Predictions of US COVID-19 Pandemic: Comparisons Among Different Models

Article · December 2020

CITATIONS
0

READS
46

2 authors, including:

Yuzhou Wang
University of Wisconsin–Madison
4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Predictions of US COVID-19 Pandemic: Comparisons Among Different Models

Principal Investigators:

Yichen Zhu
zhu389@wisc.edu

Yuzhou Wang
wang2555@wisc.edu

December 2020

1 Introduction and Motivations

As reported by WHO, there have been more than 70 million confirmed cases of COVID-19, including over 1.5 million deaths globally as of 9 December 2020. While the pandemic in most Asian countries has been currently under control, the infection and mortality rates in many countries like the US are still unstable. Recently, new daily coronavirus cases in the United States even reached their latest all-time high which should draw people's close attention. Other than broad health concerns, COVID-19 has also driven an unusual economic downturn with soaring employment rates and uncertainty. We believe that it is necessary to conduct accurate predictions on the infection rate and provide instructive suggestions for residents and governments in the US which form our main tasks. Consider US for being a large import and export entity, we are also interested in how its economic factors vary accordingly with the COVID-19 pandemic.

Our main task for this project is to make a forecast of future pandemic situations. For the sake of selecting the best model to accomplish this task, we compared different models by their predicting performance. We also segregate our analysis into two parts: forecasting on cumulative confirmed cases and on the daily increase in new confirmed cases. For the first part of our study, time-series models including AR2, MA2, ARIMA(2,2,2), and the VAR2 model are applied to predict the one week cumulative US infection rate. In this part, we find VAR2 model has the strongest predicting power. In the second part, besides the ARIMA model, Prophet model from Facebook and a machine learning algorithm XGBOOST have been implemented to forecast daily increase in US new confirmed cases. As a result, we find Prophet algorithm with logistic model and seasonality settings has the strongest predicting power.

The rest of this paper is structured as follows. The next section covers detailed information about the COVID-19 pandemic data used in our project. In Section 3, the basic framework of our proposed method, assumptions, pros, and cons of each model are introduced. Forecasting results and predicting performance analysis are given in Section 4. Section 5 presents the main contribution of this paper and provides alternative directions for future research.

2 Data

We collected data from the "JHU CSSE COVID-19 Data" which can be accessed from `https://github.com/CSSEGISandData/COVID-19`. The time-series COVID19 data file contains both spot and cumulative daily confirmed cases and covers information from 189 countries and regions around the world from $Jan.22^{nd}$ to $Dec.9^{th}$ 2020. The models that we selected, including AR, MA, ARIMA, VAR, etc., will be implemented on time-series data from $Jan.22^{nd}$ to $Nov.18^{th}$ which only cover a very short term of current wild infection. Our study can be segregated into two parts. First, we implement multiple predicting models with the US cumulative confirmed cases. Then we switch to the data set of daily confirmed cases and work with the daily difference of new confirmed cases in US. For a supplemental study, we adjust the data used in the second part of our study into weekly-based value to remove the impact of weekend missing data which could improve forecasting performance. We treat it as a test for our main result.

3 Models Specification

3.1 AR, MA, ARIMA and VAR Model

In this study, AR, MA, ARIMA, and VAR are basic time-series models used to investigate data of US confirmed cases. The lag orders in the AR model and the MA model are decided by the PACF, and the ACF test in the analysis of cumulative data. According to the plot of PACF and ACF, the truncate in the PACF, and the tail in the ACF show that the AR model is better than the MA model. These results also indicate that the AR2 model should be used. As a comparison, MA2 is applied to see whether the PACF and ACF tests are crucial in the model building process. The ADF test of these data shows that the US cumulative confirmed data is stable since the P-value of the ADF test is $2.2e^{-16}$, which rejects the Null hypothesis. From the result of the ADF test, it is unnecessary to take difference upon the original data, which means the ARIMA model is not helpful. However, ADF testing has the disadvantage of low test efficacy. In other words, it is very easy to reject the hypothesis that the data are not smooth. Hence, the result from ARIMA model has been preserved in the paper. The employment rate has a strong relationship with COVID-19, since the growth of the number of the COVID-19 may cause the decrease of the employment rate. The VAR2 model, which includes the employment rate and the cumulative number of the COVID-19, will be established.

