Traffic Forecasting using Time-Series Analysis

Conference Paper · January 2021

DOI: 10.1109/ICICT50816.2021.9358682

5 authors, including:

Muhtadi Zubair
BRAC University
1 PUBLICATION **10** CITATIONS
SEE PROFILE

Sarowar Hossain
BRAC University
1 PUBLICATION **10** CITATIONS
SEE PROFILE

Muhammad Iqbal Hossain
BRAC University
60 PUBLICATIONS **156** CITATIONS
SEE PROFILE

Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021]
DVD Part Number: CFP21F70-DVD; ISBN: 978-1-7281-8500-2

Traffic Forecasting using Time-Series Analysis

Mohammad Asifur Rahman Shuvo, Muhtadi Zubair,
Afsara Tahsin Purnota, Sarowar Hossain, Muhammad Iqbal Hossain
Department of Computer Science And Engineering
Brac University
Dhaka, Bangladesh
shuvoasif28@gmail.com, muhtadiz23@gmail.com,
afsara.purnata@gmail.com, mishu.sarowar@gmail.com, iqbal.hossain@bracu.ac.bd

***Abstract*—Traffic jams are a common phenomenon all over the world, especially in a densely populated country like Bangladesh. Due to this, people try heart and soul to tackle this problem by any means necessary to save time to reach their desired destination. Hence the traffic related research is a hot topic now a days which will be quite beneficial for all people living in congested cities. We also tried to do some research on the traffic network to find the most suitable traffic forecasting model to forecast or predict the future traffic value using time-series forecasting models. The only topic which deals with both, traffic prediction and traffic control is traffic time-series analysis for which it is essential. In this paper, we have obtained a suitable dataset containing data of the number of various vehicles for each hour for seven days straight. We have used this dataset to feed into a few time-series forecasting models of our choosing. The models or algorithms considered are ARIMA, ETS, SNAIVE, PROPHET and the last one is the combination of all models we named it "mix". The study shows us the significant difference between each of the models and which one produces a more reliable and accurate prediction.**

***Index Terms*—Traffic Forecasting, Time-series forecasting models, 'Auto Regressive Integrated Moving Average, Seasonal Naive, Exponential Smoothing, Prophet**

I. Introduction

We all know that traffic jams are a significant issue for a big population country like Bangladesh. A traffic jam refers to a situation in a road when the bus, truck, rickshaw, and other vehicles are stuck for an extended period. Sometimes vehicles move very slowly because of the traffic jam. In the whole world, there are different countries that face traffic every day. However, our country has a large population. This is a preeminent problem for Dhaka city.

We decided to use time-series analysis for making Traffic forecasts and also choosing a model that will be quite beneficial for all people living in congested cities as it helps them get an overview of the traffic hours or a day before they plan to go to a particular place.[6] The analysis of traffic time series is very important to predict the travel time and usually, it is also using a traffic control system. This is a system that operates with time-series data. The meaning of the time series is, in a fixed time of period data would generate in a series.[13] Our aim is to the prediction of future traffic that will help us to know the traveling time. On the other hand, it impacts on our financial growth also. For example, if we forecast the electricity bill for the next month, then we will easily assume that how we manage our usage of electricity, it also helps us to save money. So that we want to decide that we will follow the time series analysis and forecasting, which fits our prediction model perfectly.

For this purpose, we are working to find out the prediction of the next 24hours traffic. We are trying to use four types of time series models, which are to Auto Regressive Integrated Moving Average (ARIMA), Seasonal Naive (SNAIVE), Exponential Smoothing (ETS) and Prophet.

To implement our idea at first, we collected the historical data of traffic flow. Next, we applied some of the time-series analysis techniques and machine learning models to acquire our desired results. We have predicted the next 24 hours of traffic.

II. Background

Time series analysis may be a statistical procedure that analyses and manipulates statistic data. It is made of collected data points at constant time intervals.[1] As we know, a collection of data at regular intervals is time-series. Another technique that adds a bit more complexity but elegance to the system is forecasting. A technique that determines predictive estimates in determining future trends using historical data as input is called forecasting.[2] Now, this gets interesting when we combine this with time series. It is a machine learning technique. Here we are using historical time series data as input.[3] To understand time series analysis and time series forecasting with more clarity and depth, we looked into various published papers. The paper helped us with our thesis is discussed in the Literature Review section.

