Forecasting Myocardial Infarction occurrences in Chittagong Metropolitan Area: An application of the Time Series Model

Course Name: Data Analysis Project

Course Code: PM-ASDS07

Submitted to:

Department of Statistics

Jahangirnagar University

Savar, Dhaka

Submitted by:

Mohammad Saiduzzaman Sayed

ID: 20215063

Batch: 5th

Sec: A

Professional Masters in

Applied Statistics and Data Science (PM-ASDS)

JAHANGIRNAGAR UNIVERSITY

To

My Beloved Parents

and

Respected Teachers

CERTIFICATE OF APPROVAL

This is certified that Mohammad Saiduzzaman Sayed (ID: 20215063, Section: A and Batch: 5th), a student of the department of Statistics at Jahangirnagar University has completed the project work under my supervision.

His report entitled **"*Forecasting Myocardial Infarction occurrences in Chittagong Metropolitan Area: An application of the Time Series Model*"** is prepared after performing in the field work as requirement of obtaining a Masters' (PM-ASDS) degree.

He has completed the report by himself. He has been permitted to submit the report thereby.

Dr. Md. Abdus Salam
Professor
Department of Statistics
Jahangirnagar University

CONTENTS

List of Figures

Abstract

Myocardial Infarction (Heart Attack) is a grave public health issue for citizens of Chittagong Metropolitan Area in terms of both prevention and treatment, and its emergence has caused great suffering among residents. Early warning systems must be addressed in order to manage epidemics as soon as possible, given the increasing burden of disease in older years and the difficulties in assessing present and future demands.

The goal of the study was to identify the best suitable model for heart attack incidence and, using that model, forecast the prevalence of heart attacks in the Chittagong Metropolitan Area in the future.

This study considered a supplementary data set of daily heart attack incidence from January 1, 2020, to December 31, 2021. With the aid of the chosen model selection criteria, we first identified the most appropriate model from the Autoregressive Integrated Moving Average (ARIMA), Error, Trend, Seasonal (ETS), and Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal (TBATS), and Long-Short Time Memory (LSTM) models, and then we used the chosen model to predict the occurrence of heart attacks in the Chittagong Metropolitan Area. The LSTM model outperforms the ARIMA, ETS, TBATS models in terms of performance.

It is crucial to comprehend the likely future course of heart attack occurrence before developing a thorough plan for future battling measures. According to the research, heart attacks are predicted to rise in frequency in the next years.

Keyword: Heart Attack Cases, ARIMA, ETS, LSTM, TBATS, Time series forecasting

Chapter 1

Introduction

In Chittagong, Heart-Attack is a significant contributor to disease and fatalities. Even during the COVID-19 pandemic, record patients were admitted to hospitals and deaths due to heart attacks. We collected data on daily heart attacks that happened from January 1, 2020, through December 31, 2021, and were recorded at the hospitals in Chittagong City Corporation in order to forecast the number of instances.

Several types of research have been conducted on the topic of heart attack outbreak prediction. I seldom come across someone who forecasts heart attack cases in Chittagong Metropolitan Area using time series data. Chittagong is the second largest city and only seaport in Bangladesh. The increasing rate of Myocardial Infarction disease in Chittagong is the concern of the study.

Statistical modeling is one strategy that may be used to anticipate heart attack outbreaks. The discipline of epidemiologic research made great use of the time series approach. We applied Autoregressive Integrated Moving Average (ARIMA), Decomposition method Seasonal and Trend Decomposition (STL) with Exponential Smoothing (ETS) model, Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal (TBATS), and deep learning-based time series model Long Short-Term Memory networks (LSTM) to establish statistical forecasting models.

The primary goal of this research is to create the best time series model for projecting heart attack cases in the Chittagong Metropolitan Area by utilizing daily data from 1st January 2020 to 31st December 2021.

Chapter 2

Methodology

2.1 Data Source and Variable Descriptions

Data Source:

The secondary information on the amount of heart attack patients was taken from various hospitals in the Chittagong metropolitan area between January 1, 2020, and December 31, 2021. The data's format is suitable for time series analysis.

Study design for collecting the sample:

We are using "one-stage cluster sampling" to select hospitals. The hospitals are "Primary Sampling Units"- the clusters. Five hospitals were chosen at random from a list of hospitals in Chittagong. Then, between 1 January 2020 and 31 December 2021, we collected data from heart attack cases at the chosen hospitals.

Dataset Details:

Description: Daily heart attack cases admitted in hospitals. The dataset contains 731 observations and heart attack cases is our study feature. The time range is 1st Jan 2020 to 31st Dec 2021.

Format: A daily time series data

Cluster: 5 group of clusters in Chittagong City Corporation area namely

1. CSCR
2. Ibn Sina
3. Parview
4. Surgiscope
5. Metro

Flow of process:

This study investigated using time series model to forecast the number of reported heart attack cases.

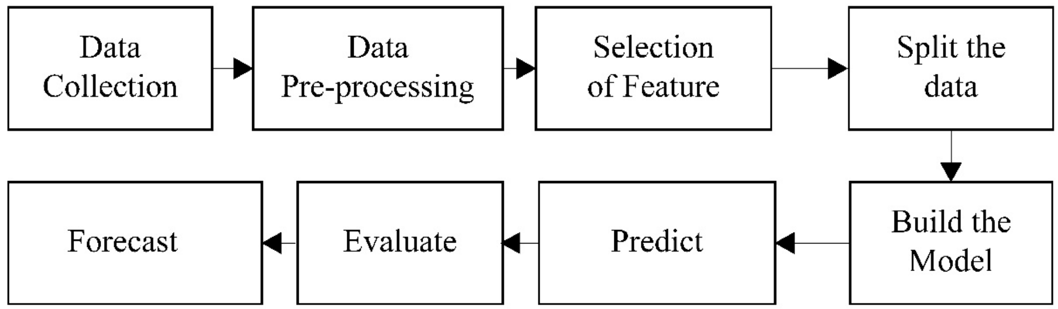

Fig 1: Flow of process for prediction of heart attack

Data Collection:

The information was gathered by using an electronic patient management tool to retroactively retrieve administrative data. The Patient Manager is the Patient Administrative System for time series data that is most frequently utilized. Medical professionals from many locations within the same health service can access a centralized database using PM to securely track all administrative components of an inpatient session.

Data Pre-processing:

Before it can be used in the model, the dataset is pre-processed. The first procedure deals with missing data in the dataset. And estimate or remove outlier and high influential points. Then normalizing data to become unit free.

Selection of feature and splitting:

Our dataset has one feature which is heart attack Cases. We are using this feature to predict future. Now, we split our data into the training set and test set. We build the model using training data and check accuracy with test data.

