Application of Machine Learning techniques in building a time series forecasting model of the number of COVID-19 infections

Linh My Thi Tran[1,2,3], Hanh Hong Thi Duong[1,2,3], Thien Long Nguyen[1,2,3], and Trong-Hop Do[1,2,4]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]*{18520999, 18520711, 18521046}@gm.uit.edu.vn*
[4]*hopdt@uit.edu.vn*

Abstract. Building predictive models is a common scientific task that enables humans to plan or detect anomalies. In particular, a lot of scientific research related to Covid-19 has been and is being carried out. Many aspects related to this pandemic are included in the development of infection prediction models. In this project, we focus on learning about time series prediction methods, combined with various data processing techniques and machine learning algorithms to solve the problem of predicting the number of cases. infected with Covid-19. Through the experimental process, we initially propose the Extreme Learning Machines model with respective MSE and sMAPE of 33.10^9 and 13%. Through this study, many new features of the data and models were discovered and allowed us to discuss in more depth the potential future development avenues of this problem.

Keywords: Time series · Forecasting · Covid19

1 Introduction

Appearing at the end of 2019 in Wuhan, China, the Covid-19 pandemic has so far swept through most countries in the world, demonstrates its terrible destructive power on the development of the entire humanity, specifically in areas such as economy, health, society,... In order to prevent the spread of the disease and propose methods to respond to the pandemic, a practical problem has been posed - predicting the development of the disease.

To build a good predictive model for this problem, it is necessary to rely on many factors, including natural features (such as geographical location, climate,...) and social features (such as education level, population density, indigenous culture,...). But in this report, our team relies on only one attribute - the time series - to predict future disease development.

Therefore, the report use prediction techniques based on time series data - the method of analyzing the statistical data of processes that have taken place, recorded at successive intervals of time with the goal of using past experience to predict what will happen in the future . Our goal is to be able to build predictive models for the number of infections worldwide or in large regions using both individual and aggregate approaches. However, with limited resources, we initially carried out experiments with a small number of specific countries that are currently experiencing complicated Covid situations.

In this report, we focus on introducing information related to performance goals in the 1 section and mention information about related works in section 2. Next, the 3 and 4 sections are, respectively, the place to present information on collection, data set construction and the posed problem. In the section 5, we review the preprocessing methods, models, and evaluation methods that have been applied during the experimental process. In addition, the actual approach, testing and results will be evaluated and analyzed in the 6 section. Finally, the conclusion and future development direction in section 7.

2 Related works

Time series data is one of the most common data types found in the collection and storage of information in many fields. Because of the popularity of time series, there has been a growing need for performing different tasks on time series data. That's why Time series modeling for predictive purpose has been an active research area of machine learning for many years. However, the appearance of too many models of the same type will greatly hinder newcomers in accessing this field. To solve that problem, Fatoumata Dama and Christine Sinoquet[1] have released a scientific report: "Time Series Analysis and Modeling to Forecast: a Survey" (2021) - an extremely valuable document for people who do jobs that use time series. In this report, the authors have synthesized many time series prediction models, explained many concepts, presented how it works, and provided instructions for performing various operations.

Besides, academic works related to predicting the number of Covid-19 infections appear more and more and add a lot of references for researchers. It is hard to not mention Kayode Oshinubi et al[2]. In this report, the authors refer to the national approach to training models with many deep learning methods such as LSTM, GRU, CNN, and DNN.

3 Dataset

Face with the increasingly unpredictable Covid-19 situation in Vietnam, and around the world in general, we decide to apply what we had learned to build a forecasting model based on Time Series data.

My dataset is collected from DATAHUB, a well-known platform for sharing and publishing high-quality datasets. It includes data on the number of confirmed infections every day from eight countries: China, the United States, the United

Kingdom, Italy, France, Germany, Spain, and Iran. The table 1 has more information.

Table 1: Detailed information about dataset.