A. Literature Review

It is urgent for us for understanding the inner vehicle framework development rule, for expectation and control.In paper [4], a permeability chart calculation is utilized to change over traffic stream time arrangement into networks. Because of the changed over organizations, a few attributes will be found from the first traffic stream to recognize traffic states. To extensively examine traffic stream time arrangement in various thicknesses, two broad strategies have been presented as follows: multiracial detruded vacillation examination (MFDFA) and permeability diagram. They have taken around 3000 examples

from an assigned street of just the left side which went about as their inspecting point. Their graphical information shows results dependent on 2000 of their examples.

In paper [5], they have examined how to group traffic time arrangements that have comparative vacillation designs. They have utilized a basic normal detrending strategy and just investigation the remaining time arrangement. Second, they have utilized standard investigation (PCA) on crude information and utilize the heaviness of the main d-parts as the highlights of the time arrangement. Third, they have utilized the k-implies calculation to group the traffic time arrangement. The traffic stream time arrangement is recorded at 1000 road traffic stream stations from August 1, 2011, to August 31, 2011. The example timespan crude information is 5 minutes. In this paper[5], they have talked about the grouping of the traffic time arrangement that has comparative variance designs.

In paper [6], they have made far-reaching learn about the key innovations including applying remote sensor organizations to the traffic observing organization, its traffic stream conjecture dependent on the dim determining model, and gridlock control. They additionally utilize the Adaptive GM (1,1) Model which makes some genuine memories moving estimate for city traffic and have a superior conjecture result. They have planned a calculation of traffic stream blockage control and booking for traffic organizations, which is called TRED. They have utilized it for continuous traffic planning and have opened up another approach to consider and take care of gridlock control issues.

Paper [7] presented an improved transient traffic stream determining calculation dependent on the ARIMA model. Forecast of traffic volume, i.e., the normal number of vehicles out and about in the following five minutes to one hour is a significant exploration issue in Intelligent Transportation Systems (ITS) applications.

This paper [8] proposed a model that predicts the traffic stream in blockages. One of the stars of this model is that it depends on the time arrangement. This model will have the option to anticipate the forthcoming traffic conditions in a specific crossing point. The creators guarantee that the model is equipped for anticipating the traffic stream with an exactness of 88.74% and 81.96% for 15minutes previously and 1hr before separately.

In the diary [9], this model was created to destroy the truck gridlock issue in a specific zone in Bombay. The specialists did a fundamental examination and gathered information.A while later, they did an otherworldly examination. At that point, they shaped normalized information.The chose ARMA and ARIMA model for the issue dependent on MMSE and MLR standards. These two were utilized for the choice of the best model for every hallway. From that point forward, they displayed approval.

To improve the predicting exactness of the traffic stream, this paper [10] proposes two combinational conjecture models dependent on GM, ARIMA, and GRNN. At that point, it proposes the utilization of neural organizations to decide variable weight coefficients and builds up the Elman combinational estimate model dependent on GM, ARIMA, and GRNN, which accomplishes the coordination of these three people.

We find in paper [11] that ARIMA performs in a way that is better than all different cases. Notwithstanding, this precision is accomplished to the detriment of computational multifaceted nature.

B. Algorithms/ Models

Time series can be trend cycles, cyclic, seasonal, or irregular. These four components are discussed below:

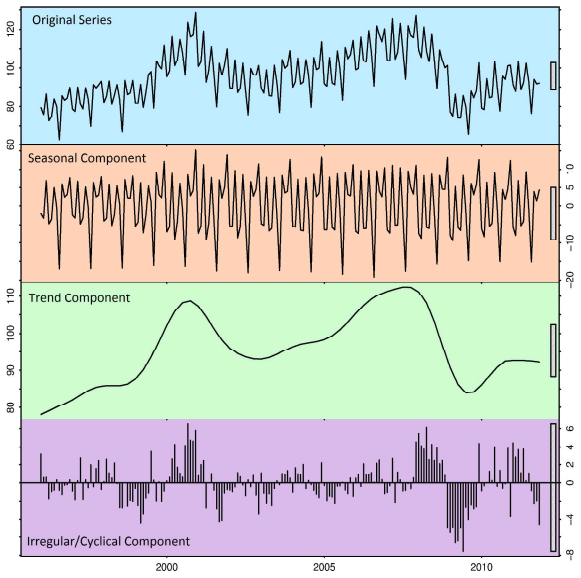

Fig. 1. Trend, Seasonal, Cyclical and Irregular graphical representation. [3]

