IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.005, LIIR204
DOI: https://doi.org/10.52731/liir.v005.204

Optimized Website Traffic Forecasting with Automatic Models and Optuna : A Study in Machine Learning dan Big Data Analytics

Rifqi Mulyawan *, Ayu Ahadi Ningrum †

Abstract

Although time-series forecasting has emerged as a key area of interest in recent years for analyzing historical data to make predictions about future trends, accurately forecasting web traffic can be challenging due to the dynamic nature of the internet and the many factors that can influence user behavior. Existing traffic flow prediction approaches primarily use simple models that are often inadequate for real-world applications. This research aims to develop an optimized machine learning model using FB-Prophet and NeuralProphet for forecasting website traffic and to compare their relative performance and effectiveness in predicting web traffic. This study aims to develop an optimized machine learning model using FB-Prophet and NeuralProphet for forecasting website traffic and to compare their performance in predicting web traffic. Our study found that both FB-Prophet and NeuralProphet models performed better than the simple models used in existing traffic flow prediction approaches. However, the NeuralProphet model outperformed FB-Prophet in terms of accuracy and computational efficiency. The best result obtained from the study was achieved by the NeuralProphet model, which had a Mean Average Error (MAE) of 25.61, Mean Square Error (MSE) of 1354, Root Mean Square Error (RMSE) of 5.060, and a Coefficient of Determination (R2) of 0.882, indicating its superior performance in accurately forecasting website traffic. The results suggest that an optimized machine learning model using NeuralProphet can be an effective way to forecast website traffic and help businesses and organizations better understand web traffic patterns.

Keywords: Facebook Prophet, Foreasting, NeuralProphet, Website Traffic, Time Series.

1 Introduction

With the growing number of internet users worldwide, it's inevitable that website traffic will increase. However, this surge in traffic can lead to various problems, and companies that manage traffic effectively are more likely to succeed. To do this, websites need to predict the number of page visits to allocate computer resources and forecast future income and advertising growth [1].

The internet has become an essential aspect of daily life, and as more and more people gain access to it worldwide, it is inevitable that website traffic will increase. However, increased website traffic can create problems, such as website crashes or prolonged loading times, and

* Universitas Amikom Yogyakarta, Indonesia
† Universitas Muhammadiyah Banjarmasin, Indonesia

the company that manages traffic variation most efficiently is better off [2]. For example, during festivals or special events, various e-commerce sites often crash because more people try to log into the website than it was initially capable of handling. Such events cause many inconveniences for the users, which could decrease the user's rating of the site, and they may end up using another site, reducing the company's business. Therefore, it is essential to implement a traffic management approach or plan to limit the likelihood of such catastrophes, which might be devastating to the company's survival.

In the past, there was no need for such solutions because most servers could manage the traffic flood. However, the smartphone age has increased the demand for some websites to such a high degree that firms could not respond quickly enough to maintain their usual customer service level [3]. Many approaches for projecting online traffic have been presented in the literature. They are roughly grouped into two groups: non-linear forecasting and prediction along a straight or linear line.

Non-linear forecasting models are more complex and allow for more sophisticated predictions, taking into account trends, seasonality, and other factors. In contrast, linear forecasting models are simpler and assume that the trend will continue in a straight line. Companies need to analyze their website traffic patterns and use the most appropriate forecasting models to predict traffic levels accurately [4]. This information can be used to allocate computer resources effectively and forecast future income and advertising growth. By managing website traffic efficiently, companies can improve their customer satisfaction and increase their business success.

The two main groups of models that are commonly used for forecasting purposes are the linear and non-linear models [5]. Within the linear category, the Holt-Winters model is a popular choice. On the other hand, the augmented reality model is another well-known model within this field. When it comes to non-linear prediction, repeating neural networks are often utilized for forecasting purposes [6].

The forecasting model called Prophet was developed by Facebook and was considered the most advanced model in 2018. This model is particularly effective for datasets that exhibit pronounced seasonal patterns [7]. The Prophet is a highly efficient forecasting model developed by Facebook in 2018 that has gained popularity among researchers in illness forecasting. It requires less hyperparameter adjustment and smaller datasets than LSTM. In comparison to the ARIMA model, the Prophet has outperformed with more than 90% results score in forecasting instances in Indonesia, according to Satrio's research [2].