Summarized Workflow:

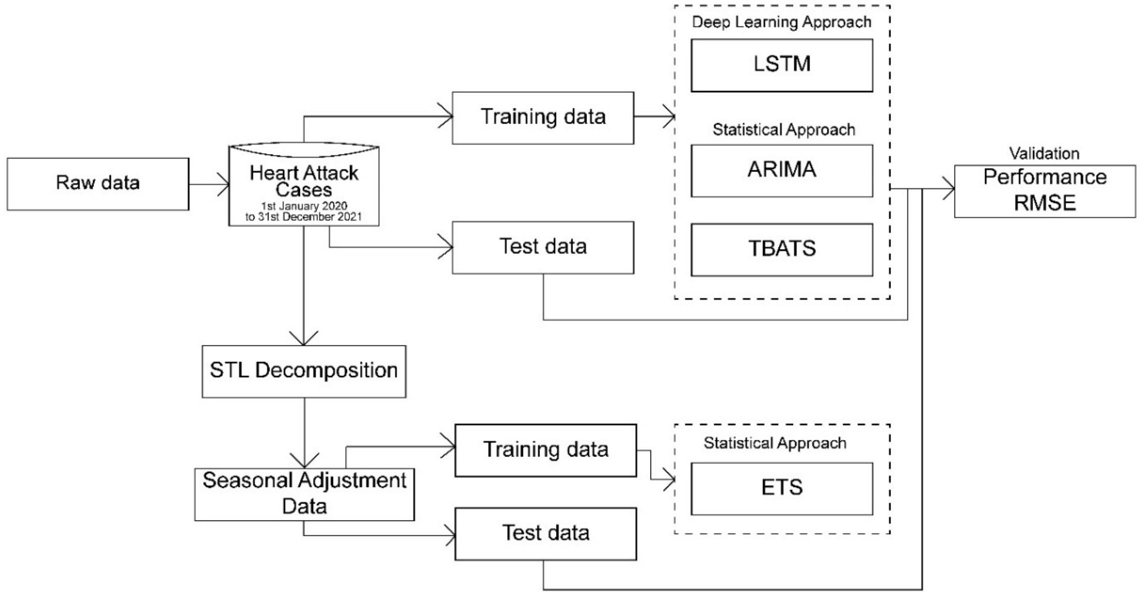

Fig 2: Summarized Workflow

2.2 Statistical Models

Forecasting Scenario for Time Series data:

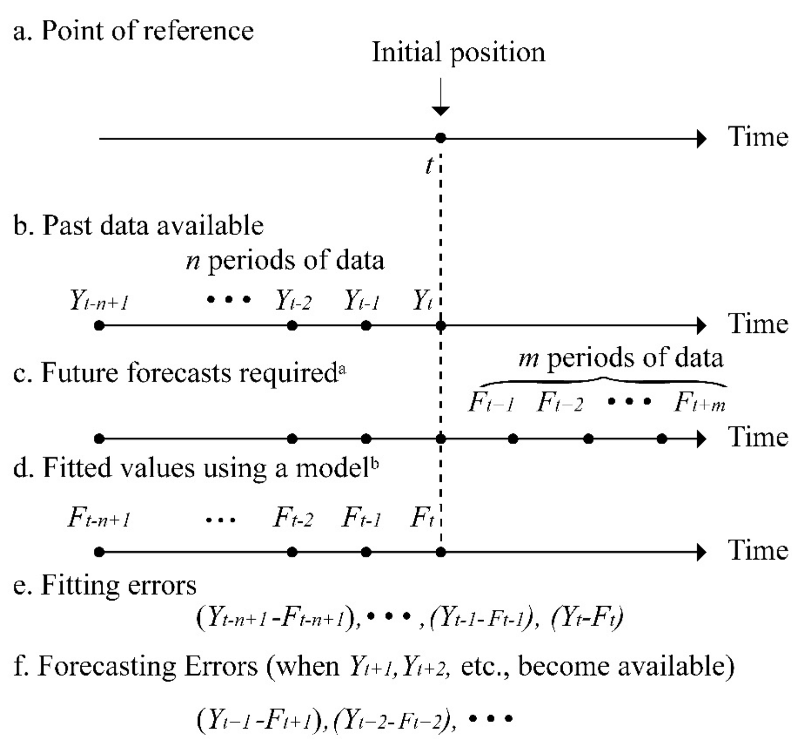

Fig 3: Forecasting Scenario

(i) Forecast with Autoregressive Integrated Moving-average (ARIMA):

This model is one of the statistical tools used to forecast and analyze time series data. Seasonal impacts are frequently seen in various time series data.

ARIMA (p, d, q)

$$\left(1 - \emptyset_1 B - \cdots - \emptyset_p B^p\right)(1 - B)^d y_t = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t$$

Intercept Form

$$\left(1 - \emptyset_1 B - \cdots - \emptyset_p B^p\right)y_t' = c + \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t$$

Mean Form

$$\left(1 - \emptyset_1 B - \cdots - \emptyset_p B^p\right)(y_t' - \mu) = c + \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t$$

- $y_t' = (1 - B)^d y_t$
- μ is the mean of y_t'
- $c = \mu(1 - \emptyset_1 - \cdots - \emptyset_p)$

The Box-Jenkins Methodology for ARIMA model:

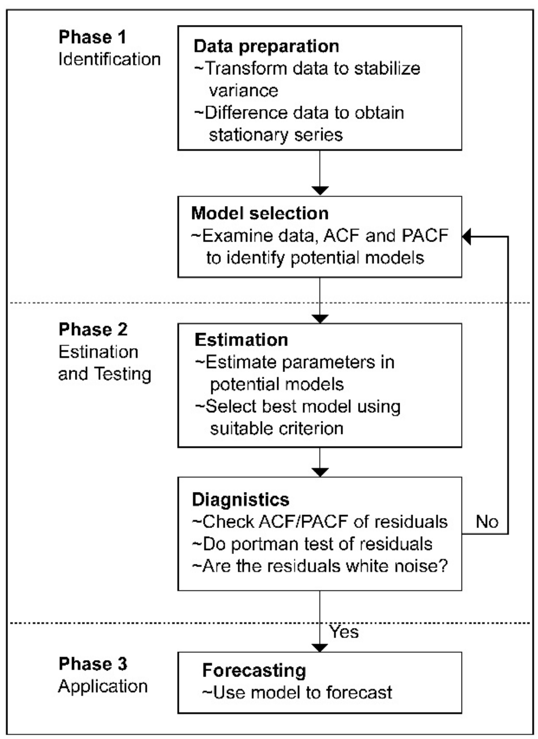

Fig 4: Schematic representation of the Box-Jenkins methodology

(ii) Forecasting with Seasonal adjustment: Seasonal and Trend decomposition using Loess and Exponential Smoothing Method

STL Decomposition:

A flexible and reliable way for breaking down time series is STL. Loess is a technique for estimating nonlinear connections, and STL stands for "Seasonal and Trend decomposition using Loess."

The three STL analysis components have the following relationships with the original time series:

$$y_t = S_t + T_t + R_t$$

$where,$

y_t = data at period t

S_t = seasonal component at period t

T_t = trend-cycle component at period t

R_t = remainder component at period t

Exponential Smoothing:

The ETS model is a method for univariate time series forecasting that focuses on trends as well as seasonal elements. This model may construct trending and nonstationary seasonal components for a variety of attributes. Error, Trend, and Seasonal are the three parameters of this model. It contains four values for each parameter: A for Additive, Ad for Additive Damped for Trend, M for Multiplicative, N for None.

Hence, ETS (A, M, N) means Additive for Error, Multiplicative for Trend, and None for Smoothening.