3.1.1 ARIMA model

ARIMA consists of three parts, AR, I, and MA. AR stands for autoregression, which is an auto-regressive model, and I stands for integration, which is a single integer number. MA is the moving average model. It can be seen that the ARIMA model is a combination of the AR model and the MA model. The ARIMA model is no exception as a time series model, so the unit root test should be performed first, and if the data is a non-stationary series, it must be converted into a stationary series by differential.[3] The order of the ARIMA model is decided by the AIC test in the code.

The ARIMA(2,2,2) model can be represented as:

$$\hat{Y_t} = \mu + Y_{t-1} + Y_{t-2} - \theta_1 * e_{t-1} - \theta_2 * e_{t-2} \tag{1}$$

$\hat{Y_t}$ is the predicted cumulative covid-19 of moment t. $\hat{Y}_{t-1}$ is the predicted cumulative covid-19 of moment t-1. Y_t is the true cumulative covid-19 of moment t. Y_{t-1} is the true cumulative covid-19 of moment t-1.

3.1.2 VAR model

Vector auto-regression is the process of predicting several economic variables together as a system to make the predictions mutually consistent. The more the variables in the VAR system, the more coefficients need to be estimated. The VAR2 model is based on the simple relationship between the employment rate and the cumulative COVID-19. From the result of code, the lag 10 is chosen.

The VAR2 model is then:

$$Y_{c_t} = \beta_0 + \beta_{c1}Y_{1,t-1} + ... + \beta_{c10}Y_{1,t-10} + \gamma_{c1}Y_{2,t-1} + ... + \gamma_{c10}Y_{2,t-10} + \delta_{c1}Y_{3,t-1} + ... + \delta_{c10}Y_{3,t-10} + \varepsilon_{1t} \tag{2}$$

$$Y_{u_t} = \beta_0 + \beta_{u1}Y_{1,t-1} + ... + \beta_{u10}Y_{1,t-10} + \gamma_{u1}Y_{2,t-1} + ... + \gamma_{u10}Y_{2,t-10} + \delta_{u1}Y_{3,t-1} + ... + \delta_{u10}Y_{3,t-10} + \varepsilon_{2t} \tag{3}$$

Y_{ct} is the cumulative COVID-19 in the t period. Y_{ut} is the employment rate in the t period.

3.2 FB Prophet Model

Prophet Forecasting model is a business-driven model constructed by Taylor and Letham (2017)[4] which allows scale forecasts. Different from Time-series models like ARIMA, the Prophet formulation frames the prediction problem as a curve-fitting exercise rather than simply rely on dependence within periods. The main reason that we include this model is that this model is more flexible than the traditional time-series

approaches. It was built on the decomposed time series model (Harvey and Peters 1990) and have trend, seasonality, and holidays being considered as three main components. They can be interpreted as:

$$y_t = g_t + s_t + h_t + \epsilon_t$$

g_t indicates non-periodic changes in the value of the time series, s_t represents periodic changes such as the periodicity of a week or a year. h_t is the influence of holidays. ϵ_t is an error term.

In this case, users can customize parameters easily with the FB Prophet model. Moreover, both the linear model and the nonlinear logistic growth model are available under this framework.

4 Result

4.1 Predictions on Cumulative Confirmed Cases

In this section, we use the data from $Jan.22^{nd}$ to $Nov.18^{th}$ 2020 to build the AR, MA, ARIMA, and VAR models. Due to the recent explosive growth of COVID-19 cases, possibly due to Thanksgiving, data after late November will be extreme values in the model, making the model less reliable, so the AR, MA, ARIMA, VAR model will not include data after November 18. The data before November 7 will be used as a train set in the model, and the data from November 7 to November 18 will be used as a test set. The comparison of these several models is shown in the figure [**?**].In this figure, it is clear that the predictive effect of the ARIMA(2,2,2) model is the best. It is extremely close to the true data. The error rate is only 0.41%. The VAR2 model's results are very close to ARIMA model. The MA 2 model performs worst. It mainly because the truncate in the PACF, and the tail in the ACF(in the R code), which means MA model is not suitable in this data set.