Moreover in Wuhan, China, Liu [8] conducted a study in Wuhan, China, and found that Prophet outperformed ARIMA in predicting medical-related forecasts. Moreover, since the data used in this study is based on actual daily website views, there are fewer occurrences of website visitors on weekends, which can be effectively handled by Prophet due to its ability to handle seasonality in the data caused by reporting schedule.

NeuralProphet [9] is an advanced version of the Prophet model developed by Facebook, which combines the advantages of Prophet with the scalability of neural networks and the interpretability of AR models. The model is built on PyTorch, a popular machine learning framework, and utilizes an autoregressive network called AR-Net to improve accuracy and scalability. This enhanced backend architecture enables NeuralProphet to effectively capture complex patterns and trends in website traffic data, making it a powerful tool for forecasting and analysis [10]. This forecasting approach has been utilized in various fields, with impressive estimation outcomes. Velásquez, for example, utilized NeuralProphet to estimate PV solar plant energy that exceeded ARIMA-LSTM by more than 10%. In theory, NeuralProphet should consistently outperform Prophet, where we also employed the NeruralProphet as one of our recommended models.

In our research, we developed an optimized time series model for predicting website traffic using the Prophet and NeuralProphet models, with Optuna used to tune the parameters. In the next section, we review similar research efforts and describe the success of our proposed strategy using FB-Prophet and NeuralProphet, with and without Optuna. We conclude with some remarks and suggestions for further research, with Figure 1 illustrating the framework used in our investigation.

2 Materials and Method

Previous studies on online traffic forecasting utilize Wavelet pattern analysis and neural networks to predict web traffic [11]. However, this study is laborious and time intensive for real-time website traffic analysis. In addition, Shelatkar et al. research on forecasting online traffic with Arima and LSTM has been shown to improve system performance and efficiency [2].

In recent years, the field of time-series forecasting has seen a surge of research, with a variety of proposed techniques and models. Among these, one study proposes a novel approach for predicting agricultural product quality and harvest amount using smartphone images [12].

Another approach utilizes physical meaning and remote sensing satellite data to predict linear precipitation zones, offering a useful tool for climate analysis and agricultural planning [13]. Additionally, the RLS algorithm is employed as a method for time series data prediction with multiple missing data points, with the aim of improving the accuracy of predictions [14].

Regression analysis is also presented as a useful tool for predicting isoflavone concentration in beans, using Sentinel-2 optical sensor data to improve the accuracy of predictions [15]. Overall, these studies demonstrate the wide range of applications and techniques being developed in the field of time-series forecasting, highlighting the importance of continued research and innovation in this area.

Prophet is a relatively new algorithm in the field of time-series forecasting, but it has been used in various applications. The technique has demonstrated its modeling capability with numerous approaches, including temperature forecasting [16]. For example, Prophet was utilized to model the daily average temperature from a local meteorological station in Myitkyina, Myanmar, from 2010 to 2017, and the predicting model fit the data well with an RMSE of 5.75 (for 2012 and 2014).

Prophet has also shown promising results in the health field. In a study predicting the second wave of COVID-19 patients, Khurana et al. found that NeuralProphet produced performance nearly equal to that of other machine learning algorithms like Random Forest and Passion Distribution [14]. Another study by Hyun et al. employed a NeuralProphet model to detect diabetic foot sores. The model used multiplicative future regressors and demonstrated strong modeling power [18].These studies highlight the diverse applications and potential of Prophet and NeuralProphet in the field of time-series forecasting.

3 Materials and Method

This section explains the methodology used to create a website traffic forecasting model. The literature review results and domain experts' knowledge were used.
The system design flow is depicted in Fig. 1 for analyzing the performance of the projections, and the obtained website traffic forecasts were assessed using established evaluation metric(s).

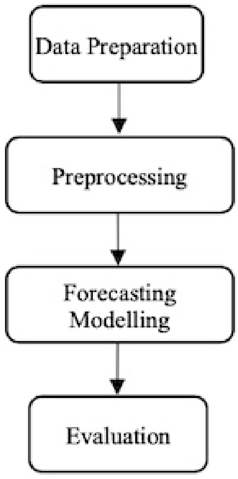

Figure 1: Optimized Website Traffic Modeling System Design

3.1 Data Preparation

We used a dataset from a year's worth of website data captured on Google Search Console for our investigation. The data set includes web page traffic statistics for the previous 16 months, organized by date.