Forecasting evolution:

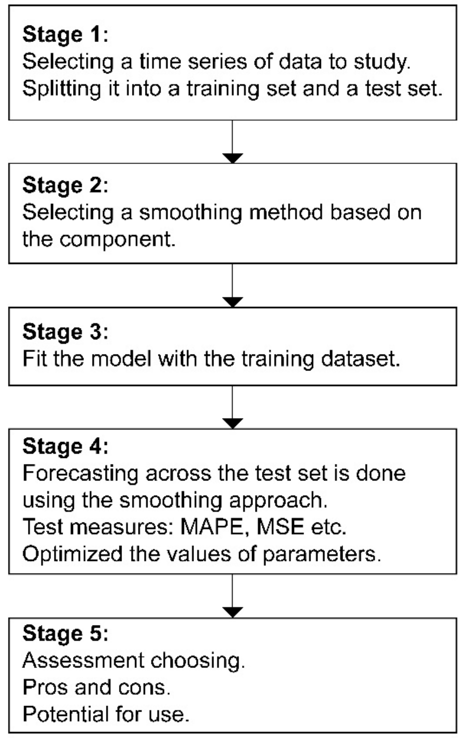

Fig 5: Forecasting Evaluation

(iii) Forecast with Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal (TBATS) models:

In contrast to dynamic harmonic regression, TBATS models let seasonality to gradually alter over time, whereas harmonic regression terms need seasonal patterns to repeat regularly without changing. However, TBATS models have the disadvantage of being slow to estimate, especially for lengthy time series.

$$y_t^{(\lambda)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = \phi b_{t-1} + \beta d_t$$

$$d_t = \sum_{i=1}^{p} \varphi_i \, d_{t-i} + \sum_{i=1}^{q} \theta_i \, e_{t-i} + e_t$$

Seasonal part:

$$s_t^{(i)} = \sum_{j=1}^{(k_i)} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos(w_i) + s_{j,t-1}^{*(i)} \sin(w_i) + \gamma_1^{(i)} d_t$$

$$s_{j,t-1}^{*(i)} = -s_{j,t-1}^{(i)} \sin(w_i) + s_{j,t-1}^{*(i)} \cos(w_i) + \gamma_2^{(i)} d_t$$

$$w_i = 2\pi j / m_i$$

$where,$

Time series at moment t: $y_t^{(\lambda)}$	Seasonal Component: $s_t^{(i)}$
Level: l_t	Trend with damping: b_t
White Noise: e_t	ARMA(p,q) process for residuals: d_t
Amount of seasonality: T	Length of seasonal period: m_i
Box-Cox transformation: λ	Amount of harmonic seasonal period: k_i
Smoothing: α, β	Trend damping: ϕ
ARMA(p,q) coefficients: φ_i, θ_i	Seasonal smoothing: $\lambda_1^{(i)}$, $\lambda_2^{(i)}$

(iv) Forecast with Long short-term memory:

The LSTM network model is an enhancement to the RNN architecture that permits bigger reference windows for sequence prediction. An LSTM layer consists of recurrently connected memory blocks with a larger capacity than RNNs. This is achieved by including a forget gate into an LSTM cell, allowing the model to keep just the relevant prior information.

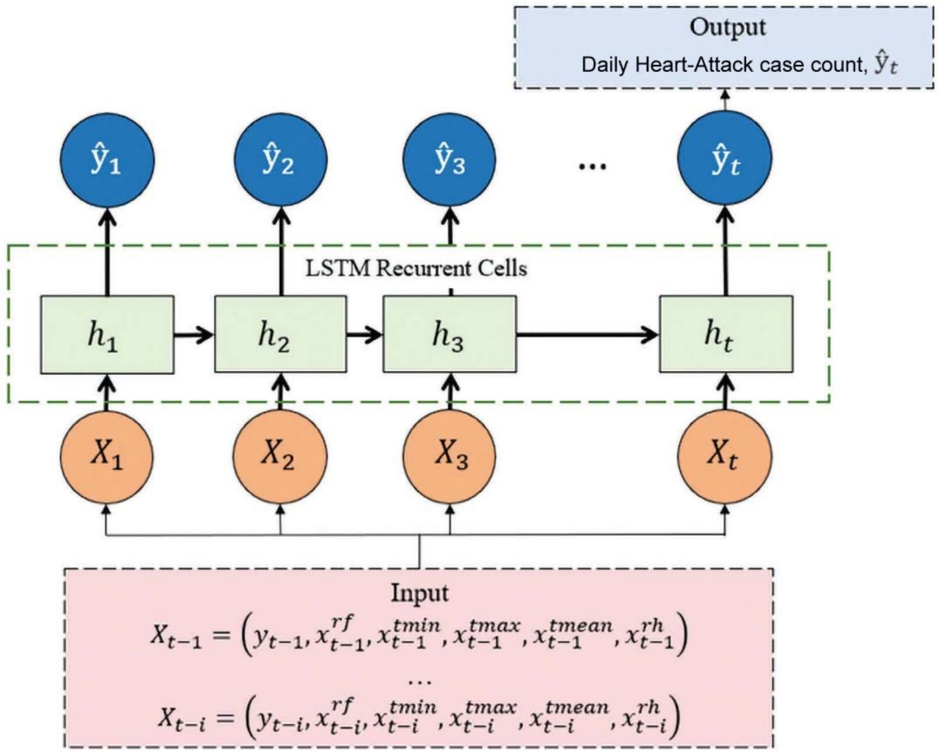

Fig 6: The architecture for forecasting using the LSTM network

Model Selection Criteria:

The Akaike's Information Criterion (AIC), Akaike's Information Criterion correction (AICc), and Bayesian Information Criterion (BIC) were employed here since it is preferable to choose the proper model before predicting.

Akaike Information Criterion:

$$AIC(\beta) = -2\ln L_X(\beta, S_X(\beta)/n) + 2(p + q + 1)$$

Akaike Information Criterion corrected:

$$AICc(\beta) = -2\ln L_X(\beta, S_X(\beta)/n) + \frac{2(p + q + 1)n}{(n - p - q - 2)}$$

Bayesian Information Criterion:

$$BIC = (n - p - q)\ln\left[\frac{n\hat{\sigma}^2}{(n - p - q)}\right] + n\left(1 + ln\sqrt{2\pi}\right) + (p + q)ln\left[\frac{\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}^2}{p + q}\right]$$

Forecasting Accuracy Measures:

Additionally, the Mean Error (ME), Mean Square Error (MSE), Root Mean Square Error (RMSE) are used to assess the fitted model's correctness.

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

$$MPE = \frac{1}{n}\sum_{t=1}^{n} PE_t$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} |PE_t|$$

$$PE_t = \left(\frac{Y_t - F_t}{Y_t}\right) * 100$$

One way to evaluate forecasting model by calculating the commonly used forecasting accuracy measurement is Root Mean Squared Error (RMSE). The equation of RMSE shown below, where n is the total of data, y_t is actual value, and y_t' is predicted value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_t - y_t')^2}{n}}$$

2.3 Compactional Tools:

This whole project computing is done in RStudio and using R programming language. The packages we used in the project are TensorFlow, Keras, ggplot, forecast, itsmr, tseries etc.