- **Trend:** It is the change in data over a period of time. Underlying rationale, stochastic, random time series character is provided by a trend that can be deterministic.
- **Seasonal:** It is often of known and fixed frequency. A seasonal pattern occurs when seasonal factors affect the time series—seasonal factors like a particular time of the day or year.
- **Cyclic:** It is the fluctuation of data. Cyclic does not have a fixed frequency, unlike seasonal. [12]
- **Irregular:** Data which are represented in a random manner which is due to unknown, unpredictable, non-seasonal and short-term factors for which the future data cannot be predicted here.

Our data exhibits seasonal behaviour, which is of time series data type. The traffic data plotted for each street shows us an increase in slope in a particular pattern, that is, seasonal pattern. In this work, we present a medium scale comparison study for time series models which would show the highest accuracy in forecasting the traffic data of our seasonal time series dataset. The models or algorithms considered are ARIMA (Auto-Regressive Integrated Moving Average), SNAIVE (Seasonal Naïve), PROPHET, ETS (Error Trend and Seasonality, or exponential smoothing) and the last one is the combination of all models we named it ”mix”.

ARIMA: This deals which various types of time-series data for which it uses its resourceful method which is quite simple but sophisticated in making forecasts for the data. It stands for Auto-Regressive Integrated Moving Average which is a generalisation of the Auto-Regressive Moving Average along with the integration element added. They are a class of statistical algorithms which is used for forecasting and analysing data which are based on time-series.

ETS: Exponential Smoothing Method handles weighted averages of previous observations to forecast future data. As observations becomes aged (in time), these values' priority gets smaller at an exponential rate (since the current vales are given more importance in the series). They are a family of forecasting models.

Seasonal naïve (SNAIVE): Sometimes the data available is not enough and data which is of time-series type. In such circumstances, the naïve method is used which uses previous data from the last observation to make the prediction for the next data. Now Seasonal naïve (SNAIVE) works for data which are very seasonal in nature. SNAIVE makes prediction in the same way as naïve month expect for it predicts the last observed data from the same season of the year.

Prophet: The Prophet is an open-source library, developed by Facebook, which is developed for building forecasts for univariate (single variable) time-series datasets. It is also known as Facebook Prophet which was published by Facebook's Core Data Science team. It is very easy to use. It is built in such a way that it can make accurate forecasts for data having seasonal and trend behaviour. This model is a perfect match for time-series data containing having rich seasonal traits. Also, seasonal historical data works like a charm for this model.
The study shows us the significant difference between each of the models.

III. Proposed Model and Implementation

We have used a dataset from a project called Future Melbourne 2026 Plan. This data set was released as a public resource to use. A contractor was hired by The City of Melbourne to conduct the traffic counts on roads through different roads in the city. The vehicles were categorized into 12 categories and recorded count on per hour. Our data exhibits seasonal behaviour, which is of time series data type. The traffic data plotted for each street shows us an increase in slope in a particular pattern, that is, seasonal pattern.The dataset has been preprocessed and feature extracted to move towards the Implementation stage.

A. Implementation

1) Programming tools and IDEs : The forecasting models are implemented using the R programming language in RStudio. We have used liberties like lubrdate, xts, tibble, fable, fabletools, fable.prophet, feasts, fasster to implement the models and rmdformats, dygraphics, apexcharter, etc. for formatting, plotting graphs, charts.

