alternating with smooth fluctuation in the next term. The second is mean reversion. In general, the

volatility will move around the average rate. The third one is time variations. Volatility changes by time. The fourth is called jumping swift, meaning the volatility may go through a sudden change when black swan events happen. The last is called long memory, meaning that for one volatility, values in different periods' relationship will decline as the time goes by [7,8].

First, considering the time series may not be stable, we first change the data into logarithmetics. Next, we use Augmented Dickey-Fuller test to check whether the time series have unit root. For the function (6),

$$y_t = by_{t-1} + \alpha + \epsilon_t \tag{6}$$

The coefficient b describes the relationship of y_t and y_{t-1}, when b=1, the effect of residual will be unpreventable, and the process will actually become a random walk. If the time series is stable, there will not be a unit root. After that, we use the LB test to test the auto-correlation of the residuals and the square of the residuals.

If r_t refers to a logarithmic revenue time series, then define $a_t = r_t - E(r_t|F_{t-1})$ as the new interest rate at t time. The F_{t-1} means the information in the t-1 period. $\{a_t\}$ meet the GARCH(m,s)distribution, if

$$a_t = \sigma_t \epsilon_t \tag{7}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_1 a_{t-i}^2 + \sum_{j=1}^{m} \alpha_1 a_{t-j}^2 \tag{8}$$

Where σ_t^2 refers to variance in t period [9-13].

4. Data

4.1. Original data

The data we use is the daily new cases of Covid-19 infection cases from an online World Health Organization open source (https://github.com/), which time plot is shown in figure 2:

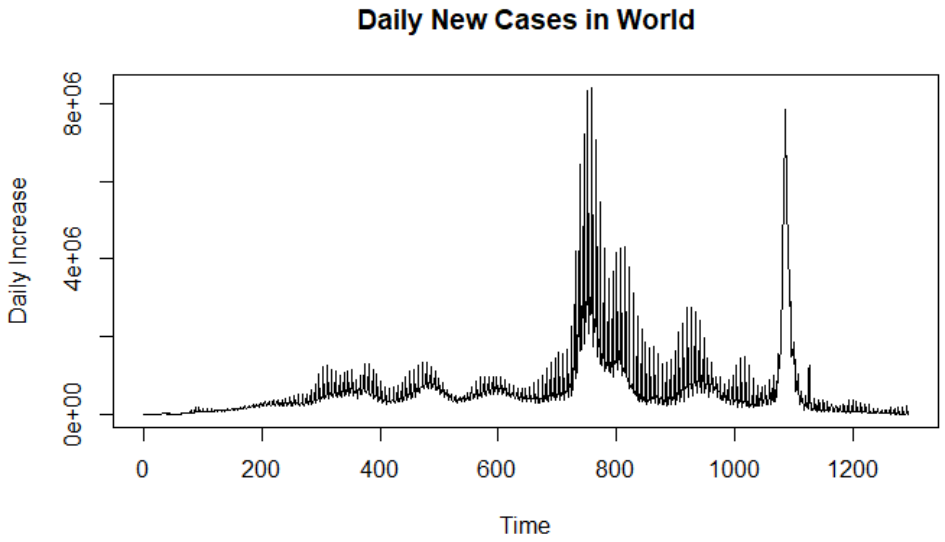
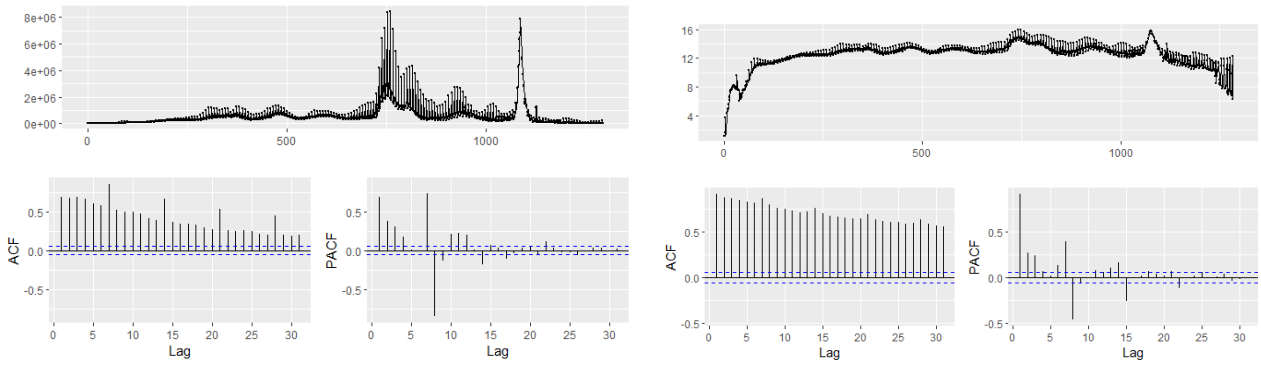