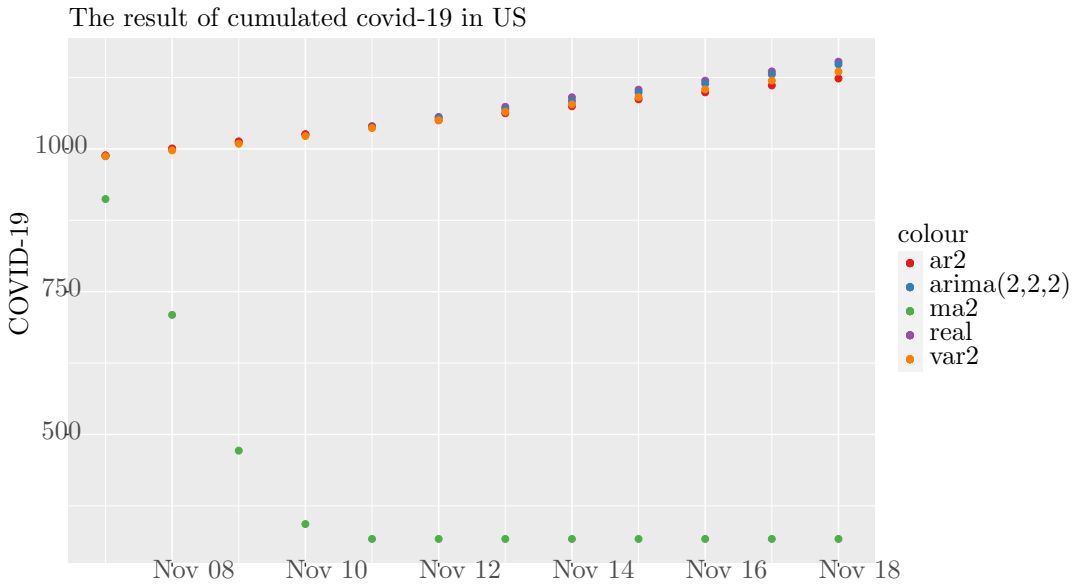

Figure 1: The predictive results of different models

Note: The Y-axis of figure 1 is in tens of thousands of units. The X-axis includes the date. Since the plots of the AR2, ARIMA(2,2,2), VAR2 model are very close to each other, the figure 2 is displayed to show the difference between each other.

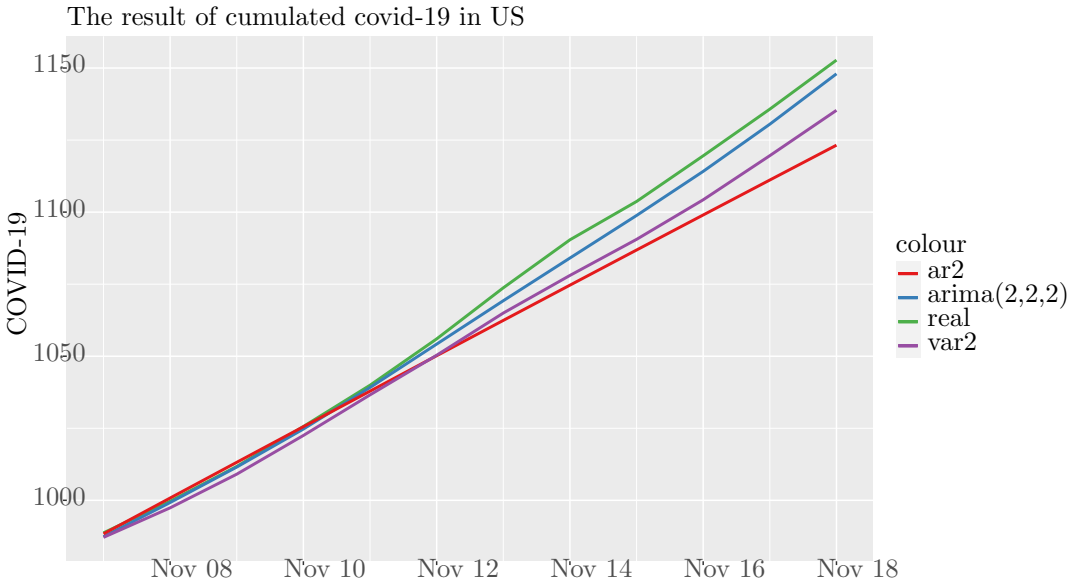

Figure 2: The predictive results of different models

Note: The Y-axis of figure 2 is in tens of thousands of units. The X-axis is the date. It is clear to see the difference between the AR2, ARIMA2(2,2,2), and the VAR2 models. The ARIMA(2,2,2) model is the best among all the predicting models.