The time series dataset, which has 487 rows, has five columns: (1) The date on which the website data was recorded, (2) Clicks is a column that shows how many times a website has been clicked in search results. (3) Impressions represent the number of times our website has appeared in search engine results. (4) The click-through rate (CTR) is the percentage of clicks to websites appearing in search results. (5) Position is a number that indicates the level at which a website page is displayed in a query. The columns chosen for inclusion in the model are the "Date" and "Click" columns, which are the most relevant to website visitors. To summarize, this dataset represents the daily views of various online pages. UrlWebsite is a particular data set from September 9, 2021, to January 8, 2023. Fig. 2 shows a time-series data visualization of the dataset used.

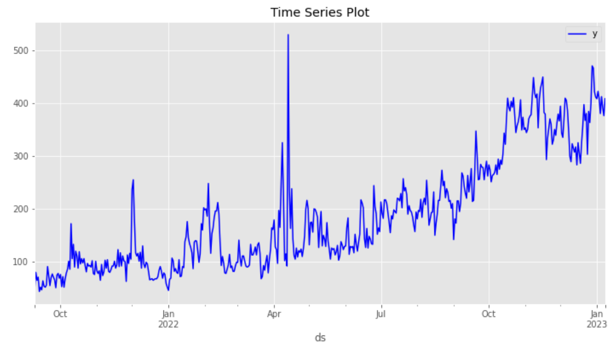

Figure 2: UrlWebsite Time-Series Dataset Plot Image

3.2 Preprocessing

The data extracted from Google Search is a file with the .xlsx format, where the data is then inserted into a DataFrame. The main preprocessing stage is to remove duplicate data. Before fitting the model, the 'date' and 'views' columns must be renamed to 'ds' and 'y', respectively, a standard introduced by Facebook.

NeuralProphet, like Facebook Prophet, must also reformat the input DataFrame submitted to the model when training the forecasting time series model. The DataFrame should have two columns: (1) 'ds' for the date and (2) 'y' for the target variable we want the model to forecast.

3.3 Website Traffic Forescasting Modeling

As described in the previous part, the suggested model in this work was developed using an optimized Prophet and NeuralProphet. These algorithms demonstrated good interpretability in a variety of study areas. However, earlier research on NeuralProphet and Prophet algorithms is still lacking in many sectors, particularly power generation forecasting. These algorithms were only recently revealed, and academics have yet to experiment with them in other domains.

The implemented time-series methodology involves three steps for obtaining forecasting values using a hybrid model built with the machine learning techniques Prophet and NeuralProphet. These methods generally require two inputs: (1) a (ds) input representing the date-time property, and (2) a (y) input representing the target (forecasted) value, which were modified accordingly based on the supplied dataset column heads.

To use Prophet for predicting, the first step involves defining and configuring a Prophet() for FB-Prophet and NeuralProphet() for NeuralProphet object, setting the type of model, such as the growth and seasonality, among other parameters. By default, the model works independently to automatically figure out practically everything. The next step is to fit the model on the dataset by invoking the fit() function and providing the data. This function requires a DataFrame containing time series data, where the first column must be labeled "ds" and contain date times, while the second column must be labeled "y" and contain the observations. Therefore, the column names in the dataset must be changed, and the first column must be converted to date-time objects if it is not already in the correct format, such as during the dataset loading stage.

3.4 Evaluation

In time-series modeling, the data is often split into two parts: training data and test data, in an 80:20 ratio. The purpose of this split is to assess the accuracy of the time-series model for forecasting website traffic. Model evaluation is an essential aspect of machine learning as it helps determine how well a model performs. Typically, evaluation metrics are chosen to determine the optimal model that produces accurate predictions for future outcomes. To objectively evaluate model performance, unseen data, which the model has not been trained on, is used.

In this research, two commonly used evaluation metrics for time-series modeling, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), were employed. The MAE provides the average value of the anticipated error, with all errors weighted equally. In contrast, the RMSE gives more weight to larger errors, which is particularly useful when significant errors are undesirable, such as in this proposed model. The mean squared error and R-squared error are two other evaluation metrics often used in this type of algorithm [19].

