Web Traffic Time Series Forecasting

Gayatri Gattani
MS Data Science

Rishikesh Kakde
MS Data Science

Yashwanth Vanama
MS Computer Science

December 12, 2023

project-ggattani-rkakde-yasvana

Abstract

The project aims to develop a robust predictive model for forecasting web traffic using data mining and deep learning approaches. Recognizing the dynamic nature of web user behavior, this study addresses the challenge of accurately predicting web traffic, a critical aspect for efficient web resource management and user experience optimization. The methodology encompasses data preprocessing, exploratory data analysis, and the implementation of and deep learning models. The results demonstrate the effectiveness of these models in forecasting web traffic with a significant degree of accuracy.

Keywords

Web Traffic, Time Series Forecasting, Deep Learning, LSTM, ARIMA, Data Preprocessing, Outlier Detection

1 Introduction

In the digital age, where the internet serves as a cornerstone for businesses, media, and communication, understanding and predicting web traffic has become essential. The "Web Traffic Time Series Forecasting" project, aims to tackle the complex challenge of forecasting web traffic patterns using advanced data mining techniques. The primary objective is to develop a predictive model that can reliably forecast the number of future visits to a website. This is crucial for optimizing website performance, managing server capacity, and enhancing the overall user experience.

Web traffic prediction is inherently challenging due to the unpredictable nature of human behavior online. Factors such as seasonal trends, marketing campaigns, and changing user preferences can cause significant fluctuations in web traffic. This project aims to analyze these patterns using a combination of machine learning and deep learning models. By effectively processing and analyzing web traffic data, the goal is to provide accurate, actionable forecasts that can help web service providers to anticipate and prepare for future demand, thereby ensuring smoother operation and better service delivery.

Previous work

For a comprehensive understanding, consider reading

[1] The authors of this paper introduce a deep learning approach to forecast web traffic on Wikipedia, employing LSTM networks. Their model accounts for seasonal trends in time

series data and suggests that the abundance of big data enhances the performance of neural net- works in forecasting tasks. Despite its struggle with unexpected traffic peaks, the proposed LSTM architecture, which includes a unique windowed dataset function and a Conv1D layer, demonstrates potential for accurately predicting web trends.

[2] In this study, the importance of web traffic forecasting has been underscored due to increased internet usage during the pandemic. The paper evaluates the performance of ARIMA and LSTM models, with a focus on a specific ARIMA(4,1,0) instantiation. The authors highlight the challenges in predicting complex web traffic trends and the necessity of preprocessing to address outlier effects.

[3] This research targets the precise forecasting of web traffic to prevent website performance issues, within the context of a Kaggle competition. It details the use of ARIMA models and LSTM networks, emphasizing the significance of feature engineering. The bidirectional LSTM model, complemented by wavelet transforms, was chosen for its ability to handle both past and future data in time series analysis. The paper also discusses the role of statistical analysis in feature engineering and model development.

