

Website Traffic Forecasting Using Python and Machine Learning

Ram Babu Jaiswal¹, Madhusmita Thakuria², Akhilesh Yadav³, Neha Bagga⁴, Dr. Sheetal Kalra⁵, Dr. Sartaj Singh⁶

Student^{1,2,3}, Assistant Professor⁴ Associate Professor⁵, Associate Professor⁶
School of Computer Science and Engineering^{1,2,3,4}, Department of Computer Science and Engineering⁵, School of Computer Application⁶
Lovely Professional University-Phagwara, Punjab India^{1,2,3,4,6}, Guru Nanak Dev University, Regional Campus, Jalandhar, Punjab, India⁵

Abstract: In contemporary times, websites serve as digital storefronts worldwide, comprising the largest segment of internet traffic. The forecasting of website traffic involves predicting future visitor numbers, which is beneficial for formulating marketing strategies, allocating resources, and optimizing websites. Measured in terms of sessions within a specific time frame, website traffic varies considerably based on factors such as the time of day, day of the week, and other variables. The capacity of a platform to handle web traffic depends on the size of the servers supporting it. An increase in website visitors may lead to crashes or slow loading times, resulting in potential disruptions. The accuracy of internet traffic flow forecasting is heavily reliant on historical and real-time traffic data collected from various sources that monitor network flow.

IndexTerms: ARIMA, machine learning, time series analysis, regression, prediction.

I. INTRODUCTION

A website is a collection of web pages that a user visits every second, depending on the reason for the visit, the visitor's objectives, and how they found the website. Websites need to forecast view counts for individual pages in order to manage computer resources and project growth in future sales and advertising. Nevertheless, bot-generated traffic is not taken into consideration in this estimation. The bulk of internet traffic since the mid-1990s has been generated by web traffic, which is measured by the quantity of visitors and pages viewed. Websites use traffic analytics to track inbound and outbound traffic to see which pages and sections are most popular and whether any trends are apparent, like a page being viewed primarily by visitors from a particular nation. There are many ways to monitor traffic, and the information gathered is utilized to find security flaws, improve the structure of websites, or highlight bandwidth shortages. While some businesses offer advertising schemes where they pay for screen space on the website in exchange for increased traffic, not all web traffic is desirable. Additionally, "fake traffic" is produced by bots for the advantage of a third party.

Future trend forecasting using historical data is a common task in business analytics. An analyst must have a thorough understanding of the field as well as a firm grasp of intricate mathematical theories in order to tackle the difficult topic of forecasting. Many business forecasting techniques rely on intuition and linear regression, but more complex models can produce better results at the expense of more difficult implementation.

II. LITERATURE REVIEW

Using real-time traffic data, a machine learning algorithm based on ARIMA predictive analysis is employed. A machine learning algorithm using actual daily, weekly, monthly, and annual traffic data trains the proposed supervised model. Findings show that the model's accuracy is higher than 70%, and with ongoing improvements, it will reach higher than 95% until the project's final evaluation. Studies indicate that ARIMA-based machine learning models can be used to analyze traffic data in real-time and accurately estimate traffic flow conditions that are influenced by seasonality, SEO, and marketing tactics, among other factors. This research does have certain limitations, though. For instance, only a certain amount of data traffic could be used for a certain amount of time due to memory capacity constraints. The availability of data for analysis also presents some difficulties. Redefining the extent of the data used for predictive analysis is required to improve the model's accuracy. The results of the prediction have been analyzed using the ARIMA model to ascertain the range of its accuracy. Data that exhibits evidence of a non-stationary process in the mean but not in the variance or autocovariance is analyzed using the autoregressive integrated moving average model, or ARIMA. Predicting sequential data is one use for it.