Chapter 3

Data Analysis and Result

Analysis and Result for ARIMA model:

Time plot:

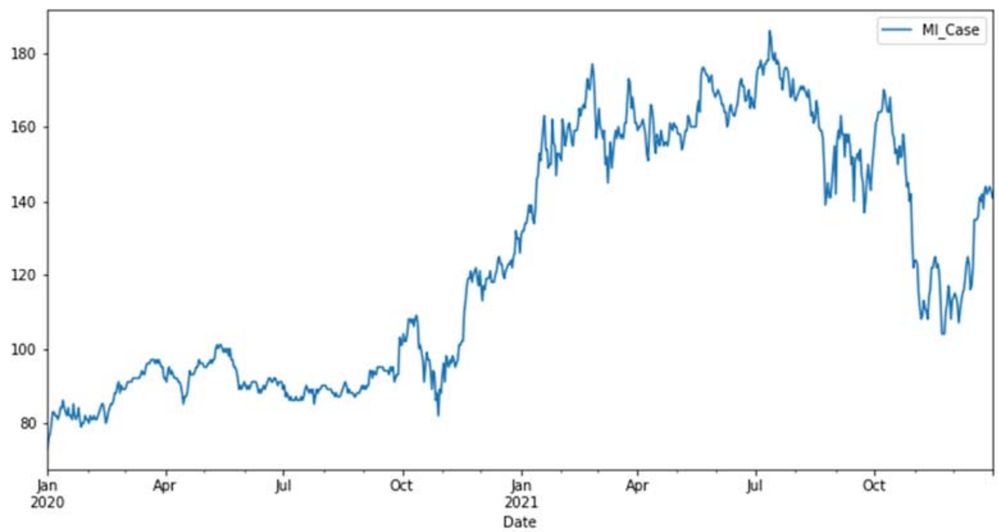

Fig 7: Time Plot

Comment on Time-plot:

The above time-plot describes some main features of data which is given below:

Trends: The above time series plot shows a clear upward damped trend.

Seasonal pattern: These data show an additive seasonal pattern because the data is daily data. The frequency is 365.35, and therefore period is daily, so a seasonal component is present.

Sharp Changes: The above time series plot shows that there is no sharp change in our data.

Outliers: Following the above time series plot, we can assume that there are no outliers in the data. We can check outliers with the help of forecast package and tsoutliers function in R.

ACF Plot:

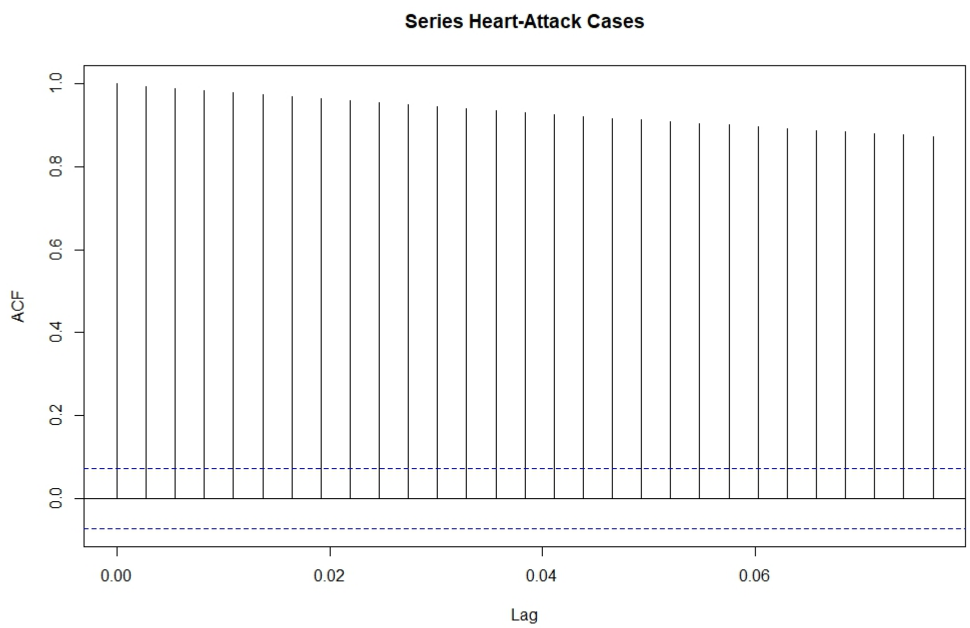

Fig 8: ACF Plot

Comment on ACF plot:

ACF plot shows that trend and seasonality are present in the data. The spikes are significant, and the coefficient is high in each lag. Hence, the clear view of ACF shows that this data set needs to make stationary to fit the model.

Checking stationary:

After estimate outliers, now we are checking the data is stationary or not. Augmented Dickey-Fuller (ADF) test and The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test can be used to test for stationary.

In the ADF test,

H_0 : the series is not stationary

H_a : the series is stationary

p-value: 0.7518

Hence, a small p-value (i.e., less than Alpha=0.05) suggests that the series is stationary. Here, the p-value is greater than alpha (0.05). So, we cannot reject H_0. That means the ADF test shows that our dataset is not stationary.

In the KPSS test,

H_0 : the series is stationary

H_a : the series is not stationary

p-value: 0.01

Hence, a small p-value suggests that the series is not stationary, and a differencing is required. Here, the p-value is less than Alpha (0.05). So, we can reject H_0. That means the KPSS test shows that our dataset is not stationary.

So, here the seasonal differences are required. Now, we can check stationary after taking 1^{st} differences.

After taking 1st difference the ADF test shows that the p-value (0.01) is less than Alpha (0.05) means the data is stationary and the KPSS test shows that the p-value (0.1) is greater than Alpha (0.05) means the data is stationary. So, we can say that one nonseasonal difference is enough for the ARIMA model for our dataset to get stationary series.

Time series display plot:

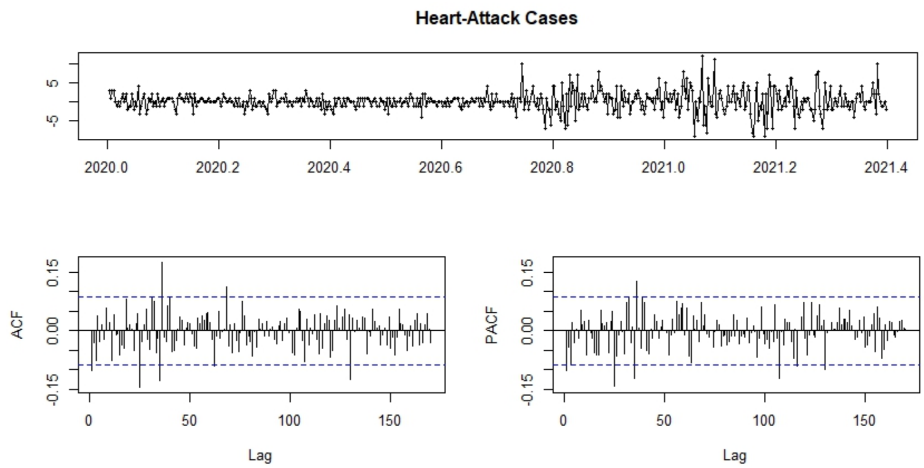

Fig 9: Time Series Display Plot

Comment on time series display plot:

After taking one nonseasonal difference the time series display plot shows that the trend is not present in the time plot. Also, ACF and PACF plot shows that there is only one significant spike in both plots.