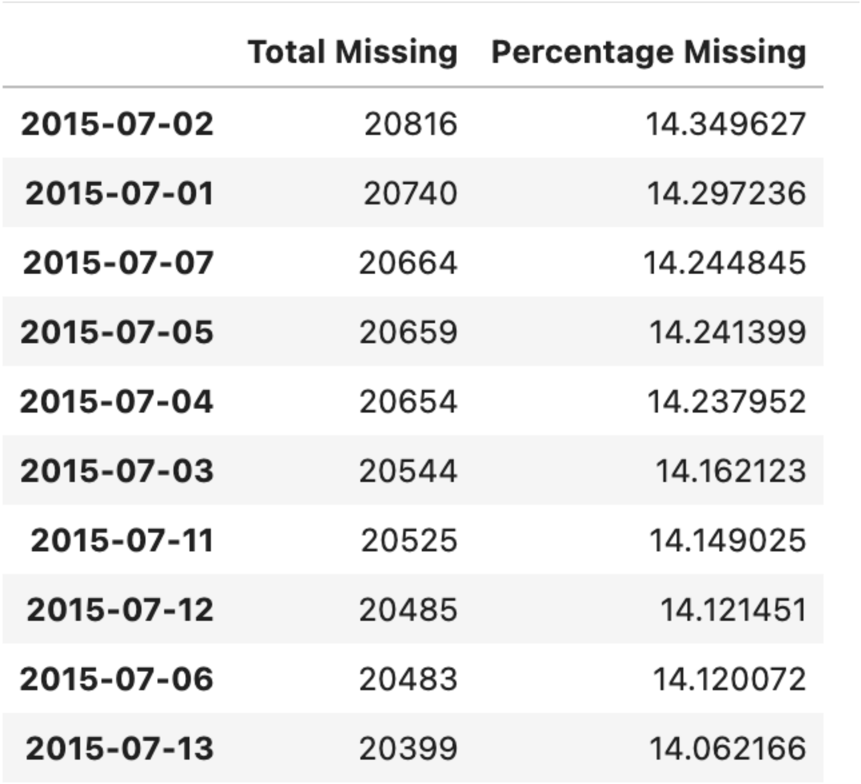
[4] Shelatkar et al. present a hybrid method combining ARIMA models and LSTM RNNs for web traffic forecasting. This approach processes time series data through Discrete Wavelet Transform to separate it into linear and non-linear components. The authors find that this hybrid model surpasses individual ARIMA or LSTM models in capturing traffic trends and spikes, marking a significant step forward in the field.

2 Methods

The project involved data preprocessing, including handling missing data through interpolation and outlier detection via capping. Exploratory data analysis was conducted to understand web traffic trends. The models implemented were ARIMA, for its effectiveness in linear pattern recognition, and LSTM, for its ability to handle non-linear complexities in time series data.

Data Preprocessing

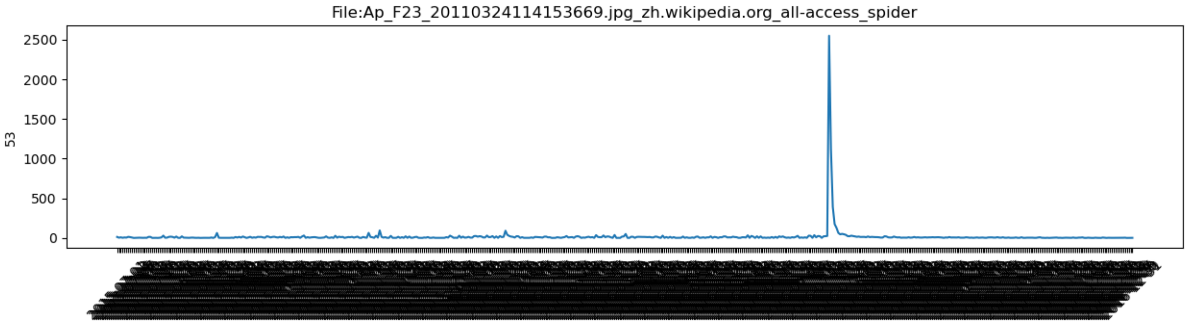
In the data preprocessing stage, a significant focus was placed on handling missing values and outliers, which are critical for maintaining the accuracy and reliability of our forecasting models.