The Box-Jenkins approach is another name for the flow chart that is provided. We examine the time series and give special attention to determining p , d , and q in order to construct this model. • d -integrated stationary; • q -moving average; • p -auto-regressive. We use ACF, PACF, and the unit root tests to find the proper values of p , d , and q in order to create a class for identification. D is in the integration order, p stays in the AR order, and q stays in the MA order. Based on the most appropriate values of p , d , and q , the estimation is completed. The most accurate values and a sample are used for forecasting. The process of forecasting website traffic involves using business analysis to forecast future traffic on websites based upon the available data from the past. Most of the focus is on collecting daily history, cleaning, and processing the historical data while adjusting it as per holidays and aggravating it on a weekly or monthly basis. After this analysis, we will apply different prediction models to it. After the model analysis, selection and aggregation of data have to be done, and it is converted into a daily and weekly forecast after implementing daily and holiday adjustments to the analyzed information.

III. METHODOLOGY

➤ Data Collection and Preparation

The initial stage required gathering two years' worth of historical website traffic data. This dataset comprised daily totals of distinct users, divided into those who were returning and those who were new. To deal with missing values, outliers, and inconsistencies, preprocessing and data cleaning were done. The dataset was properly formatted and organized for analysis.

➤ Exploratory Data Analysis (EDA)

An analysis of the website traffic data using EDA was done to find underlying patterns and trends. We examined the traffic distribution, looked for trends and seasonality, and found any anomalies or patterns using descriptive statistics, visualizations, and time series plots.

➤ Model Selection

The ARIMA (Autoregressive Integrated Moving Average) model was selected because of how well it handled time series data when predicting website traffic. ARIMA is a useful technique for forecasting future traffic patterns because it effectively captures trends, seasonality, and autocorrelation found in time series data.

➤ Parameter Identification

P , D , and Q are the three crucial parameters that must be determined in order to use the ARIMA model. To find the proper values of p , d , and q , the unit root tests, the Partial AutoCorrelation Function (PACF), and the AutoCorrelation Function (ACF) were employed. The ARIMA model is composed of three components that are controlled by these parameters: the moving average (MA), differencing (I), and autoregressive (AR).

➤ Model Training and Evaluation

With an emphasis on fitting the model to capture the underlying patterns and dynamics of website traffic, the ARIMA model was trained using the prepared dataset. In order to evaluate the trained model's accuracy and predictive power, suitable performance metrics like Mean sq\,d Error (MSE) and Root Mean sq\,d Error (RMSE) were used.

➤ Forecasting

Website traffic forecasting for the required time period was carried out following model evaluation and training. To confirm the ARIMA model's accuracy and dependability, the predicted values were contrasted with the real traffic statistics.

➤ Visualization and Interpretation

Time series plots and forecasted versus were among the suitable graphs and charts that were used to illustrate the forecasting process' outcomes. genuine comparisons of traffic. In order to derive practical insights and facilitate

decision-making concerning marketing tactics, resource allocation, and website optimization, the results were interpreted.

Mean squared error: A popular metric for assessing the effectiveness of regression models, including time series forecasting models like ARIMA, is the mean squared error (MSE). It calculates the average squared difference between the ground truth values, or actual values, and the model's predicted values. Because the model's predictions are more closely aligned with the actual values, lower MSE values are indicative of better model performance.

$$MSE = \frac{1}{n} \sum (y - y')^2 \quad (1)$$

RMSE: Regression models, which include time series forecasting models like ARIMA, are frequently evaluated for prediction accuracy using the Root Mean squared Error (RMSE) metric. It gauges how much the average error magnitude is between the dataset's actual values and predicted values. With respect to the observed values, the model's predictions' typical deviation, or "error," is quantified by RMSE.

$$RMSE = \sqrt{\frac{\sum (y - y')^2}{N - P}} \quad (2)$$

Traffic Estimation and Forecasting



Fig 1. Traffic Estimation and Forecasting

❖ Traffic Estimation

The effectiveness of a website is heavily influenced by the ability to predict and understand website traffic accurately. Estimating and forecasting traffic are crucial elements of website strategizing, involving activities like developing content, determining advertising budgets, and managing resources. Through analyzing and predicting traffic trends, website owners and marketers can make well-informed decisions to enhance their digital footprint.

❖ Analyzing Historical Data

Analyzing historical data is one of the main techniques used in traffic estimation. Website owners can identify trends and patterns that help them understand the factors influencing the traffic to their website by closely examining previous traffic patterns.