4.2 Predicting Performance Analysis on Cumulative Confirmed Cases

In order to compare the performance of different models,the RMSE(Root Mean Squard Error) is used as a criterion to determine the effect of a model. The Root Mean Square Error is a measure of the deviation between the observed value and the true value.

The formula of the RMSE:

$$RMSE = \sqrt{\frac{1}{m}\sum_{n=1}^{m}(y_i - \hat{y_i})^2} \tag{4}$$

In table 1, the RMSE of the ARIMA(2,2,2) model is only 3.76, while the RMSE of the worst MA2 model is 690.35. The RMSE of the AR2 model is much smaller than the MA2 model, which means the PACF and the ACF tests are very important when we establish the time series models. The phenomena that the ARIMA(2,2,2) model is better than the AR2 model indicates that the ADF test is not believable. The disadvantage of the ADF test is that it has low test efficacy (high probability of making Type II errors), especially when the sample size is not large, as in our dataset. We believe that the ADF test rejects the hypothesis of the existence of a unit root in this case. Usually, the VAR model, which is based on economic theory, is superior to the ARIMA model in terms of predictive performance. But here the VAR model is slightly worse than the ARIMA model, probably because we do not have a very scientific theory yet to identify the variables that mainly affect the growth of COVID-19, and if the theory that affects the growth of COVID-19 is established in the future, The VAR or BVAR model may outperform the ARIMA model.

4.3 Prediction on Daily Increase of New Confirmed Cases

4.3.1 ARIMA, Prophet and XGBOOST

In addition to the study on cumulative confirmed cases, we explore the data set which includes US daily new confirmed cases. To examine the severity of the spread, we take the first difference and pass an ARIMA model to the new data. Different from previous study, the Auto non-seasonal ARIMA Algorithm[1] select ARIMA(2,2,1)(0,0,0) to perform the prediction. We allocate 90% of the full samples to the training set. In other words, the remaining samples from 10-18-2020 to 11-18-2020 constitute a testing set. From the figure 3 below, we observe an obvious departure from

Table 1: MSE and RMSE Report

MODEL	MSE	RMSE
ARIMA(2,2,2)	14.1173	3.7573
AR2	216.5252	14.7148
MA2	476593.4721	690.3575
VAR2	105.5965	10.2760

* The MSE and RMSE in the table are mean square error and square root mean squared error correspondingly.

the true observation for our predicting outcome, which has been verified by a large SARIMAX MAE of 50.926. We suspect recent seasonal flu and heated social activities to be the causes.