Number of data points: 732.			
Number of data points: 10.			
Feature	**Data type**	**Definition**	**Data range**
Date	Datetime	Date.	01/23/2020 to 01/23/2022.
China	Integer	Number of confirmed cases by day in China.	[0,15136]
US	Integer	Number of confirmed cases by day in US.	[0,1368382]
United_Kingdom	Integer	Number of confirmed cases by day in UK.	[0,225760]
Italy	Integer	Number of confirmed cases by day in Italy.	[0,228123]
France	Integer	Number of confirmed cases by day in France.	[0,456270]
Germany	Integer	Number of confirmed cases by day in Germany.	[0,140870]
Spain	Integer	Number of confirmed cases by day in Spain.	[0,372766]
Iran	Integer	Number of confirmed cases by day in Iran.	[0,50228]
total	Integer	Total number of confirmed cases by date of 8 countries.	[0,2107541]
See more about the dataset here: Dataset			.

4 Task

In this research, we use the time aspect to forecast the number of confirmed cases by day of countries. With my dataset, our main goal is to solve a common problem of forecasting the total confirmed cases of all 8 countries.
Input: A time mark by day.
Output: A respective number of confirmed cases.

5 Methodologies

5.1 Data preprocessing

Time Series data is a special type of data, so for the convenience of interacting with the data, we take advantage of the support of TSDataset - a Chronos module that permits an abstract of the time series dataset. TSDataset provides a plethora of preprocessing methods as well as Feature Engineering approaches to give effectively support Time Series prediction. With this project, we use some basic functionalities from TSDataset such as : *impute()*, *deduplicate()*, *gen_ dt_feature()*, *roll()*, *scale()*. Specifically:

- *impute()*: resolve missing data.
- *deduplicate()*: remove duplicate data points.
- *gen_ dt_feature()*: generate new datetime feature(s) for each record from the existing datetime feature. Those features aid both humans and the model in better understanding the data.
- *roll()*: this is a sampling by rolling support for Machine Learning and Deep Learning models.
- *scale()*: allows performing feature normalization. We mainly use two methods of normalization from *sklearn.preprocessing* [1], are *MinMaxScaler* và *StandardScaler*.

The preprocessing steps do not be performed in a fixed way, but always be adjusted and changed to best suit each specific data.

5.2 Model

LSTMForecaster[2] Long short-term memory (LSTM) is a type of RNN whose main idea is to connect previous data to predict current data. Because of this, LSTM is widely used in time series forecasting. LSTMForecaster implements a basic version of LSTM, called VanillaLSTM, with 2 LSTM layers, 1 dropout layer and 1 dense layer.

TCNForecaster[3] Temporal Convolutional Networks (TCN) is based on CNN to learn from data over time. It also allows large-scale parallel computation, which takes less time than LSTM.

AutoTSEstimator[4] As one of the critical components of the pipeline that AutoTS provides, helping with distributed hyperparameter tuning is supported by Orca[5] automatically by providing functions and time series optimization. We can use AutoTSEstimator to find the best models for the three built-in model types LSTM, TCN, Seq2Seq, or a third-party model.

AutoRegression A time series model predicts future data using data from a previous period as input into a regression equation.

[1] https://scikit-learn.org/stable/modules/classes.htmlmodule-sklearn.preprocessing

[2] https://analytics-zoo.readthedocs.io/en/latest/doc/PythonAPI/Chronos/forecasters.htmllstmforecaster

[3] https://analytics-zoo.readthedocs.io/en/latest/doc/PythonAPI/Chronos/forecasters.htmltcnforecaster

[4] https://analytics-zoo.readthedocs.io/en/latest/doc/PythonAPI/Chronos/autotsestimator.htmlid1

[5] https://analytics-zoo.readthedocs.io/en/latest/doc/ Orca/Overview/orca.html

XGBoost XGBoost is based on gradient boosting algorithm and is mainly used in classification and regression problems. However, XGBoost has also been used for time series forecasting, with transformed data as input to a supervised problem. XGBoost makes predictions by combining multiple decision trees that work together to form a more accurate model.

Extreme Learning Machines(ELM) ELM is a neural network model that can be used for object recognition, feature learning, clustering, classification, regression, and time series modeling. ELMs are much faster to compute than traditional neural models because their weights and biases are assigned at random. They still capture the nonlinear characteristics of the data to predict the output.