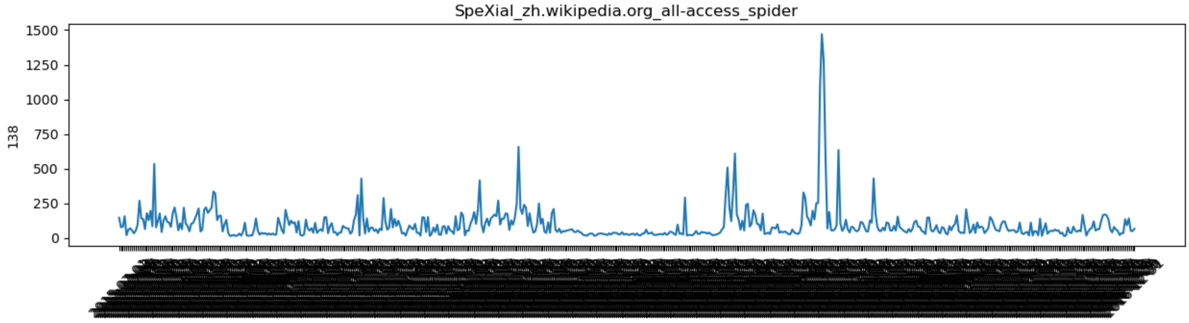


	Total Missing	Percentage Missing
2015-07-02	20816	14.349627
2015-07-01	20740	14.297236
2015-07-07	20664	14.244845
2015-07-05	20659	14.241399
2015-07-04	20654	14.237952
2015-07-03	20544	14.162123
2015-07-11	20525	14.149025
2015-07-12	20485	14.121451
2015-07-06	20483	14.120072
2015-07-13	20399	14.062166

For missing values, we implemented a strategy to identify and interpolate these gaps in the web traffic data. This approach ensured that our analysis and subsequent modeling were based on a complete dataset.

In addition to this, we addressed the challenge of outliers, which can skew the results and lead to misleading forecasts.



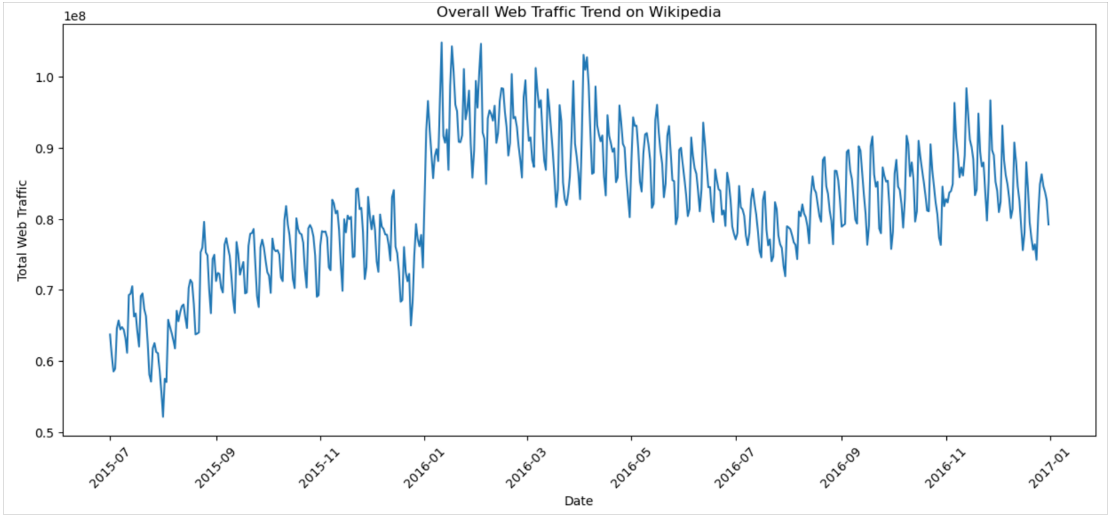


The two visualizations above emphasize the presence of outliers on two web pages. In the case of the first one, the traffic exhibited consistency, with the exception of a brief period during which it surged to a notably high level. As for the second web page, its web traffic activity displayed erratic patterns throughout, with a significant increase observed during a short timeframe.

We employed a capping strategy instead of removal to detect and handle outliers in the traffic data. Outliers are an essential part of forecasting and simply removing them would've affected the ability of the forecasting model to capture seasonal trends. However, having significant outliers also reduces the accuracy of a forecasting model.