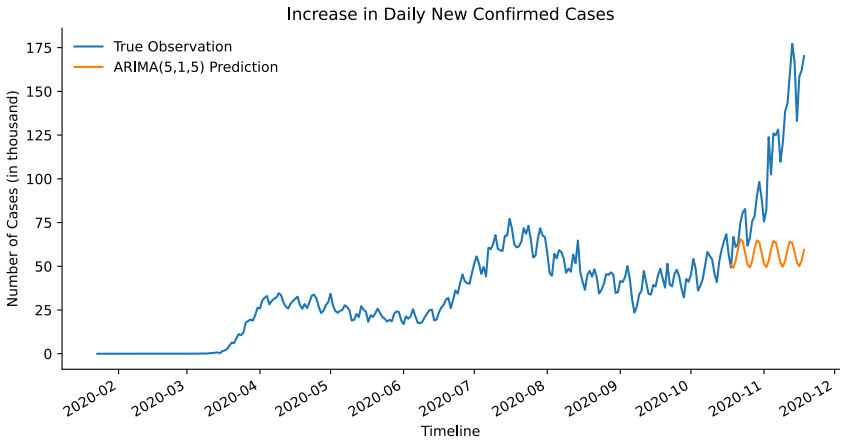

Figure 3: ARIMA: US Daily New Confirmed Cases Increment

Considering the periodical feature of coronavirus testing (closure of test centers on weekends), we include Facebook Prophet Model in our study. It works better with time series that have strong seasonal effects.[4] More importantly, it is robust to missing data and can handles outliers well. Under our expectations, the basic Prophet model exhibits a similar fitting performance while the one with seasonal adjustment fits better for US daily incremental confirmed cases, comparing with the corresponding ARIMA outcome. Detailed predicting performance analysis will be covered in the following section.

Figure 4 illustrates the forecasting outcomes generated by a basic Prophet model. It highly depends on linear relations of temporal periods which may be neither realistic nor suitable for the prediction of infection growth. We consider Logistic Growth as a better pattern in fitting COVID-19 infection data as it is characterized by an exponential growth in the early stage but transforming to a decreasing trend later, which can be presented as an S-shape. As Wang(2020) [5] suggests in their COVID-19 forecast study, the epidemic situation presented an explosive exponential growth trend as soon as the infection reached a certain point, and then with regulation and public's cooperation, the infection gradually slowed down, approaching to its upper bound. Hence, we conduct another prediction with a Logistic fitting model under Prophet and add customized seasonality which is shown in figure 5. It indeed fits better than both the ARIMA and the basic Prophet model by being able to catch an upward trend of increase in daily new confirmed cases.

Different from the time-series models and the Prophet method we have mentioned in previous sections, XGBOOST is a Machine Learning Process.[2] It is usually used for supervised learning problems in ML, where we use the training data to predict a target variable. Instead of making assumptions on the growth pattern of infection, XGBOOST is built on a decision tree framework.[1] Here in our analysis, we transform our time-series data into multiple segments (day, month, quarter, day of the week, etc.), treating them as a set of independent variables X to predict the dependent variable - daily increment value of new confirmed cases. We compare the outcome generated by this tool to traditional time-series methods and Prophet to check whether the ML process can scratch more information from the data to obtain a better predicting performance. The result in figure 6 shows that the forecasting result is very similar to the product given by Prophet.

We made both 20-period and long term forecast basing on the Prophet model which is shown in figure 7 and 8 in the

[1] Check https://xgboost.readthedocs.io/en/latest/index.html for detailed information.

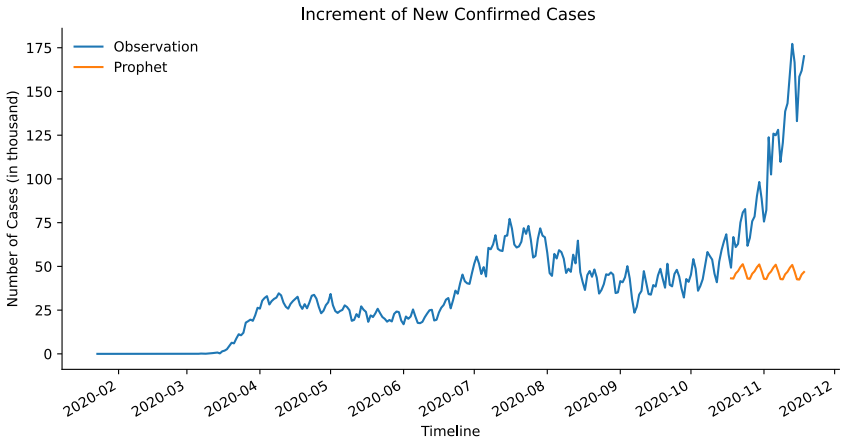

Figure 4: Basic Prophet

Appendix. The former predicted a daily increase in new confirmed cases to reach 190 thousand in the following days. It seems to have fitted better and caught the high level of infection because it has been trained on a larger data set which includes the most recent explosion growth period. A longer forecast is presented in figure 8 in the Appendix section. We examine a soaring trend in the following season.