Fig. 2 Daily New Cases in World

4.2. Data Preprocessing

From figure 3(1) time series, we see that the variance between adjacent data is still high, therefore we decide to use the logarithm to scale our data. From figure 3(2), we can see that it stabilizes variance, brings data closer to a normal distribution, linear relationships between variables, shrinks data to reduce the impact of outliers, provides more intuitive interpretation, converts multiplicative effects to additive effects, and ensures non-negative value constraints. These properties make the logarithmic transformation a useful tool for our statistical models and analysis methods.



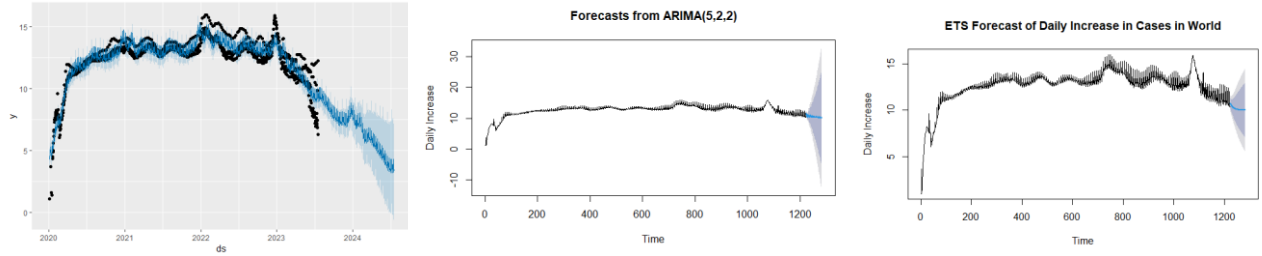
(1)New cases plot (2)Scaling by Logarithm

Fig. 3 Preparing scaling data for fitting model

5. Model Fitting

5.1. ARIMA

In this case, we are going to use the ARIMA model, exponential smoothing model, and the Prophet model to fit the data and do the forecast. At last, we will compare them to the machine-learning method. From the figure 4 and table 1, we can see that the uncertainty in the prophet method is the least and has only 0.776 for mean absolute error. Also, from figure 4(2), the uncertainty interval is huge and gets a higher error, 1.448, than Prophet mode. ETS model has a similar MAE to the ARIMA model which is 1.437.



(1)Newcases Prophet (2)Newcases ARIMA (3)Newcases ETS

Fig. 4 Prophet vs ARIMA vs ETS

Table 1. Model performance

Model	RMSE	MSE	MAE
Prophet	1.0219	1.0444	0.7763
ARIMA	1.8146	3.2928	1.4483
ETS	1.7357	3.0127	1.4371

For the prophet model, we can see that the prediction has lots of fluctuation. First of all, the Prophet method still has lots of disadvantages. Prophet assumes that seasonality is constant, that is, future seasonal patterns are the same as past seasonal patterns. This assumption may not hold in some cases. For example, if your business is undergoing large structural changes, past seasonal patterns may no longer apply in the future. Therefore, the prophet model may provide some wrong peaks in the prediction. Also, beyond the overall trend, there are some small fluctuations in the graph which are caused by weekly seasonality and those fluctuations should not be discarded. In a real situation, the population density will change as the different days in a week. For instance, there will be higher population density on the weekend due to the day off from work and school. Those weekly fluctuations can help people to have proper predictions and do further planning and analysis. For

ARIMA model, it follows the trend but has large uncertainty because this data is more complicated than normal time series which involve more features than the normal one. Moreover, it may have over-fitting when we apply the autoarima function since the training set error is much smaller than the test set which is obviously the result of over-fitting. For ETS model, it has smaller uncertainty range than ARIMA model since it has better performance in dealing with non-stationary data. Due to the data complexity, however, ETS have larger error than the Prophet model.