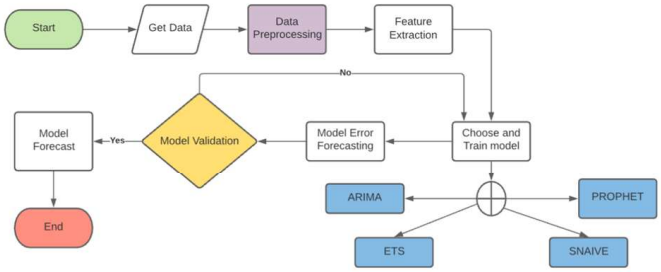

Fig. 2. Work Flow Diagram

2) Target variables for Time Series: As we know, there must be a time variable, based on which the time series models are implemented, and the should be a target variable that would be predicted eventually. In the data preprocessing, we discussed how we altered hour data, which means merging date and time column into one and converted the data type from 'char' to 'timestamp'. Furthermore, we choose 'Total Vehicles' as the target variable since we want to predict the number of vehicles. Finally, we removed the anomalies in road names and used the 'road name' as a unique identifier.

3) Plotting after cleaning the data: After cleaning up the dataset as per our requirement, we plotted the 'Total Vehicles' variable with time where time is the independent variable. The figure shows that the seasonality is very strong, and all the series seem stationary. Almost all the streets have seasonality, and it is quite familiar with traffic data since the amount of traffic passing through a street depends a lot on time.

4) Forecasting models: We have over a hundred unique roads. Though most of them have seasonal data, some of them have unpredictable data too. That is why we want to implement four different forecasting models to see which one performs better for which kind of data. We have chosen four different prevalent and advanced forecasting models: 'ARIMA', 'ETS', 'SNAIVE' and 'PROPHET and lastly, we have a mixed model that gives us the average of all four models. For implementing these models, we have used two frameworks, namely fable, and prophet. The fable package is capable of handling many time series all at once.

There are a total of 66000 traffic data in our dataset from we have taken 29 days as train data and 1 day for test data out of 30 days of traffic data for each street. Hence our training data is to test data ratio is 29:1.

5) Implementing models using fable package : From our primary dataset, we set the entire data as training data except for the last 24 hours since we want to set the forecast horizon as 24 hours, and the time step the data has is an hour. So the last 24 hours historical data is our test data.

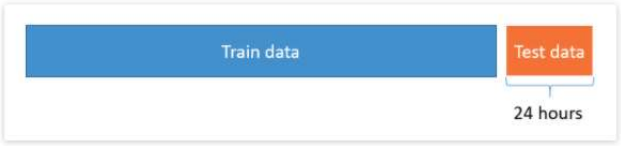

Fig. 3. Training data and test and forecast horizon.

We have built our target variable y as the aggregation of road traffic per road and per hour. Fable package consists of many inbuilt time series models like ARIMA(), ETS(), SNAIVE() etc. Therefore we have used ARIMA(), ETS(), SNAIVE() and prophet() time series functions to our use. We have passed the y parameter through those in built functions to obtain the models variable. The fable package is capable of handling several time series all at once, we decided to use on a single time series. We have fitted all the models in a single call.

```
Agnes %>%
  model(
    arima = ARIMA(y), # fit best ARIMA model using auto.arima
    ets = ETS(y) , # fit the best ETS model
    snaive = SNAIVE(y), # fit seasonal naive model
    prophet = prophet(y~ season("day")+season("week")+season("year"))#fit prophet with multiple seasons
  ) %>%
  mutate(
    mixed = (ets + arima  + snaive +prophet) / 4
  )-> models
```

```
{r}
print(models)
```

Fig. 4. Fitting all models.

After we obtaining the ”models” variable from the previous functions, we pass it through another function from the fable package which is called forecast(). Using the forecast function we place our models variable with a parameter which is the horizon of the forecast (in the case it is 24 hours). From this we obtain the fc variable which gives us our forecasts for a particular street.

To forecast using all the models on our time-series dataset, we then send the object to the ‘forecast‘ function with the horizon of 24 hours.

```
{r eval=FALSE}
fc <- models %>%
  forecast(h = 24)
```

```
{r}
print(fc)
```

Fig. 5. Forecasting with a horizon of 24 hr.

The object returned is known as a ‘mable’ which is a model table which consists of:

- the ‘.model’ column becomes an additional key;
- the ‘.y’ column contains the estimated probability distribution of the response variable in future time-periods;
- the ‘.mean’ column contains the point forecasts equal to the mean of the probability distribution.