4.3.2 Weekly Adjustment

Testing centers and data maintenance groups may close on holidays and weekends, this may account for missing data in our data set. Moreover, the testing service may highly depend on the availability of testing center appointments and detection reagents. The above causes make further adjustment of data necessary. Figure 9 and 10 in the appendix suggest that Prophet model fits a lot better than the ARIMA model for weekly adjusted data which verifies the result on daily data in the previous subsection.

4.3.3 Predicting Performance Analysis

Table 2: MAE and RMSE Report

MODEL	MAE	RMSE
ARIMA(5,1,5)	50.9260	62.6179
Linear FB Prophet	61.2131	71.4905
Seasonal Adjusted Logistic FB Prophet	50.3364	61.4346
XGBOOST	54.0431	64.7467
ARIMA(2,2,1) with Weekly Data	8232.0251	14542.4407
Logistic FB Prophet with Weekly Data	8367.0465	13433.6650

* The MAE and RMSE in the table are mean absolute error and square root mean squared error correspondingly. Comparing with MSEs, RMSEs are less sensitive to units.

From the table above, we first observe that the Seasonal Adjusted Logistic FB Prophet shows the best performance in prediction. Although MAEs and RMSEs for the last two models with weekly data are not comparable to the first three, Prophet model still predicts better than ARIMA within both groups. Moreover, the predicting power of ML process XGBOOST does not perform better than the adjusted Prophet model which contracts our expectations. It even performs worse than the ARIMA model. This unpleasant forecasting result maybe attributes to unnecessary correlations caught by the learning process and the underrating of periodical dependency. We believe this algorithm can predict better with more messy data.

We do not think these forecasting models which are trained on the training set perform very well for the test set. And this problem can be attributed to the following reasons. Firstly, the supremum of daily increase in US new confirmed cases is 177 till Nov. 18^{th}. We examine large gaps between the ceiling of predictions and the true maximum value for all forecasting models. Prophet gets a max value of 63 with control of seasonality and with the logistic fitted model. This gap may appear because the fitting model behind Prophet counts on the cumulative cap value of historical data.

Secondly, the true infection are extremely sensitive to recent unexpected increments in number of patients, which can further widen the prediction gap.

5 Conclusion

In this study, we have accomplished the task of making forecasts on future coronavirus infection. In the short-run, the daily increase of new confirmed cases is predicted to keep rising at a high level. Prediction for a longer period reflects a slow-down trend in terms of daily increasing cases. Moreover, our model suggests that the daily increase of new confirmed cases will reach its next peak around March 2021. This forecasting result sets up an alarm for all US residents and policy-makers. Even with the conservative Prophet model which is limited by cumulative maximum values, the upward trend is steep and will take more than three months to recover. Although this can be partially attributed to the issue of seasonal flu, we should be careful about whether the current explosion of infection is related to relaxation of local prohibition, worse implementation of social distancing orders, and residents' negligence on the pandemic issue. The government and public institutions need to stay vigilant to the dynamic situations and help residents keep safe and guarded against the coronavirus. After all, the public physical and mental health condition is highly correlated with local security and political stabilization.

Although there are a great number of existing papers that accomplished tasks of making predictions for the COVID-19 pandemic, few of them focused on the cases in US and made comparisons among models that we chose for analysis. The main contribution of this paper is that we use different methods, including the traditional time series models and the newly built models, containing the Prophet model, XGBOOST to predict the COVID-19 data. For our exploration in cumulative confirmed US COVID-19 data set, ARIMA model presents the best predicting power. When it comes to the daily increase of new cases, the FB Prophet model with logistic and seasonality out-performs ARIMA model and the XGBOOST machine learning process. Another unique feature of this study is the combination of R and Python programming which have made models more versatile and have enriched the visualization of different pandemic outcomes.

There are imperfections in our forecasting models. And here we summarize alternative works to accomplish in the future. For now, most methods used in this study are highly dependent on the split of the training set and test set. This problem can be addressed by the machine learning process. Further improvements in a method that can better combine time-series models with the machine learning process might be considered. Also, all methods being selected in this paper are not able to catch the exact dramatic up-ward trend of infection. More works can be done to fix this issue. Furthermore, more accurate and robust VAR models can be built if the factors influencing COVID-19 are found.

References

[1] Ansari Saleh Ahmar and Eva Boj del Val. Suttearima: Short-term forecasting method, a case: Covid-19 and stock market in spain. 2020.