The evaluation equation for the performance metrics used in the study can be expressed as follows:

$$MAE = (1/n) * \Sigma|Yi - \hat{Y}i| \quad (1)$$

$$MSE = (1/n) * \Sigma(Yi - \hat{Y}i)^\wedge 2 \quad (2)$$

$$RMSE = sqrt(MSE) \quad (3)$$

$$R2 = 1 - (\Sigma(Yi - \hat{Y}i)^\wedge 2 / \Sigma(Yi - \bar{Y})^\wedge 2) \quad (4)$$

Where: (1) Yi is the actual value of the i-th observation. (2) $\hat{Y}$i is the predicted value of the i-th observation. (3) n is the total number of observations. (4) $\bar{Y}$ is the mean of the actual values.

It is essential to note that the training dataset used in this study cannot be used for objective evaluation to avoid model overfitting. Therefore, the model's performance was assessed using the test dataset, which was not used during the model training process. Additionally, the model's performance was visualized to provide a better understanding of its behavior and how well it performed in forecasting website traffic.

4 Results

The process of forecasting website traffic is a complex one, as it requires understanding and analyzing time series data. One of the essential steps in this process is plotting the data. By plotting the data, we can identify any patterns, trends, seasonal cycles, or anomalies in the data. It also allows us to visualize the data and get a better sense of its behavior over time. In the case of website traffic forecasting, we can plot the data to identify traffic trends based on clicks over time and monthly seasonal sales patterns. These trends need to be considered in the forecasting model to make accurate predictions. To optimize the FB-Prophet-based and NeuralProphet-based algorithms for forecasting website traffic, the default model parameter with the "Country Holidays" function was used. This function informs the model about holidays in different countries, which can impact website traffic. After creating the optimized models, it is essential to compare their performance to determine which one performs better. This comparison helps us choose the best model to use for forecasting website traffic accurately.

4.1 Default FB-Prophet and NeuralProphet

Facebook prophet may be fitted by creating a Prophet or NeuralProphet object. In this case, the dataset's frequency is 'day,' and the default model is used to project the future of 1 month (30 days) data. The FP-Prophet consists of three components: seasonality, trend and irregular components [1]. The equation 5 show the simplest form of the FB-Prophet.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t). \qquad (5)$$

The components of equation (5) are:

• g (t): piecewise linear or logistical growth curve for modeling non-periodic changes in time series S (t): periodic changes (eg weekly / annual seasonality)

• h (t): effect of holidays (user provided) with irregular schedules

• (t): the error term accounts for any unusual changes are un-accommodated by the model. The FB-Prophet and NeuralProphet both set all of the settings by default. Figures 3 and 4 depict forecasts using the model's default value.