6) Plotting the values using dygraphs: After plotting the graph using dygraphs, for a particular street, it looks like:

Fig. 6. Plot of forecast on Agnes Street.

7) Forecast error calculation : To compare these models’ forecast accuracy, we will make a new dataset from the original one for training which not contains traffic data for the last 24 hours. We aim to forecast the last 24 hours, which we have not trained, and then compare those forecasted values with the actual ones. [Fig.3]

A forecast ”error” can be defined as the difference in the forecasted value compared to the actual one. We can measure forecast accuracy by using various techniques related to forecasting errors.

The three most commonly used and famous forecasting error measures we going to use are: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

RMSE is stands for root mean squared error and it is a quadratic scoring decision that measures the average magnitude of error. It is the square root of the average of squared differences between prediction and actual observation.

$$RMSE = \sqrt{1/n \sum_{j=1}^{n} (y_j - \hat{y}_j)^2} \tag{1}$$

Mean Absolute Error (MAE) is an estimation of errors between paired perceptions which are communicating a similar wonder, in statistics. Instances of Y versus X incorporate one estimation method versus an elective procedure of estimation and correlations of anticipated versus observed, subsequent time versus initial time. MAE is determined as:

$$MAE = \sum_{i=1}^{n} |y_i - x_i|/n = \sum_{i=1}^{n} |e_i|/n \tag{2}$$

Among the three forecasting error measuring formulas, MAE is very easy to calculate and understand. Minimizing MAE subsequently directs towards the forecast of median whereas for RMSE it directs towards the forecast of the mean. Consequently, RMSE is used quire more often even

though it is more challenging to understand. The best model has to be the one that minimizes the MAE, RMSE, or MAPE.

Mean absolute percentage error (MAPE), which is also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics. For instance, in pattern assessment, it is additionally utilized as a misfortune work for relapse issues in AI. It for the most part communicates the exactness as a proportion characterized by the equation:

$$M = 1/n\sum_{t=1}^{n}|A_t - F_t|/A_t \tag{3}$$

Percentage errors are easier to interpret as a percentage is a unit-free value which can be depicted out of 100. It is quite advantageous when it comes to comparing different forecasting model's performance on a given time-series dataset.
There is a problem which comes while dealing with percentages. If the actual value or observed value is zero or close to zero then you get your MAPE to be undefined. Also, greater than 100 percent can also happen. Hence it very problematic in many cases to handle MAPE.

```
```{r eval=F}
fabletools::accuracy(fc_train, sdf) %>% select( .model, road_name, .type, RMSE, MAE, MAPE)  -> res
```

You can filter road and get the metrics by typing the name in `search`

```{r}
reactable(
  res %>% mutate(MAE= round(MAE, 2),RMSE= round(RMSE, 2)) %>% arrange(road_name),
  defaultColDef = colDef( headerStyle = list(background = "#4285F4")),
  searchable = T,defaultPageSize = 5,highlight = T)
```
```

Fig. 7. Code snippet of error calculation.

.model	road_name	.type	RMSE	MAE	MAPE
arima	Agnes Street	Test	6.22	4.14	Inf
ets	Agnes Street	Test	5.9	4.39	Inf
mixed	Agnes Street	Test	5.9	3.83	Inf
prophet	Agnes Street	Test	5.17	4.06	Inf
snaive	Agnes Street	Test	8.38	5.84	Inf

1–5 of 385 rows Previous 1 2 3 4 5 ... 77 Next

Fig. 8. Error calculation table.

B. System Implementation

For developing a live system, we have used R mark-down(rmdformats). R markdown is popular because it can be used to mark down and write R code chunks in between, and most notably, with the help of libraries, it becomes a lot powerful. For example, we used the 'knitr' library to develop dynamic report generation, 'ractables' for interactive data tables, 'shiny' for developing interactive web apps, which we eventually deployed to ShinyApps.

IV. Result Analysis

Since RMSE is mostly used for forecasting error evaluation, we have chosen RMSE for our result analysis.