[2] Trevor Hastie Robert Tibshirani. Gareth James, Daniela Witten. An introduction to statistical learning : with applications in r. *New York: Springer*, 2013.

[3] Chen Qiang. Advanced econometrics and applications stata. 2009.

[4] Taylor SJ and Letham B. Forecasting at scale. 2017.

[5] Peipei Wang, Xinqi Zheng, Jiayang Li, and Bangren Zhu. Prediction of epidemic trends in covid-19 with logistic model and machine learning technics. *Chaos, Solitons Fractals*, 139:110058, 2020.

6 Appendix I Adjusted Prophet and XGBOOST Outcome

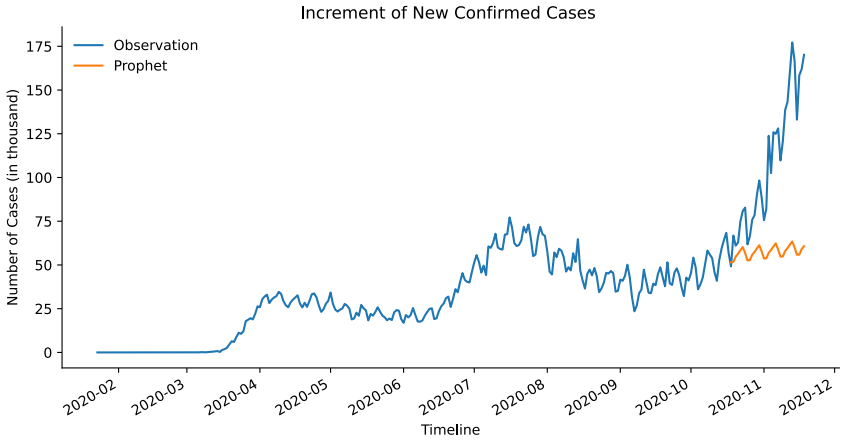