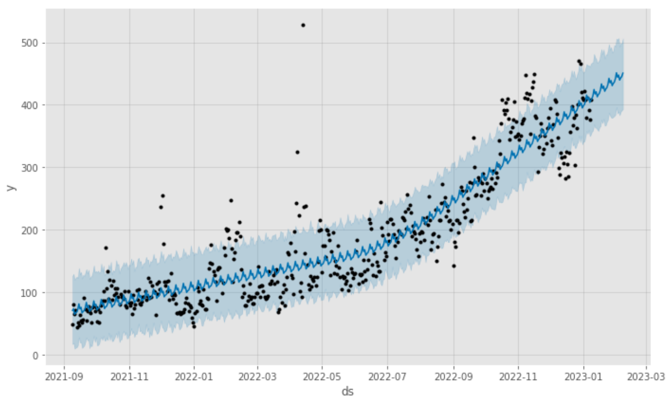

Figure 3: Default FB-Prophet Forecasting

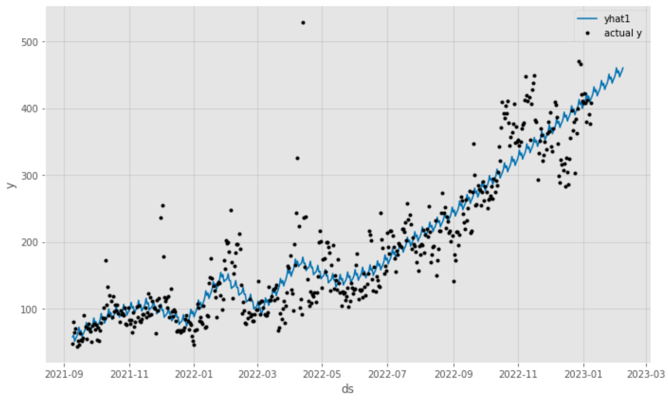

Figure 4: Default NeuralProphet Forecasting

These models create trend and weekly plots by default, as shown in Fig. 5 for FB-Prophet and Fig. 6 for NeuralProphet.

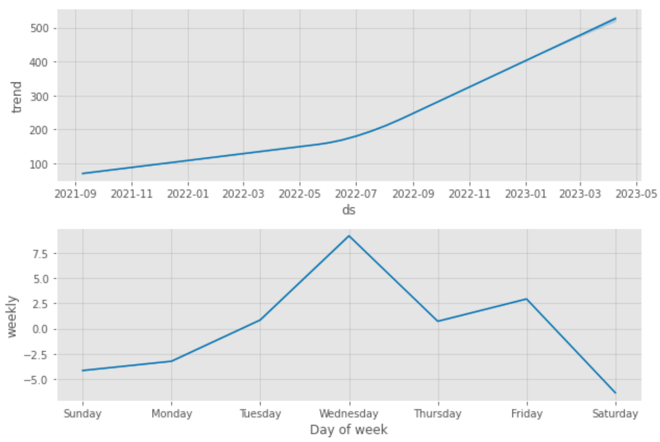

Figure 5: Default FB-Prophet Forecasting Components

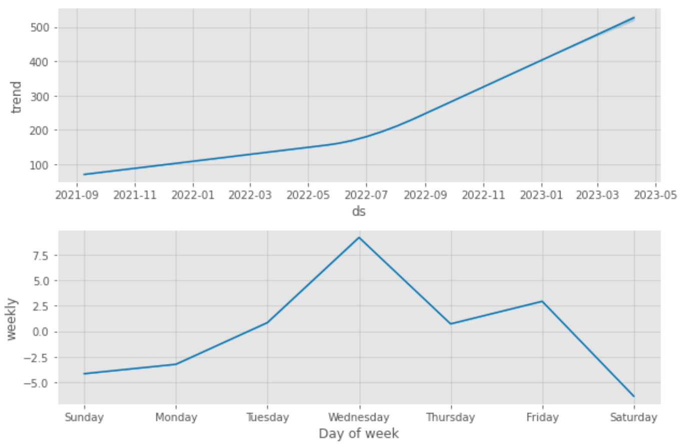

Figure 6: Default NeuralProphet Forecasting Components

The FB-Prophet-based model forecasts very well using the data, but the NeuralProphet is more detailed in catching the trends.

Quantitative results for the Default FB-Prophet and NeuralProphet models can be seen in Table I below.

Table 1: Result of The Default Model

Automatic Model	**Mean Average Error (MAE)**	**Mean Square Error (MSE)**	**Root Mean Square Error (RMSE)**	**Coefficient of Determination (R2)**
FB-Prophet	31.21	1891	5.587	**0.834**
NeuralProphet	**29.08**	**1679**	**5.393**	0.852

Based on the quantitative results from Table I above, it can be seen by default that NeuralProphet's performance is proven to be better without the need to add additional parameters for forecasting data. However, many parameters can be improved by using the Prophet model to get better forecasting results.

4.2 Optimized Fb-Prophet and NeuralProphet

In this context, the use of Optuna software framework for optimizing hyperparameters is aimed at enhancing the effectiveness of FB-Prophet and NeuralProphet in forecasting website traffic. The author employed Optuna to tune the parameters of these models and attempted to obtain reliable optimization, which is critical in ensuring that the models produce accurate results when forecasting traffic on the website. The tunable parameters include seasonal prior distribution, degree of influence, data range used to detect change points, trend influence, and others-related parameters.