.model	road_name	.type	↑ RMSE	MAE	MAPE
prophet	Little Palmerston Street	Test	0.84	0.57	Inf
mixed	Little Palmerston Street	Test	0.95	0.58	Inf
snaive	Little Palmerston Street	Test	1.06	0.4	Inf
ets	Little Palmerston Street	Test	1.09	0.7	Inf
arima	Little Palmerston Street	Test	1.14	0.65	Inf

1–5 of 385 rows Previous 1 2 3 4 5 ... 77 Next

Fig. 9. Execution of the models regarding RMSE (the lower RMSE, the better the model).

There are 76 streets in the dataset. We have calculated the best performing models and 2nd best performing model in terms of RMSE.

TABLE I
THE BEST MODEL(PROPHET)

Model name	Prophet	Snaive	ETS	ARIMA	Mixed
No. of occurrence	42	12	12	2	8

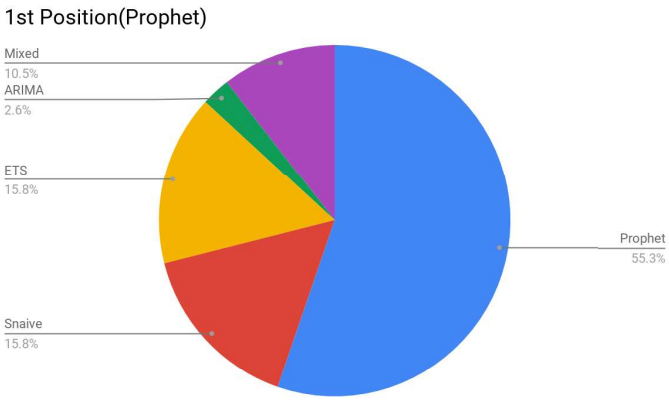

Fig. 10. Pie chart of the best performing model.

TABLE II
$2^n dbestmodel(Mixed)$

Model name	Mixed	Prophet	ARIMA	ETS	Snaive
No. of occurrence	45	11	10	9	1

A. Data with less seasonality

There are some streets where the data is less seasonal. We have observed that the models behave differently with less seasonal data. As we can see in the table below that in less seasonal data, the performance of the model 'Prophet' is the

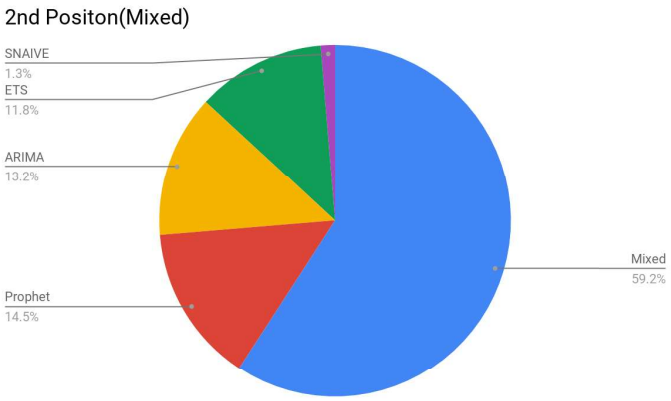

Fig. 11. Pie chart of the second-best performing model.

worst on the contrary, in seasonal data 'Prophet' is the clear winner but in our data set only 9 roads to have less seasonal data rest of the roads have strong seasonality.

TABLE III
LESS SEASONALITY(SNAIVE)

Model name	Snaive	ETS	Mixed	ARIMA	Prophet
No. of occurrence	4	2	2	1	0

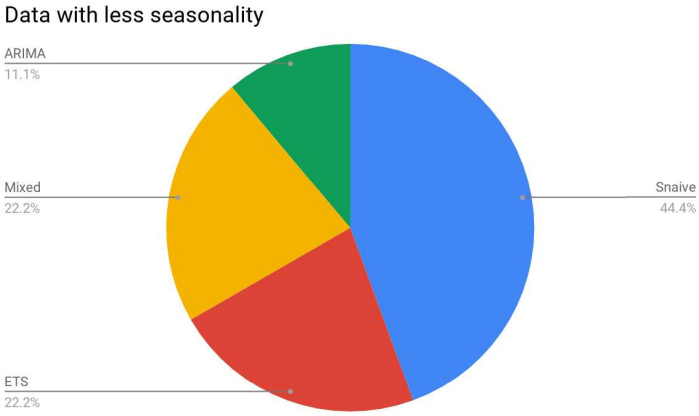

Fig. 12. Pie chart of best performing model in less seasonal data (Snaive).