❖ Utilizing Website Analytics Tools

Resources for tracking website traffic include Google Analytics and other website analytics tools. These tools facilitate user engagement measurement, traffic source identification, and visitor behavior tracking. Website owners can make

educated decisions about content creation and marketing strategies by examining these metrics, which allow them to project future traffic based on historical data.

❖ **Studying Competitor Analysis**

Valuable insights for traffic estimation can be obtained by looking at competitors' websites and doing traffic analysis. Website owners can examine their competitors' traffic sources, keywords, and overall performance with a number of tools, such as SimilarWeb and Ahrefs.

❖ **Leveraging Keyword Research**

Researching keywords is essential for both search engine optimization and traffic estimation. Website owners can assess how much potential traffic they can draw by identifying important keywords with high search volume. Search volumes can be estimated and popular keywords can be found by using tools like SEMrush or Google Keyword Planner.

IV. DESIGN AND IMPLEMENTATION

Analysis of the dataset was done using "Users" as the sample data. The dataset was further separated into testing and training sets. We plotted the quantity of hits vs. time series data. days during the testing period along with actual values and projections for the article "Users.". The page visits are shown on the y-axis as powers of 1000, and the time interval is represented on the x-axis. The forecast results for each month



Fig 2. Forecasting of daily visitors

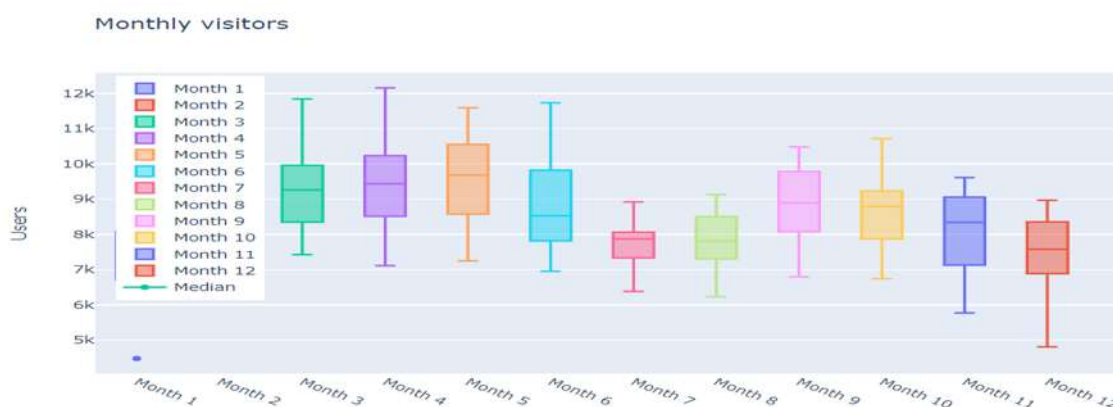


Fig 3. Forecasting of monthly visitors

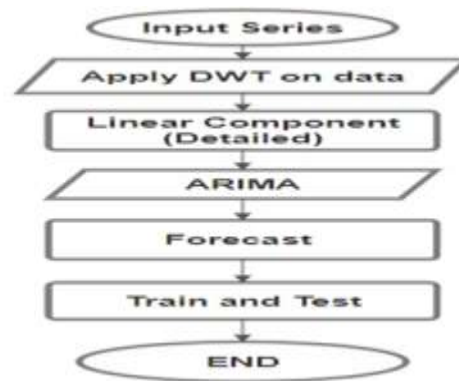


Fig 4. Flowchart of Proposed System

V. RESULTS AND BENEFIT

Accurate Prediction: Our in-depth analysis has resulted in the creation of a strong forecasting system that accurately predicts future website traffic. Our model has been verified for accuracy through thorough comparisons with real traffic data, proving its reliability.