To achieve optimization, the author added parameters to the FB-Prophet and NeuralProphet models that allow them to be optimized with a 100 percent trial by Optuna. The parameters used to achieve optimization in the second of two traffic forecasting models include n_changepoints, changepoint_range, seasonality_mode, seasonality_prior scale, daily_seasonality, weekly_seasonality, and yearly_seasonality.

The qualitative results of plotting the two optimized website traffic forecasting models can be seen in Fig. 7 and Fig. 8, which give a clear view of the improvement in forecasting accuracy.

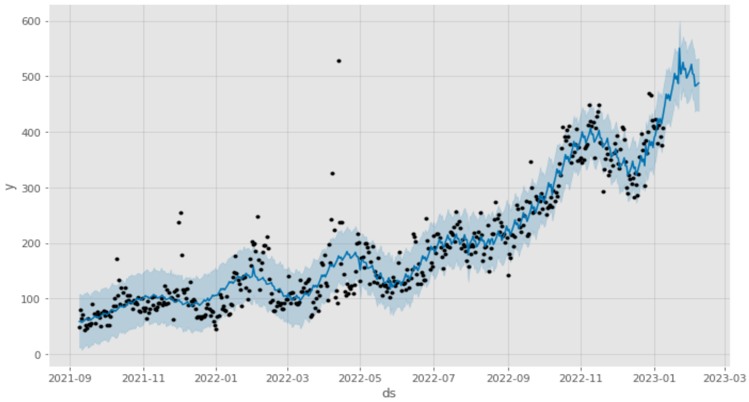

Figure 7: Optimized FB-Prophet Forecasting

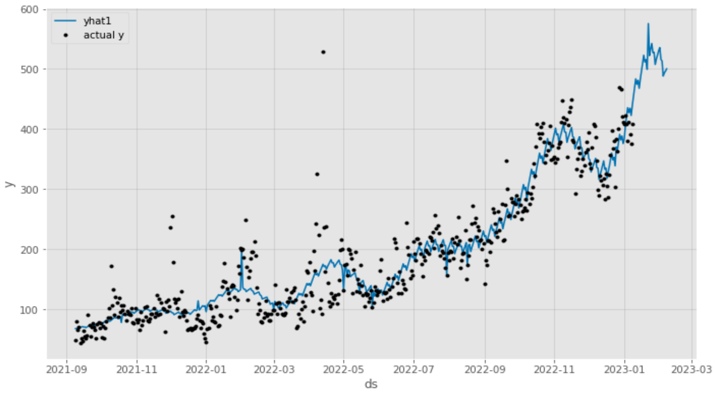

Figure 8: Optimized NeuralProphet Forecasting

With the increase in the parameters used, the models also make plots for daily, weekly and yearly trends, which can be seen in Fig. 9 for FB-Prophet and Fig. 10 for NeuralProphet.