ACF plot for stationary series:

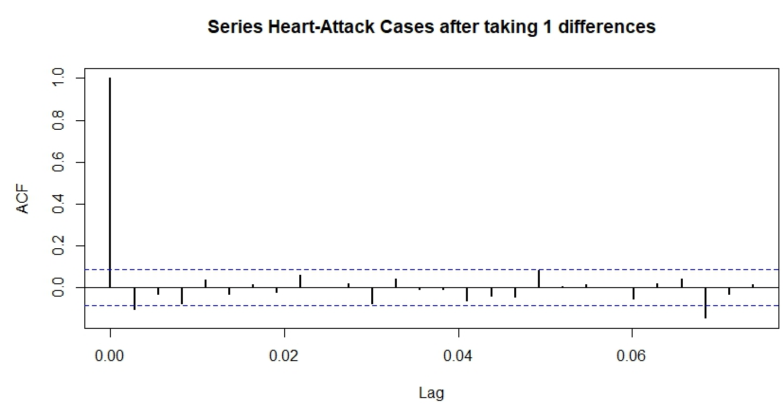

Fig 10: ACF Plot of stationary series

Comment on ACF plot of stationary series:

ACF plot shows that only the first one is a significant spike. So, here we can say the MA(q)=1.

PACF plot for stationary series:

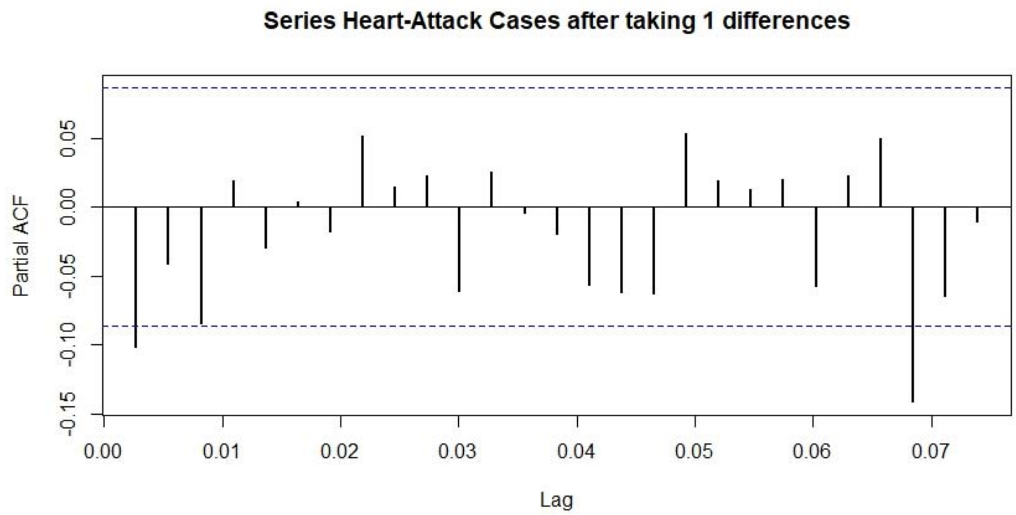

Fig 11: PACF Plot of stationary series

Comment on PACF plot of stationary series:

PACF plot shows that only the first one is a significant spike. So, here we can say the AR(p)=1.

Fit and Mathematical Expression of the ARIMA model:

Now, we fit the model with order (p=1, d=1, q=1).

ARIMA (1,1,1) with drift

Estimated Model:

$(1 - 0.4747B)(1 - B)^1 X_t = (1 - 0.5855B)Z_t, \qquad \{Z_t\} \sim WN\ (0, \hat{\sigma}^2 = 6.809)$

$X_t - 0.4747BX_t - BX_t + 0.4747B^2X_t = Z_t - 0.5855BZ_t, \{Z_t\} \sim WN\ (0, \hat{\sigma}^2 = 6.809)$

$and\ drift = 0.1928$

Residuals Diagnostics Checking:

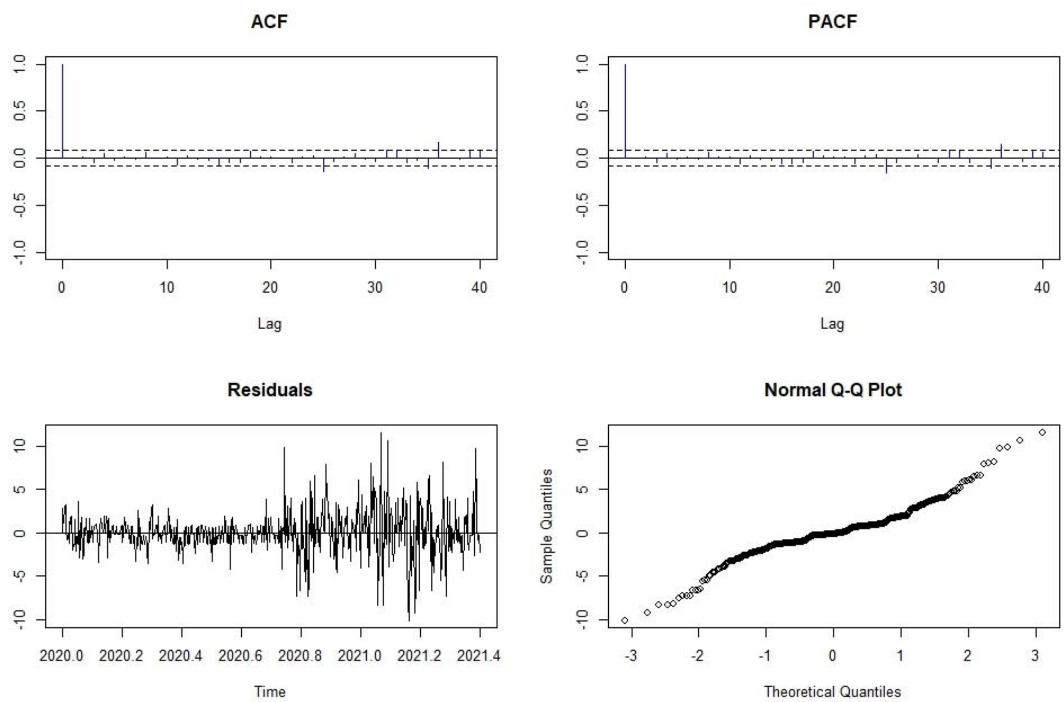

Fig 12: Diagnostics Plots of ARIMA model

Assumptions:

i. {et} uncorrelated. Here, we can see the ACF and PACF plot of residuals diagnostics, there is no significant spike. So, we can say that the {et} is uncorrelated.

ii. {et} have mean zero. From our Residuals plot, we can say that mean zero and variance is constant.

iii. {et} is normally distributed. Here, the Normal Q-Q plot shows that the residuals are approximately normally distributed because the data is near 45 degrees.

Here, H_0: Residuals are iid noise.

We can see the summary of the residuals test here all p-value is greater than Alpha (0.05). So, here we cannot reject H_0. So, the M1 model satisfied all assumptions of residuals.