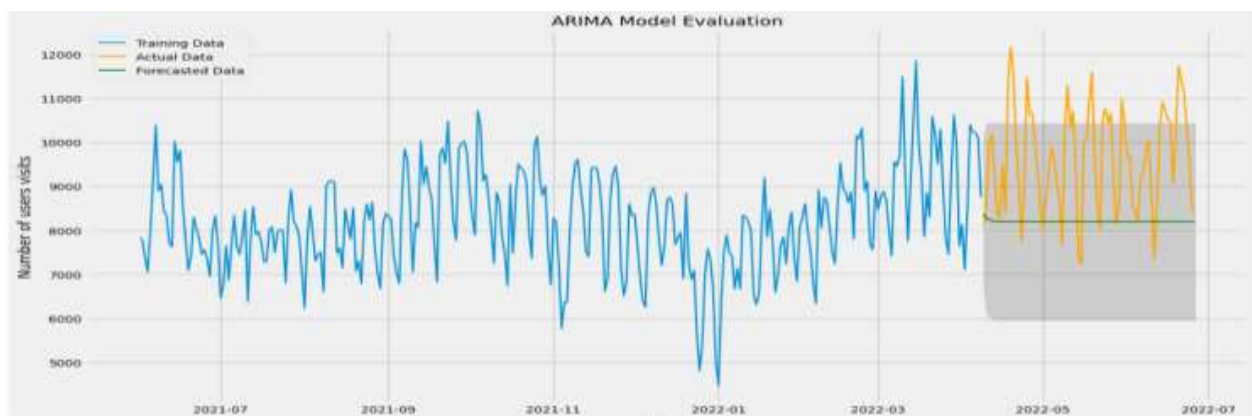


Fig 5. ARIMA model evaluation

Marketing campaigns and business strategies: Marketing is a major factor that can determine the growth of your website. Marketing strategies can also influence the type of traffic present on a web application. For example, an ad campaign can target certain sections of society, such as gender groups, ethnic groups, and age groups. It's important to understand your target audience and run promotions accordingly, and analysis of past campaign results helps in understanding your future potential customers and traffic.

Seasonality: Let's say that your website has data related to clothing and fashion. It is very likely that people are more likely to search for clothes according to the season, and thus your website should be optimized accordingly.

SEO performance: Keyword trends analysis and monitoring results of your website's ranking on search engine research pages (SERP) is essential in determining the impact on the traffic of your website. Additionally, if changes are done with the search engine's algorithm then viewership of your website can be impacted

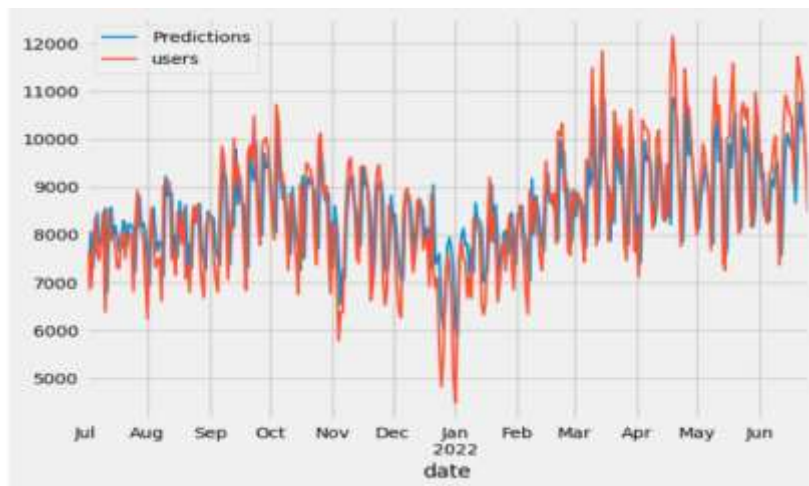


Fig 6.Forecast results for ARIMA

VI. CONCLUSION

Establishing an accurate prediction model for predicting the quantity of website visitors over a given time period was the main goal of this study. We used a trained dataset to accomplish this, and we visualized the outcomes to get a better understanding of the dynamics of website traffic. Two years' worth of monthly counts of unique users—both new and returning—were used to train our model using website traffic data.

Our analysis's conclusions showed that our forecasts of website traffic were quite accurate, demonstrating the potency of our forecasting methodology. We thoroughly assessed our prediction model's performance using the Autoregressive Integrated Moving Average (ARIMA) model. A popular method for time series analysis that takes trends, seasonality, and other temporal patterns into account in the data is the ARIMA model. We were able to evaluate the robustness and dependability of our prediction model through this assessment.