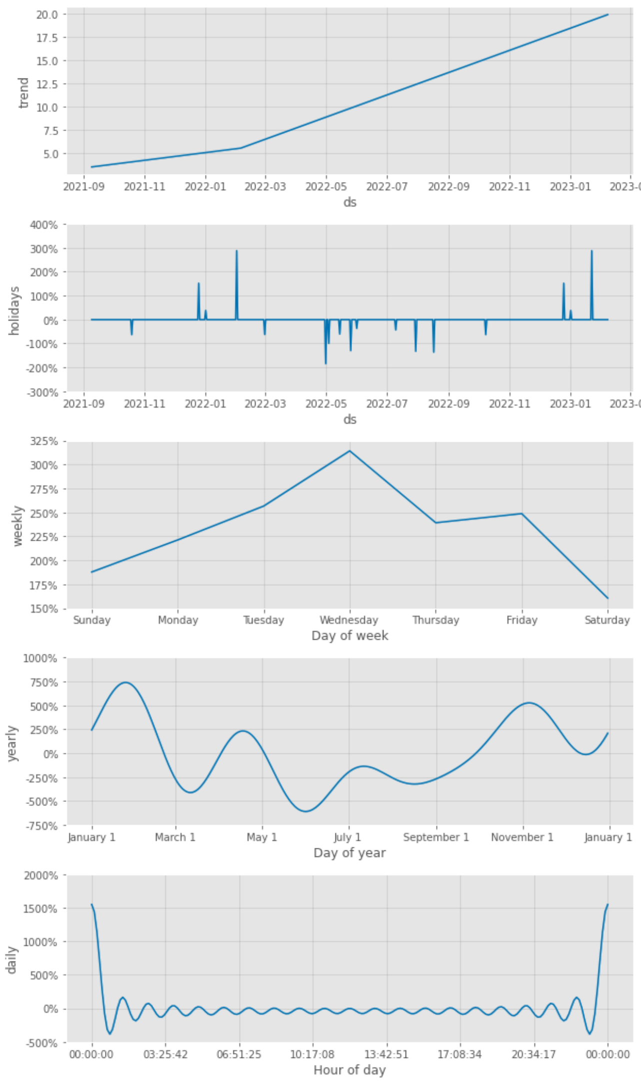

Figure 9: Optimized FB-Prophet Forecasting Components

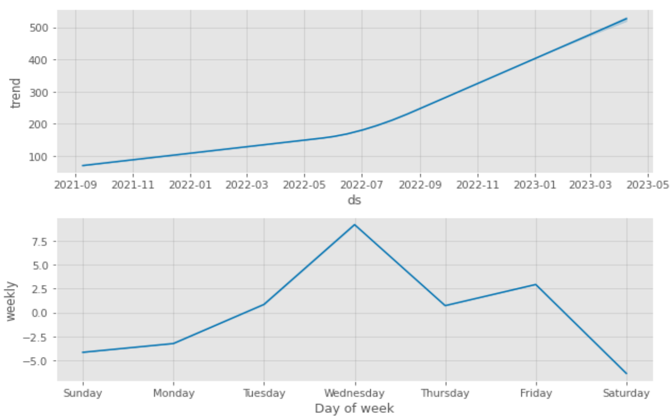

Figure 10: Optimized NeuralProphet Forecasting Components

As we can see, the results indicate that both the optimized FB-Prophet and NeuralProphet models are effective in capturing trends in the website traffic data. This is evident from the quantitative results presented in Table II, which demonstrate the accuracy and performance of the optimized.

Table 2: Result of Optimized Model

Automatic Model	**Mean Average Error (MAE)**	**Mean Square Error (MSE)**	**Root Mean Square Error (RMSE)**	**Coefficient of Determination (R2)**
FB-Prophet	25.85	1355	5.084	**0.881**
NeuralProphet	**25.61**	**1354**	**5.060**	0.882

Based on the qualitative results in Table II above, it can be said that optuna optimization works very well on both models in determining the parameters that best suit the dataset used, which in this case is website click data from Google Search Console.

5 Discussion

The results of this study indicate that both FB-Prophet and NeuralProphet models can be used for forecasting website traffic. By default, NeuralProphet performs better than FB-Prophet in capturing trends in the data. However, the Prophet model can be optimized by adjusting parameters to achieve better forecasting results. To optimize the models, the study used the Optuna software framework to tune the parameters of both models, including seasonal prior distribution, degree of influence, data range used to detect change points, trend influence, and others-related parameters. The results of optimization showed a clear improvement in forecasting accuracy for both models.

Moreover, the optimized models can capture daily, weekly, and yearly trends very well. The study's findings suggest that Optuna optimization works very well on both models in determining the best parameters that suit the dataset used, which in this case is website click data from Google Search Console.

The study's results are essential because forecasting website traffic accurately is crucial for businesses to make data-driven decisions. By using machine learning models like Prophet, businesses can get quick and accurate forecasts without the need for technical expertise in predictive modeling. Furthermore, optimization using software frameworks like Optuna can improve the model's accuracy further.

6 Conclusion

The forecasting analysis conducted using three methods, including FB-Prophet and NeuralProphet, revealed that NeuralProphet had the best accuracy results, while FB-Prophet had a slightly higher error value. However, the difference was minimal after each optimization.

The Prophet model offers an excellent option for generating quick and accurate forecasts. It has intuitive parameters that can be adjusted by people with subject expertise but no technical expertise in predictive modeling. The Prophet model also supports the concept of holidays, which involves identifying seasonal trends like retail sales during Thanksgiving and Christmas holidays to improve future forecasts. The Prophet API is easy to use, and it is compatible with the python data science workflow because it uses the standard Panda's DataFrame and matplotlib library to display data. Web traffic forecasting using the Prophet model is highly recommended for businesses. It is recommended that companies choose the results of Web Traffic forecasting with the lowest error value to make business decisions.