5.2. Machine Learning

We also try some Machine learning methods, we tried Random forest, Linear regression and regression version of the Support vector machine. The best performance method is Random forest but still has a large error, over 30%, for MAPE test (in Figure 5).

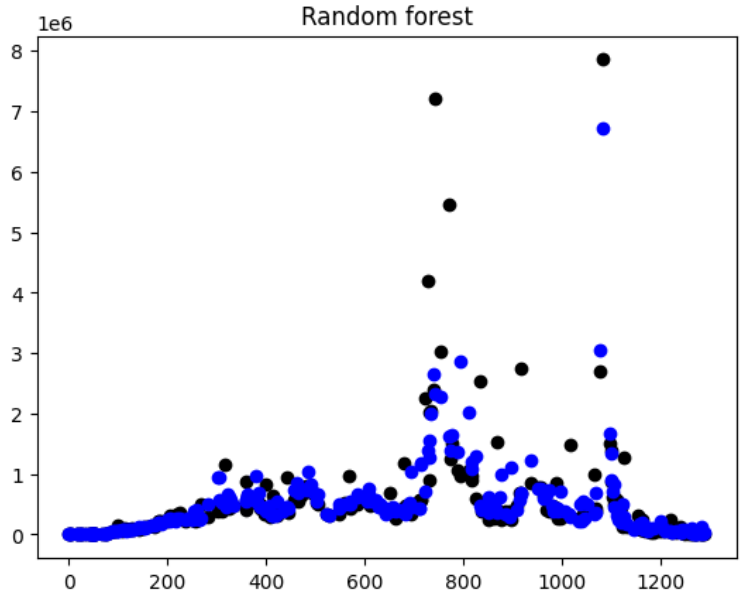


Fig. 5 Prediction VS Test set in Random Forest

5.3. Correlation with Stock Price

Epidemic prediction is not only important in clinical planning and analysis. It is also important in our economic system. Therefore, we use GARCH model to find out the correlation of infection numbers and three iconic stocks from different industry areas (in Figure 6).

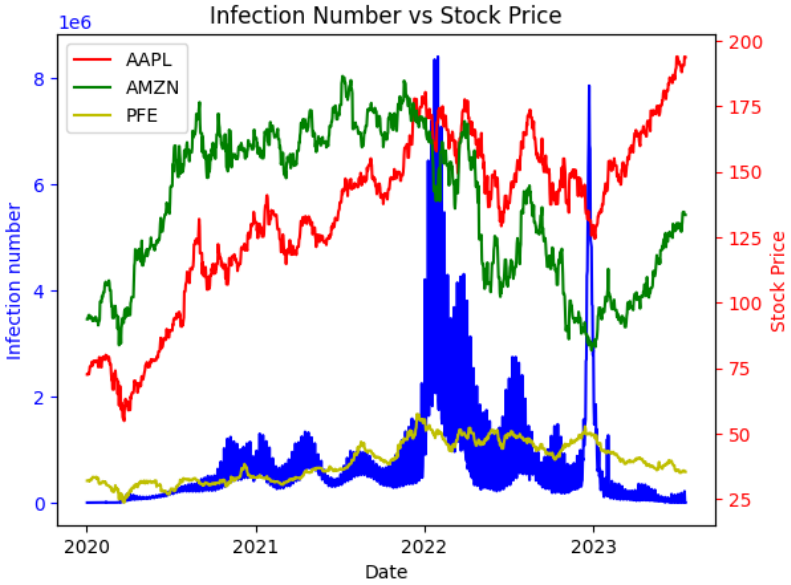


Fig. 6 Infection number vs Stock price

Table 2. Correlation index

Infection cases	AAPL	AMZN	PFE
New cases	-0.24	-0.10	0.42
Total cases	-0.53	0.48	0.17

We use the Pearson correlation index to find out the relationship between the infection number and the stock price. We can see that Apple's and Amazon's stock have a stronger relationship with total cases and Pfizer's stock has a stronger relationship with new cases. Apple's stock price has a negative correlation with the total infection number. This is because epidemics cause lots of unemployment and downturn in the economic environment. Then Apple's stock price will get a negative influence when the epidemic being more serious. Amazon's stock price has a positive correlation with the total infection number. Since when epidemics outbreak people are asked to stay at home and Amazon, which is mainly an e-commercial system, takes its own advantage and get developed. Pfizer's stock price has a positive correlation with the number of new cases of infection number have obvious reasons which is from vaccines needed (Table 2).