Facebook Prophet Prophet is an open source library published by Facebook based on separable models (trend + seasonality + holidays). It allows users to make accurate time series predictions using intuitive parameters and support for custom seasonality and holiday effects.

5.3 Evaluation metrics

To effectively evaluate the forecasting model performance with time series data for this project, we use two measures: mean squared error (MSE) and symmetric mean absolute percentage error (sMAPE). They are respectively determined using the following formula:

$$MSE = \sum_{i=1}^{N} (y_i - \overline{y}_i)^2 \tag{1}$$

$$sMAPE = \frac{100\% \sum_{i=1}^{N} (\frac{2|\overline{y}_i - y_i|}{|y_i| + |\overline{y}_i|})}{N} \tag{2}$$

The sMAPE measure is added to determine the symmetric mean absolute percentage error of the forecasting result. With above formula, this is an accuracy measure based on percentage (or relative) errors. The lower the models for the assessment index with both MSE and sMAPE, the better the model and vice versa.

6 Experiments and results

6.1 Approaching method

In addition to predicting the total number of confirmed cases per day for 8 countries, we also predict the number of cases for each country, then add it up to get the final prediction result, to compare the performance of these two experimental ways on time series prediction algorithms. The experimental process is described in figure 1

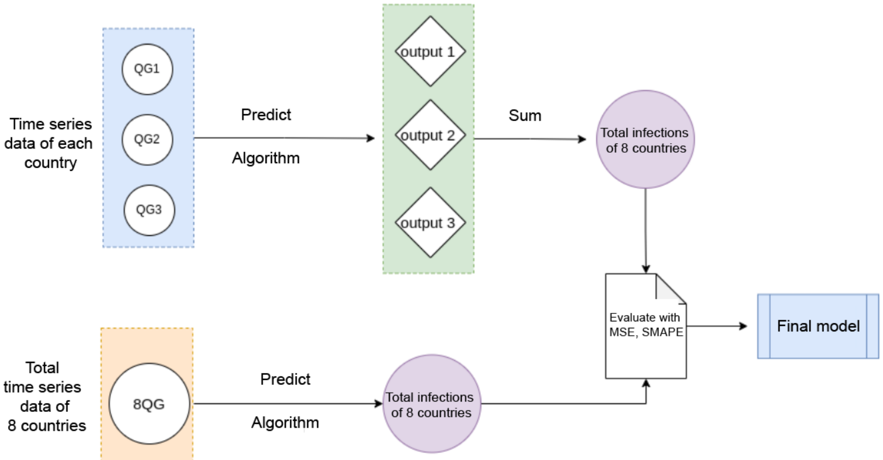
Time series data of each country
QG1
QG2
QG3
Predict
Algorithm
output 1
output 2
output 3
Sum
Total infections of 8 countries
Evaluate with MSE, SMAPE
Final model
Total time series data of 8 countries
8QG
Predict
Algorithm
Total infections of 8 countries

Fig. 1: Approaching method.

6.2 Model parameters

- **LSTMForecaster**: set the parameter past_seq_len=6 and train on 5 epochs.
- **TCNForecaster**: past_seq_len=10, future_seq_len=1, train on 1 epoch.
- **FBProphet**: train with weekly_seasonality=False, changepoint_prior_scale=1
- **Autoregression**: set lags=30.
- **XGBoost**: objective=’reg:squarederror’ và n_estimators=1000.
- **Extreme Learning Machines**: experiment with hidden_size=100, previous_time_steps=14, activation=’relu’.
- **LSTM_AutoTSEstimator**: model=’lstm’, search space=’Normal’, selected features=’auto’, and other default settings were used to train. past_seq_len and the number of epochs are adjusted depending on the data for each country.
- **TCN_AutoTSEstimator**: implemented similarly to LSTM with AutoTSEstimator but with model=’tcn’.