In the future, it may be worthwhile to explore other parameters using more extensive datasets and implementing other machine learning methods or using off-page factors not considered in this study as we hope that this research will benefit service providers such as cloud server service providers, search engine optimizers, and field software developers by encouraging them to use machine learning models to automate website traffic tasks.

References

[1] Y.-C. Tsao, Y.-K. Chen, S.-H. Chiu and J.-C. Lu, "An innovative demand forecasting approach for the server industry," Elsevier, vol. 110, 2022.

[2] T. Shelatkar, S. Tondale and S. Yadav, "Web Traffic Time Series Forecasting using ARIMA and LSTM RNN," in ITM Web of Conferences, Nerul, 2020.

[3] N. Suresh, B. Priya, M. S. Teja, B. S. Krishna and G. Lakshmi, "Historical Analysis and Forecasting of Stock Market Using FBPROPHET," South Asian Journal of Engineering and Technology, pp. 152-157, 2022.

[4] V. Kramar and V. Alchakov, "Time-Series Forecasting of Seasonal Data Using Machine Learning Methods," MDPI, vol. 16, p. 248, 2023.

[5] D. Shah, P. Sanghani and M. Rahevar, "Short-Term Traffic Flow Prediction for an Urban Highways using Time Series Forecasting Model," International Journal of Innovative Science and Research Technology, vol. 6, no. 12, pp. 134-140, 2021.

[6] B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," Philosophical Transactions of the Royal Society, vol. 379, 2021.

[7] C. B. A. Satrio , W. Darmawan, B. U. Nadia and N. Hanafiah, "Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET," in International Conference on Computer Science and Computational Intelligence, Jakarta, 2020.

[8] H. T. Rauf, . M. I. U. Lali, M. A. Khan and S. Kadry, "Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks," Springer, p. 733–750, 2021.

[9] O. Triebe, H. Hewamalage and P. Pilyugina, "NeuralProphet: Explainable Forecasting at Scale," pp. 1-40, 2021.

[10] O. Triebe, L. R and R. R, "AR-Net: A simple Auto- Regressive Neural Network for time-series," Arxiv, vol. D, pp. 1-12, 2019.

[11] Z. Cui, R. Ke, Z. Pu and X. Ma , "Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction," Elsevier, vol. 115, 2020.

[12] S. Patandung and I. Jatnika, "The FB Prophet Model Application to the Growth Prediction of International Tourists in Indonesia during the COVID-19 Pandemic," International Research Journal of Advanced Engineering and Science, vol. 6, no. 2, 2021.

[13] K. Arai, "Data Retrieval Method based on Physical Meaning and its Application for Prediction of Linear Precipitation Zone with Remote Sensing Satellite Data and Open Data," International Journal of Advanced Computer Science and Applications,, vol. 11, no. 10, pp. 56-65, 2020.

[14] K. Arai and K. Seto, "Recursive Least Square : RLS Method- Based Time Series Data Prediction for Many Missing Data," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, pp. 66-72, 2020.

[15] K. Arai, Y. Hideshima, Y. Iwak and R. Ito, "Method for Estimation of Oleic Acid Content in Soy Plants using Green Band Data of Sentinel-2/MSI," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, pp. 56-63, 2022.

[16] Z. Z. Oo and S. Phyu, "Time Series Prediction Based on Facebook Prophet: A Case Study, Temperature Forecasting in Myintkyina," International Journal of Applied Mathematics Electronics and Computers, vol. 4, pp. 263-267, 2020.

[17] A. Khurana, "Second wave of COVID-19: Steps and future vision," Indian J. Res. Homoeopath, vol. 15, pp. 75-76, 2021.

[18] J. Hyun, Y. Lee and H. M. Son, "Synthetic Data Generation System for AI-Based Diabetic Foot Diagnosis," Springer, vol. 2, no. 5, 2021.

[19] D. D. Chuwang and W. Chen, "Forecasting Daily and Weekly Passenger Demand for Urban Rail Transit Stations Based on a Time Series Model Approach," MDPI, vol. 4, pp. 904-924, 2022.