Analysis and Result for STL decomposition and ETS model:

STL Decomposition:

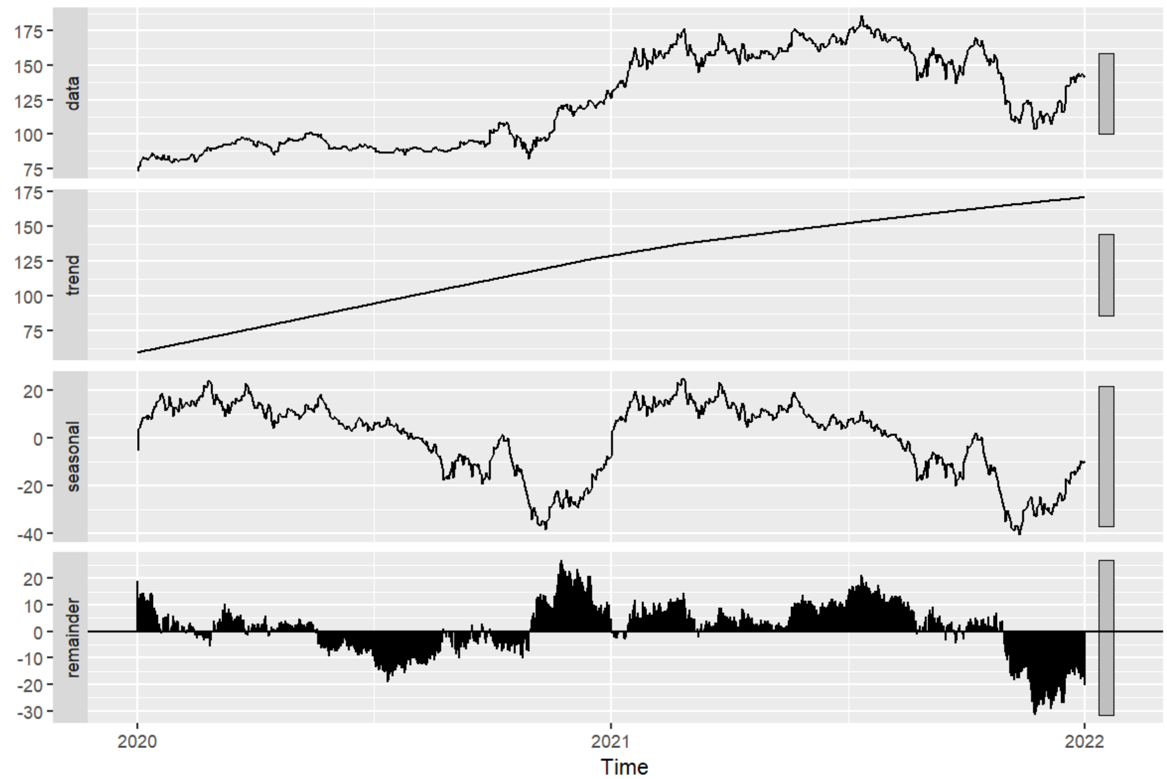

Fig 13: STL Decomposition Plot

Comment on STL Decomposition Plot:

Therefore, the STL decomposition is used to first separate the trend and seasonal components. After that we get seasonal adjustment data. Now, we are using seasonally adjusted data to fit the ETS model.

ETS model:

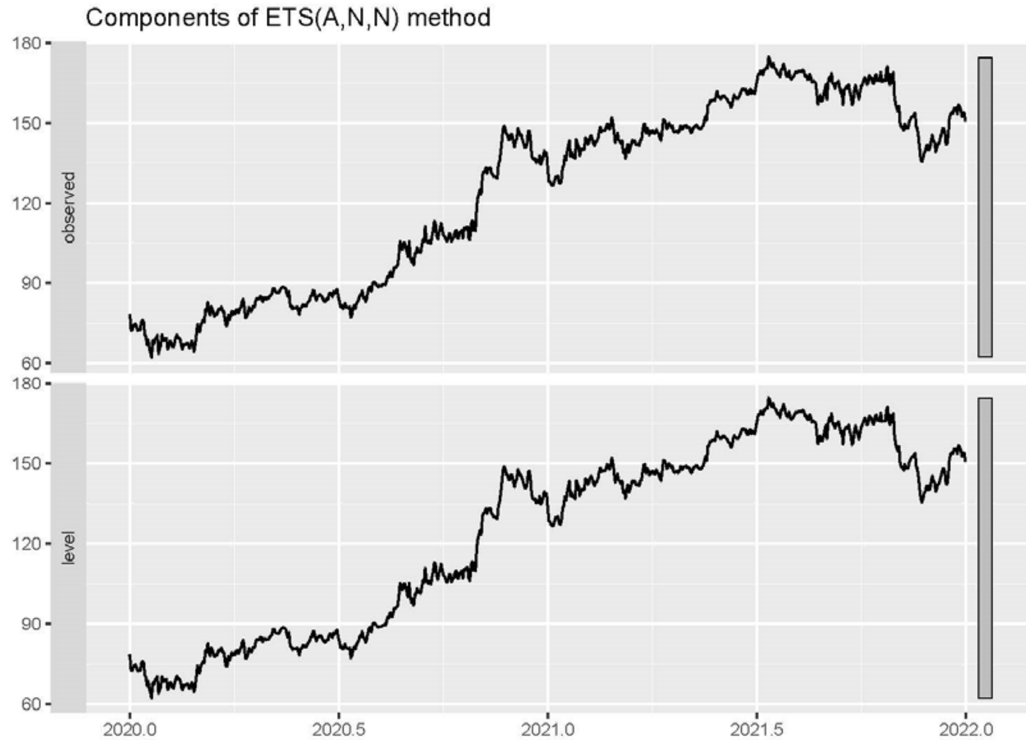

Fig 14: Observed Plots of ETS model

ETS (A, N, N): Additive Error, No Trend and No Seasonality

Model: Simple Exponential Smoothing

Mathematical Expression:

Measurement equation: $y_t = l_{t-1} + \varepsilon_t$

$$= 78.0702_{t-1} + \varepsilon_t$$

State equation: $l_t = l_{t-1} + \alpha\varepsilon_t$

$$= 78.0702_{t-1} + 0.9474\varepsilon_t$$

$$where\ \varepsilon_t \sim N(0, \sigma^2 = 4.464)$$

Diagnostic Plots:

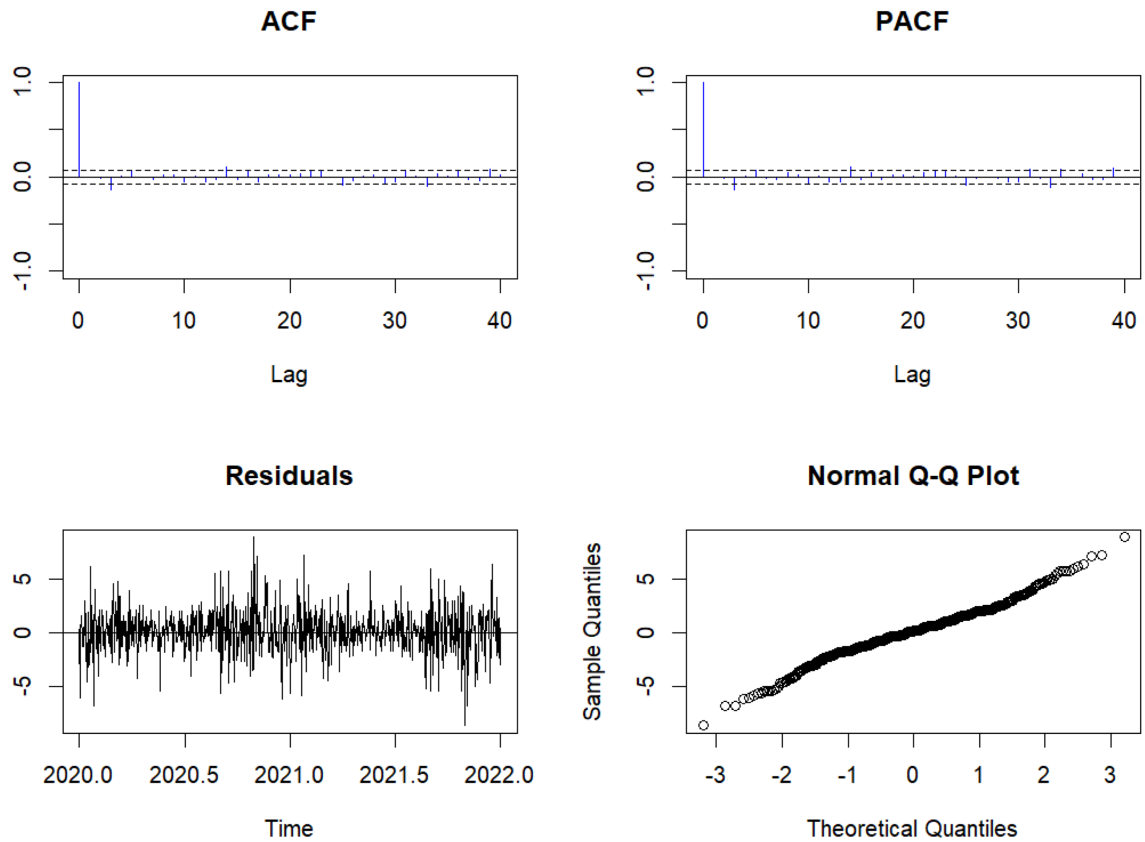

Fig 15: Diagnostic Plots of ETS model

Assumptions:

i. {et} uncorrelated. Here, we can see the ACF and PACF plot of residuals diagnostics, there is no significant spike. So, we can say that the {et} is uncorrelated.

ii. {et} have mean zero. From our Residuals plot, we can say that mean zero and variance is constant.

iii. {et} is normally distributed. Here, the Normal Q-Q plot shows that the residuals are approximately normally distributed because the data is near 45 degrees.

Here, H_0: Residuals are iid noise.

We can see the summary of the residuals test here all p-value is greater than Alpha (0.05). So, here we cannot reject H_0. So, the M1 model satisfied all assumptions of residuals.

Analysis and Result for TBATS model:

The time series with numerous seasonal periods may be predicted using TBATS. Daily data, for instance, can include a weekly trend in addition to a yearly pattern. A daily pattern, a weekly pattern, and a yearly pattern can all be present in hourly data.

The original time series is transformed using the Box-Cox method in TBATS, and this is then modelled as a linear mixture of an exponentially smoothed trend, a seasonal component, and an ARMA component. Through the use of Fourier series, trigonometric functions are used to represent the seasonal components. TBATS use AIC for certain hyper-parameter adjustment.

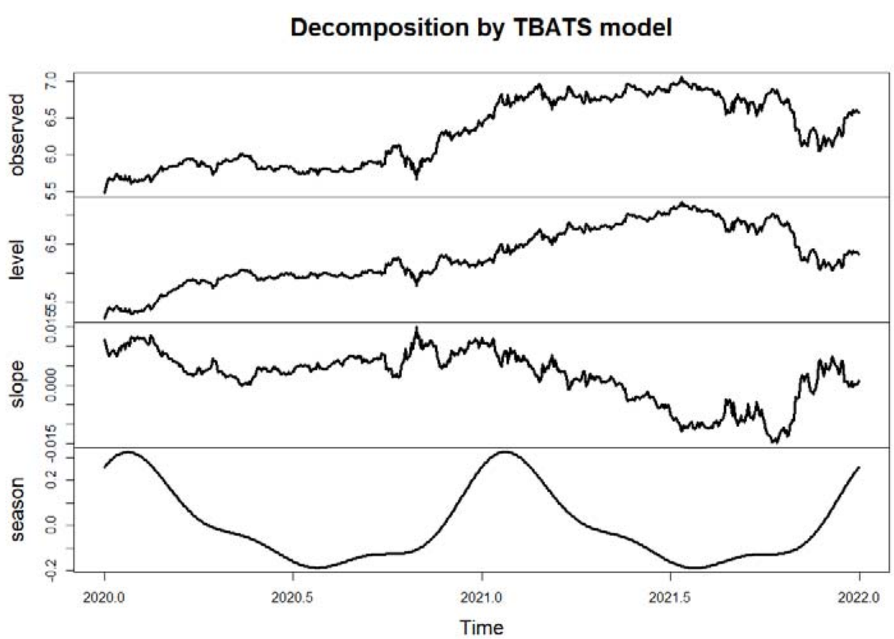

Fig 16: Decomposition by TBATS model

Comment on Decomposition by TBATS model:

Therefore, the TBATS decomposition is used to first separate the daily and seasonal components. Fig. displays a typical outcome of the disintegration of. The seasonal and daily patterns have been correctly detected, and the level term still contains the majority of the noise, as can be shown. The level phrase also makes damage information very apparent.

Analysis for LSTM model:

The LSTM-RNN-based architecture for forecasting heart-attack cases is formed of a hidden layer called the LSTM. The layer contains 64 memory cells for computing results, and the values obtained are transferred to the next iteration. Since the data is available for a period of one year, the time step of LSTM is set to a value of 12. The activation function used in the proposed methodology is Rectified Linear Unit (ReLU). The performance of ReLU is comparatively good that the other activation functions like sigmoid and tanh.