6. Conclusion

In conclusion, the Prophet model performs best in prediction and forecasting for infection cases which only has 0.78 for MAE. Furthermore, pandemics also have some impacts on our economic system. From the correlation testing, we can have a better understanding of how these affect some industries. The different time series models considered many factors that may influence the number of patients, thus they can effectively predict on many occasions. However, the methods also have shortcomings, which mainly come from the methods themselves as well as the biological features of the pathogens. The pathogens, especially viruses, can mutate frequently. The variants will be more resistant to present medical treatment, making people infected more easily. Thus, when new variants exist, they may cause a sudden rise in infection numbers, causing unpredictable fluctuations. However, the human body will generate antibodies after some time, and new medicine may be created accordingly. Thus, all variants are not possible to cause long-term deviations. In the later part of the research, we use the GARCH model to describe the dynamic correlation of assets and pandemics, providing a different view to evaluate the product value in the financial market. The fat-tail distribution can be evidently seen. Meanwhile, the volatility can be accurately predicted. However, the model also has some shortcomings. First and foremost, to confirm the model can be applied, a lot of calculations and checks should be made. These perfection and adjustment take us quite a long time and effort. Moreover, the model makes a high requirement for the data. The dataset should not only be stable to a certain extent but also be sufficient. Finally, the prediction of volatility may lose accuracy in the long term. In recent years, we can see an increasing number of unpredictable events, including pandemics. So long-term prediction cannot be exactly as expected.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Battineni, G., Chintalapudi, N. and Amenta, F. Forecasting of covid-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by FB-Prophet Machine Learning Model, Applied Computing and Informatics ,2020.

[2] Tulshyan, V., Sharma, D. and Mittal, M. An eye on the future of covid-19: Prediction of likely positive cases and fatality in India over a 30-day horizon using the Prophet Model, *Disaster Medici*ne and Public Health Preparedness, 2020, 16(3), 980–986.

[3] Sardar, I. et al. Machine Learning and Automatic Arima/prophet models-based forecasting of COVID-19: Methodology, evaluation, and case study in SAARC countries, Stochastic Environmental Research and Risk Assessment, 2022, 37(1), 345–359.

[4] Alghamdi, S. et al. Using fuzzy time series forecasting and gaussian mixture model to classify and predict new cases of COVID-19 in Saudi Arabia, Thermal Science,2022, 26(1), 261–270.

[5] Makarovskikh T .System for Forecasting COVID-19 Cases Using Time-Series and Neural Networks Models.Engineering Proceedings, 2021, 5.

[6] Gull, K., Kanakaraddi, S. and Chikaraddi, A. Covid-19 outbreak prediction using additive time series forecasting model, Trends in Sciences, 2022,19(22), 1919.

[7] Jinguan Lin, Yizhi Mao, Hongxia Hao. Realized GARCH Models with Time-Varying Leverage Effects. 2023.

[8] Zhiqiang Hu. Research on the linkage between China and the United States stock market based on the DCC-GARCH model. 2022.

[9] Abdullah, A.M. The impact of covid-19 and the Russia–Ukraine conflict on the relationship between the US islamic stock index, Bitcoin, and Commodities, Asian Economics Letters, 2023, 4(2).

[10] Jamil, I. et al. Pre- and post-covid-19: The impact of US, UK, and European Stock Markets on ASEAN-5 Stock Markets, International Journal of Financial Studies, 2023,11(2), 54.

[11] Chen, J. et al. The impact of covid-19 on the US oil stock market and currency, Advances in Economics, Management and Political Sciences, 2023, 3(1),182–191.

[12] Zhao, J. The long-term impact of covid-19 on US Stock Market: Evidence from Time Series model, BCP Business & Management, 2023, 38, 1476–1484.

[13] Taylor, S.J. and Letham, B. Forecasting at scale, PeerJ Preprints. Available at: https://peerj.com/preprints/3190/, 2017.