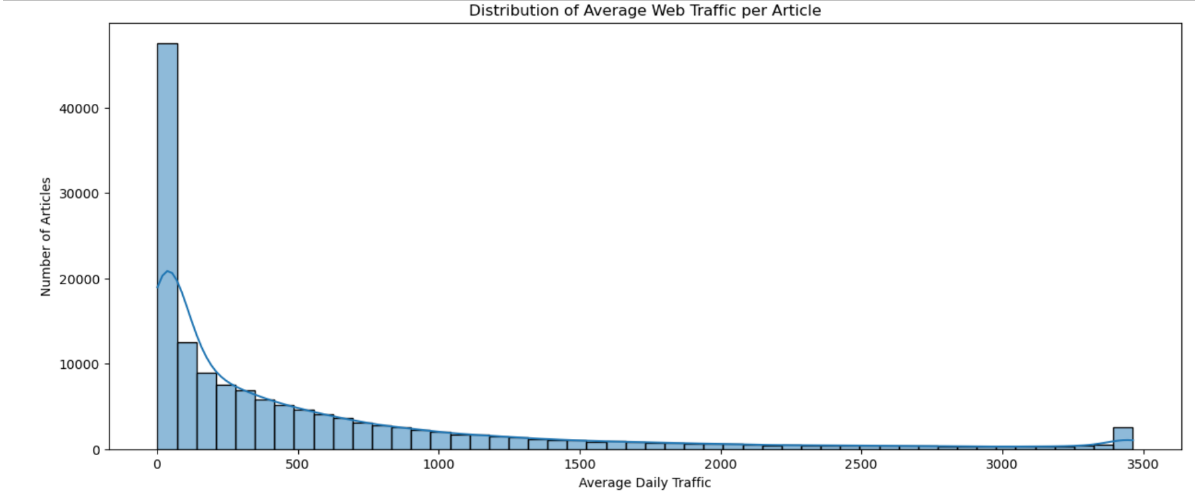
Exploratory Data Anlysis

We plotted the daily and monthly average web traffic, which highlighted the temporal dynamics of website visits. These visualizations were crucial in understanding both short-term fluctuations and long-term trends in web traffic.



As evident from the visualization above, the web traffic lacks a discernible pattern. Initially, the traffic is low, but it experiences an increase for a certain time frame before subsequently declining again for a short while.

Additionally, we calculated and visualized the mean traffic for each article across all days, presenting a distribution that showcased the variability in web traffic among different articles. This comprehensive analysis helped in identifying key characteristics of the data, such as seasonality and peak traffic periods, which were essential for informing our modeling approach.



The distribution of average web traffic per article reveals that the majority of web pages receive very little traffic, while a small number receive a disproportionately high share of the total traffic.

Model Selection and Development

In our project's model selection process, we placed a particular emphasis on two powerful and distinct approaches: the ARIMA (AutoRegressive Integrated Moving Average) and LSTM

(Long Short-Term Memory) models, each offering unique advantages for forecasting web traffic time series data.

ARIMA Model: ARIMA stands out for its ability to model a wide range of time series data with its three key components: autoregression (AR), integration (I), and moving average (MA). The AR part exploits the relationship between a current observation and a number of lagged observations, providing insights into short-term trends. The I component helps in making the data stationary, a crucial step for most time series analysis, by differencing raw observations. Finally, the MA aspect models the error of the prediction, refining the model's accuracy. For our web traffic data, ARIMA offered a straightforward yet effective way to understand and predict the dynamics of web traffic, taking into account its past values and error terms. We fine-tuned the model's parameters, such as the order of differencing and the number of lagged terms, to best capture the inherent patterns in our dataset.

The development of the ARIMA model was a critical step in our project. We meticulously fine-tuned the ARIMA model's parameters - autoregression (AR), integration (I), and moving average (MA) - to best suit our web traffic data. The AR component was calibrated to capture the influence of previous time steps on future traffic, the I component was used to achieve stationarity in the time series, and the MA part was tuned to account for the error in predictions based on historical traffic trends. The calibration process involved iterative testing and validation to ensure the model could effectively capture short-term variations and seasonal trends prevalent in web traffic data. This bespoke development of the ARIMA model allowed us to accurately forecast web traffic, taking into account its past values and the nuances of its error dynamics.