Table 2: Models results

Model	**Approach**	**MSE**	**SMAPE**
LSTM - AutoTS	Total	97.10^9	0.389
	Sum_8	98.10^9	0.248
TCN - AutoTS	Total	82.10^9	0.222
	Sum_8	144.10^9	0.284
PBProphet	Total	271.10^9	0.632
	Sum_8	272.10^9	0.630
TCNForecaster	Total	48.10^9	0.715
	Sum_8	48.10^9	0.719
LSTMForecaster	Total	46.10^9	0.737
	Sum_8	46.10^9	0.750
XGBoost	Total	441.10^9	0.412
	Sum_8	515.10^9	0.459
ELM	Total	33.10^9	0.137
	Sum_8	55.10^9	0.169
Auto Regression	Total	684.10^9	0.561
	Sum_8	664.10^9	0.544

Table 2 statistics the best results on each experimental model. In general, the predictive model does not produce very good results, especially MSE, which differs significantly from the ground truth. This can be explained by the fact that predicting the number of confirmed Covid cases is a difficult problem that is influenced by a variety of factors, not just timing. In general, the predictive model does not produce very good results, especially MSE, which differs significantly from the ground truth. This can be explained by the fact that predicting the number of confirmed Covid cases is a difficult problem that is influenced by a variety of factors, not just timing. When manually trained, LSTMForecaster and TCNForecaster produce better results than when deployed using AutoTS. However, regardless of the method of implementation, TCN and LSTM still have the potential for development on the dataset since their results are still superior to XGBoost, PBProphet, and AutoRegression. Furthermore, the experimental results reflect the success of training on each country's component data, which is then combined to forecast total confirmed cases, assisting in the improvement of model outcomes, particularly in the algorithms LSTM, TCN, and AutoRegression.

6.3 Result analysis

In this section, we visualize the forecasting results of the best model obtained after the experimental process, namely the testing results with the ELM model achieving on MSE and SMAPE are respectively above 33.10^9 and 13%.

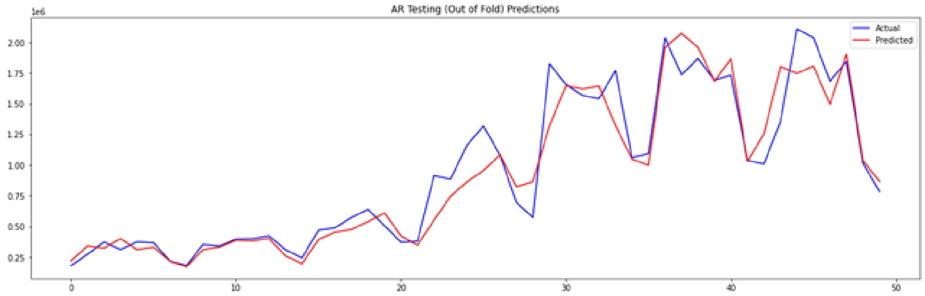

Fig. 2: The line chart shows the forecasting results of the ELM model.

We can observe from the figure 2 that the model grasps the data fairly well during the early stage, with just minor discrepancies. But more on that later, when the data gets more variable with more amplitude, the model's forecasting ability deteriorates.

7 Conclusion

At the end of the experiment, our group has built many predictive models but the results are not really good. In general, predicting the number of Covid-19 infections based on time series is not an easy problem. In fact, the spread of the disease depends on many other factors, and capturing the up-and-down trend of the graph of the number of infections is a very difficult task even when using more features. In addition, the amount of data used is not large, so the training of value chain analysis does not have too much experience.

In terms of future development, our team aims to develop more predictive models, combine more factors and apply on larger data, more countries. In order to find a more optimal solution in predicting the infection situation, thereby providing more information for the pandemic prevention process. It is also possible to gain more experience in predicting future pandemics.

References

[1] Fatoumata Dama and Christine Sinoquet. *Time Series Analysis and Modeling to Forecast: a Survey*. 2021. arXiv: 2104.00164 [cs.LG].

[2] Kayode Oshinubi et al. "Approach to COVID-19 time series data using deep learning and spectral analysis methods". In: 9 (Dec. 2021), pp. 1–21. DOI: 10.3934/bioeng.2022001.