Out of 76 roads, 67 roads have strong seasonality in the data and the other 9 roads have less seasonality. 'Prophet' is the best model for the streets with strong seasonal data and Snaive is the best model for the roads with less seasonal data. Since the majority of the roads have seasonal data in this dataset and traffic flow on a particular road usually has strong seasonality, we can say that 'Prophet' the best performing model.

V. CONCLUSION

We wanted to find the most suitable forecasting model based on time-series which helps us to forecast future traffic data when there is enough dataset is provided. When this goal in mind, we began to search for models based on prediction, which would enable us to predict the value data. However, upon more research, we found that it, not a prediction but rather forecasting, after which we focused on that. We came across so many time-series forecasting models that it made our work both tedious and fun at the same time.While working with these determining models, we comprehended that not all the models can provoke an exact conjecture this investigation had indicated to us the critical contrast between every one of the models and how they act on a given dataset. We forecasted the daily amount of traffic flow on the major routes in Australia, finding the total of vehicles for each major street by testing and training each model to check for the accurate one.

We were able to find out the model with better accuracy with our dataset with an accuracy of 90% compared to the other models that we have used. That model is PROPHET.

REFERENCES

[1] Hira, F. I., Maruf, M. F., Hossain, A. (n.d.). Stock Market Prediction Using Time Series Analysis.
[2] Tuovila, A. (2020, September 24). Forecasting Definition. Retrieved October 04, 2020, from https://www.investopedia.com/terms/f/forecasting.asp
[3] Using decomposition to improve time series prediction Quantdare. (2019, August 29). Retrieved October 04, 2020, from https://quantdare.com/decomposition-to-improve-time-series-prediction/
[4] J. Bao, W. Chen, Y.-S. Shui, and Z.-T. Xiang, "Complexity analysis of traffic time series based on multifractality and complex network," 2017 4th International Conference on Transportation Information and Safety (ICTIS), 2017.
[5] S. Jiang, S. Wang, Z. Li, W. Guo, and X. Pei, "Fluctuation Similarity Modeling for Traffic Flow Time Series: A Clustering Approach," 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015.
[6] L. Xiao, X. Peng, Z. Wang, B. Xu, and P. Hong, "Research on Traffic Monitoring Network and Its Traffic Flow Forecast and Congestion Control Model Based on Wireless Sensor Networks," 2009 International Conference on Measuring Technology and Mechatronics Automation, 2009.
[7] G. Omkar and S. V. Kumar, "Time series decomposition model for traffic flow forecasting in urban midblock sections," 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), 2017.
[8] M. Karimpour, A. Karimpour, K. Kompany, and A. Karimpour, "Online Traffic Prediction Using Time Series: A Case study," Integral Methods in Science and Engineering, Volume 2, pp. 147–156, 2017.
[9] S. L. Dhingra, P. P. Mujumdar, and R. H. Gajjar, "Application of time series techniques for forecasting truck traffic attracted by the Bombay metropolitan region," Journal of Advanced Transportation, vol. 27, no. 3, pp. 227–249, 1993.
[10] Z. Yu, T. Sun, H. Sun, and F. Yang, "Research on Combinational Forecast Models for the Traffic Flow," Mathematical Problems in Engineering, vol. 2015, pp. 1–10, 2015.
[11] R. K. Yadav and M. Balakrishnan, "Comparative evaluation of ARIMA and ANFIS for modeling of wireless network traffic time series," EURASIP Journal on Wireless Communications and Networking, vol. 2014, no. 1, 2014.
[12] Why Time Series Forecasting Is A Crucial Part Of Machine Learning: Fingent Blog. (2020, April 19). Retrieved October 04, 2020, from https://www.fingent.com/blog/why-time-series-forecasting-is-a-crucial-part-of-machine-learning/
[13] A. A. Haider, "Traffic jam: The ugly side of Dhaka's development," The Daily Star, 13-May-2018. [Online]. Available: https://www.thedailystar.net/opinion/society/traffic-jam-the-ugly-side-dhakas-development-1575355.