Figure 5: Prophet with Logistic and Seasonality

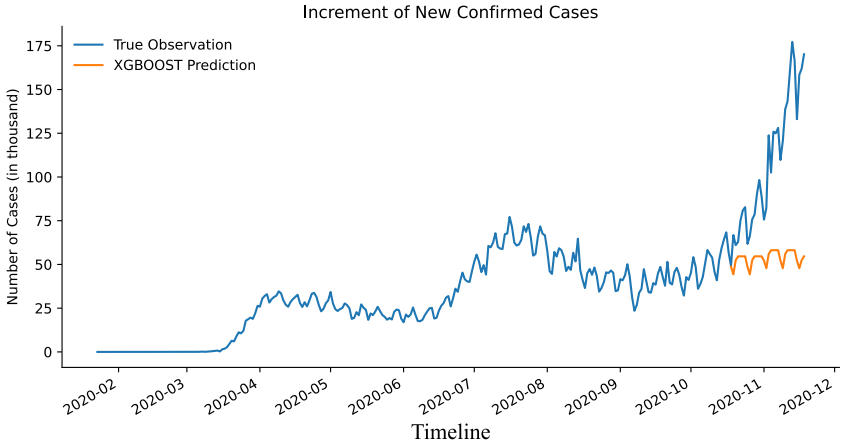

Figure 6: XGBOOST

7 Appendix II Forecast Outcome

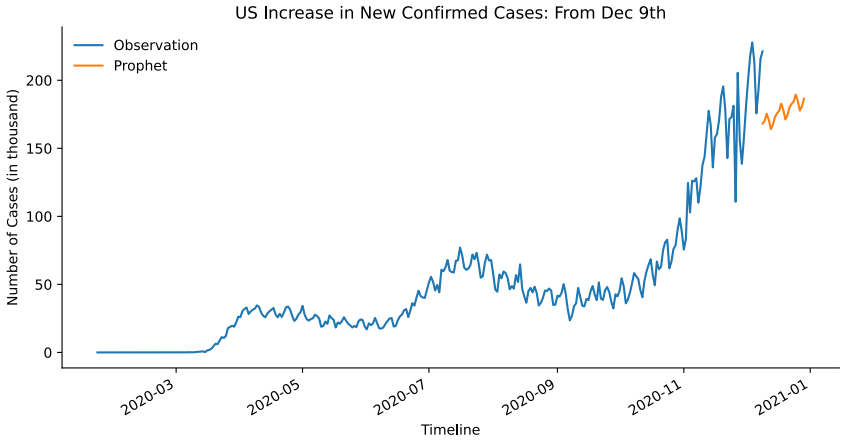

Figure 7: 20-period Forecasting: Prophet with Logistic and Seasonality

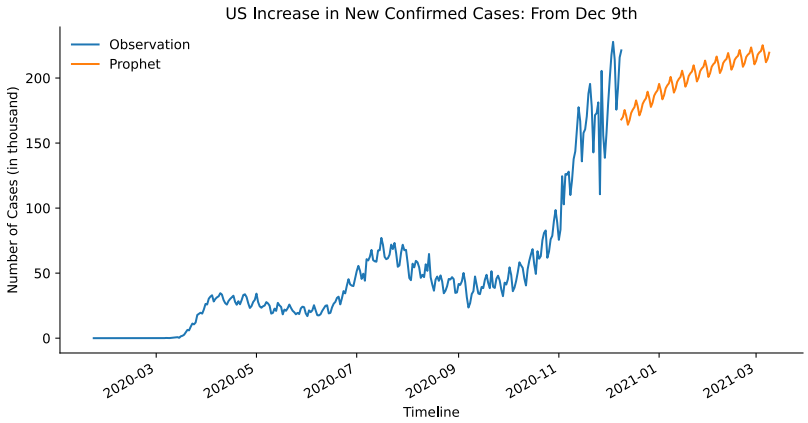

Figure 8: 90-period Forecasting: Prophet with Logistic and Seasonality

8 Appendix III Weekly-adjusted Performance

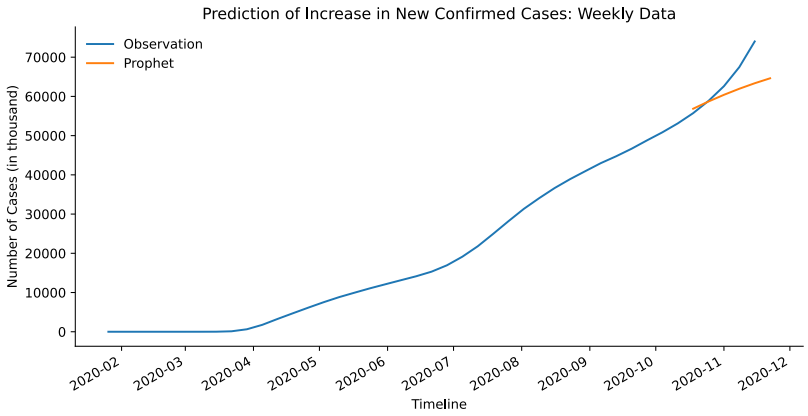

Figure 9: ARIMA: US Weekly New Confirmed Cases Increment

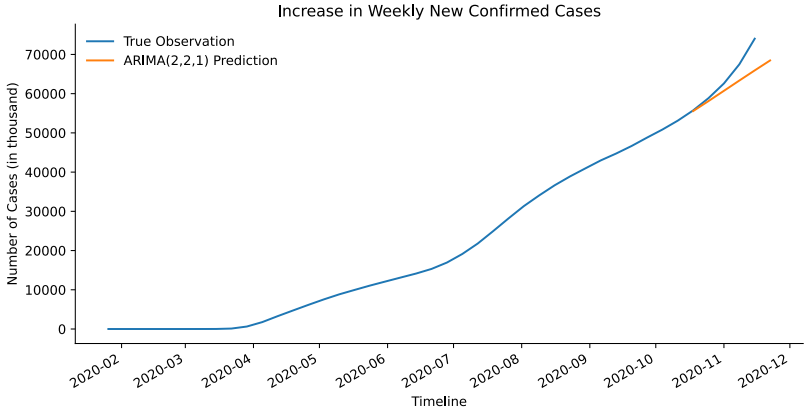

Figure 10: Prophet: US Weekly New Confirmed Cases Increment