It's crucial to recognize some of our research's limitations, though. One such restriction is the ability to forecast website traffic data for only the last two years. Future studies could examine how well the prediction system works with datasets that have longer durations, even though this timeframe offered insightful information about short-term traffic patterns. Findings from the analysis of data from more years may be more trustworthy and practical, especially when it comes to comprehending long-term patterns and trends in website traffic.

Additionally, feature selection and model refinement may be the main topics of future research projects. It may be possible to find pertinent features that greatly influence website traffic prediction by analyzing data spanning more years. We can improve the model's accuracy and predictive capacity by making adjustments based on these realizations.

To further increase the accuracy of the remaining models, we intend to investigate methods of modifying their parameters. To ensure that predictions closely match observed data, fine-tuning model parameters can help optimize performance. In order to help website owners make better decisions about how to optimize their websites for increased traffic and engagement, we constantly work to improve our prediction model.

REFERENCES

- [1]. Azzouni, A., Pujolle, G.: A long short-term memory recurrent neural network framework for network traffic matrix prediction. *Comput. Sci.* **3**(6), 18–27 (2017)
- [2]. Hongsuk Yi, HeeJin Jung and Sanghoon Bae, "Deep Neural Networks for traffic flow prediction," 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, 2017, pp. 328-331, doi: 10.1109/BIGCOMP.2017.7881687.
- [3]. Jyothi, Padma. (2017). A Study on Raise of Web Analytics and its Benefits. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING.* **5.** 61-66.

- [4]. Coursaris, Constantinos & Van Osch, Wietske & López-Nicolás, Carolina & Molina-Castillo, Francisco-Jose & Rapp, Nicolas. (2013). Driving Website performance Using Web Analytics: A Case Study.
- [5]. Shelatkar, Tejas & Tondale, Stephen & Yadav, Swaraj & Ahir, Sheetal. (2020). Web Traffic Time Series Forecasting using ARIMA and LSTM RNN. ITM Web of Conferences. 32. 03017. 10.1051/itmconf/20203203017.
- [6]. S. Mao and F. Xiao, "Time Series Forecasting Based on Complex Network Analysis," in IEEE Access, vol. 7, pp. 40220-40229, 2019, doi: 10.1109/ACCESS.2019.2906268.
- [7]. H. Agarwal, A. Singh and R. D, "Deepfake Detection Using SVM," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1245-1249, doi: 10.1109/ICESC51422.2021.9532627.
- [8]. Huang, Huajuan & Wei, Xiuxi. (2022). An overview on twin support vector regression. Neurocomputing. 490. 10.1016/j.neucom.2021.10.125.
- [9]. Pekel, Engin. (2020). Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology. 139. 10.1007/s00704-019-03048-8.
- [10]. "Efficient Prediction of Network Traffic for Real-Time Applications" by Muhammad Faisal Iqbal, Muhammad Zahid, Durdana Habib, and Lizy Kurian John, 2019.
- [11]. "Efficient Prediction of Network Traffic for Real-Time Applications" by Muhammad Faisal Iqbal, Muhammad Zahid, Durdana Habib, and Lizy Kurian John, 2019.
- [12]. "Modelling Approaches for Time Series Forecasting and Anomaly Detection" by Shuyang Du, Madhulima Pandey, and Cuiqun Xing, 2018.
- [13]. "Fast ES-RNN: A GPU Implementation of the ES-RNN algorithm" by Andrew Redd and Kaung Khin, 2019.
- [14]. "Neural Decomposition of Time-Series Data for Effective Generalization" by Luke B. Godfrey and Michael S. Gashler, 2017.
- [15]. "Web Traffic Prediction of Wikipedia Pages" by Navyasree Petluri, Eyhab Al-Masri, 2019.
- [16]. "Time Series Forecasting Based on Complex Network Analysis" by SHENGZHONG MAO AND FUYUAN XIAO, 2019.
- [17]. "Neural Decomposition of Time-Series Data for Effective Generalization" by Luke B. Godfrey and Michael S. Gashler, 2017.
- [18]. <https://towardsdatascience.com/3-facts-about-timeseries-forecasting-that-surprise-experienced-machinelearning-practitioners-69c18ee89387>.