Summary of Model:

Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 12, 150)	91200
lstm_1 (LSTM)	(None, 64)	55040
dense (Dense)	(None, 64)	4160
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 1)	65

Total params: 150,465
Trainable params: 150,465
Non-trainable params: 0

Layers details:

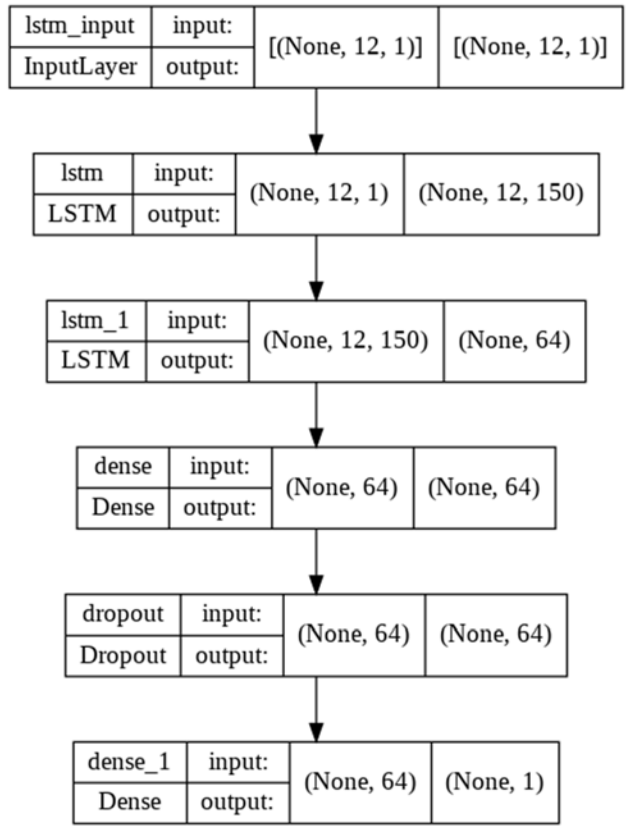

Fig 17: Layers of LSTM model

Model Selection Criteria:

Model	**AIC**	**AICc**	**BIC**
ARIMA	2435.41	2435.49	2452.36
STL+ETS	5918.14	5918.18	5931.93
TBATS	6361.85	6361.97	6376.33

Here, ARIMA is the best fitted model because its AIC and AICc value is lowest. And ARIMA model fulfilled more underlying assumptions of residuals than others.

Forecasting accuracy measures based on the test data set:

Model	MSE	RMSE	MAE	MPE	MAPE
ARIMA	2894.691	53.8	43.094	32.62	32.65
STL+ETS	2577.53	42.98	39.17	27.54	32.98
TBATS	1986.654	38.790	39.9	13,537	33.877
LSTM	1160.89	34.0718	32.954	1.6477	32.468

After checking forecasting accuracy measures for the test data set, here the LSTM model forecasting accuracy measures values are lower than other models. So, based on that LSTM model is best fitted model for forecasting.

Forecasting:

The below plot shows forecasting 1 year ahead of dengue cases using previous records with the best model LSTM.

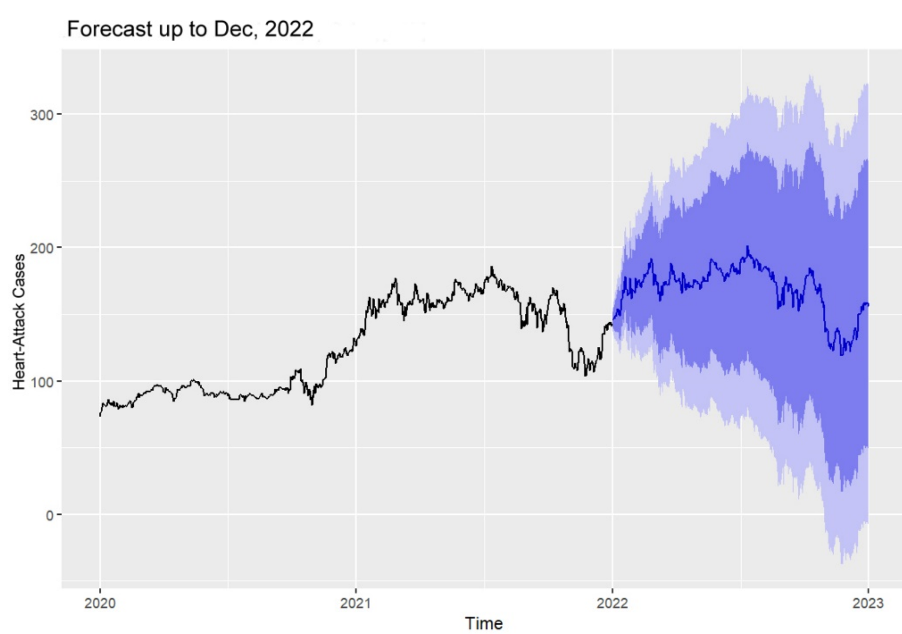

Fig 18: Forecasting Heart-Attack Cases up to Dec,2022

Chapter 4

Discussion & Conclusions

This study set aimed to identify the pattern of heart attack outbreaks and provide a short-term forecast. Despite the fact that before to now, the incidence of dengue was not given any concern. The epidemic, however, poses possible threats to residents of Chittagong. As the system dynamics model does not display the seasonality-periodic fluctuations for research, a statistical approach like the Box-Jenkins technique is often one of the preferred ways for predicting. These models do, however, have certain feedback process restrictions. If we see the AIC, AICc, and BIC value of the statistical approaches, the Arima model is best for forecasting future total cases of heart attack daily. But here we consider the RMSE value of all approaches, the LSTM model is best for forecasting future total cases of heart attack daily.

The study's limitation is that it solely takes into account heart attack incidents in Chittagong. It is thought that factors such as blood vessel spasms, drug abuse, hypoxia, low blood oxygen levels, and other factors may have an impact on the frequency of heart attacks. Future research has a lot of room to refine and validate the model in order to get more accurate results.

People and governments are pushed to consider heart-attack prevention measures as a result of the heart-attack epidemic's escalating occurrence. A comprehensive strategy for future battling methods must also consider the likelihood of heart attacks in the future. According to the outcomes of the model selection criteria taken into account in this study, LSTM is the model that is best suited for fitting heart attack data and predicting the likelihood of future heart attacks in the Chittagong Metropolitan Area.

References

Naher,S., Rabbi,F., Hossain,M.M., Banik,R., Pervez,S., Boitchi,A.B. (2022). Forecasting the incidence of dengue in Bangladesh—Application of time series model. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9178403/

Olah, C. (2015). Understanding LSTM Networks. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs

Nadda,W., Boonchien,W., Boonchieng,E. (2022). Influenza, dengue and common cold detection using LSTM with fully connected neural network and keywords selection. Retrieved from https://biodatamining.biomedcentral.com/articles/10.1186/s13040-022-00288-9

Badr,A., Makarovskikh,T., Mishra,P., Abotaleb,M., Khatib,A.M.G.A., Karakaya,K., Redjala,S., Dubey,A., Attal,E. (2021). Modelling and forecasting of web traffic using Holt's linear, bats and TBATS models. Retrieved from http://scik.org/index.php/jmcs/article/view/5400

Makridakis,S.G., Wheelwright,S.C., Hyndman,R.J. (1997). Forecasting: Methods and Applications. Retrieved from https://www.pdfdrive.com/forecasting-methods-and-applications-e188346908.html

Acknowledgements

First and foremost, I would like to express my sincere appreciation to Professor Dr. Md. Abdus Salam, who is both my supervisor and the course coordinator for the PM-ASDS program. His knowledge was crucial in developing the teaching strategy and methodology. I was inspired by your insightful comments to sharpen my reasoning and increase the caliber of my work.

My appreciation to each of my lecturers is inexpressible for their patience and insightful remarks. Furthermore, this project would not have been feasible without the kind assistance of the PM-ASDS program committee.

I should not forget to acknowledge my family, especially my parents, as a last point. Their confidence in me has sustained my enthusiasm and upbeat attitude throughout this process. Additionally, I want to thank my friends for providing me with so much amusement and emotional support.

THE END