LSTM Model: On the other hand, LSTM, a type of recurrent neural network, is adept at learning order dependence in sequence prediction problems, a common characteristic of time series data like web traffic. Unlike traditional neural networks, LSTMs have a 'memory' capability, allowing them to store past information, which is extremely beneficial for capturing long-term dependencies. This is particularly important in web traffic forecasting, where past traffic trends can significantly influence future patterns. We designed our LSTM model with multiple layers to effectively process the sequential nature of our data, allowing the model to learn from the temporal structure of web traffic. The LSTM model's ability to remember and utilize long-term patterns enabled us to forecast future web traffic with a higher degree of accuracy, especially in capturing complex, time-related trends that simpler models might miss. We designed the LSTM with multiple layers, each layer enhancing the model's ability to remember and learn from the long sequence of data points characteristic of web traffic patterns. The development process involved selecting the optimal number of LSTM layers and tuning other hyperparameters like the number of neurons in each layer and the learning rate. This careful construction of the LSTM model ensured it could effectively process and learn from the temporal structure of web traffic, allowing for a sophisticated understanding and prediction of future trends based on historical data.

By leveraging the strengths of both ARIMA and LSTM, our approach aimed to encapsulate the comprehensive temporal dynamics of web traffic, ensuring a robust and nuanced forecasting model.

3 Results

The figure 1 and 2 images depict ARIMA (AutoRegressive Integrated Moving Average) model forecasts for web traffic on specific Wikipedia pages, differentiated by their URLs and access type (all-access for the first and mobile-web for the second). These graphs compare actual historical data ("Train") with both out-of-sample tests ("Test") and the ARIMA model's forecasted values ("ARIMA Forecast"). The blue line represents the training data used to fit the model, the orange line shows the test data that was not used in model fitting, and the green line represents

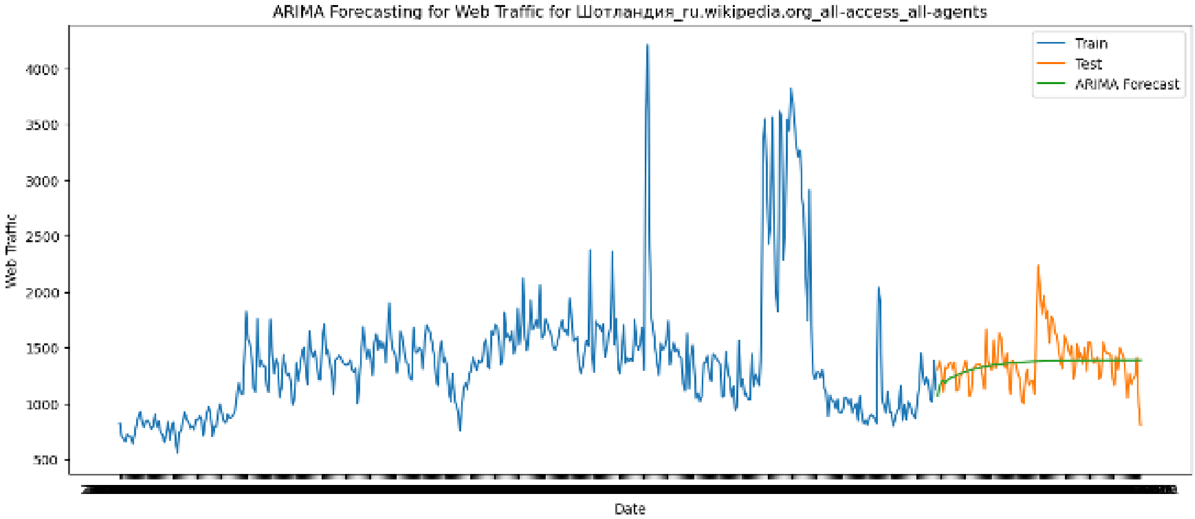


Figure 1: ARIMA forcasting on page 1

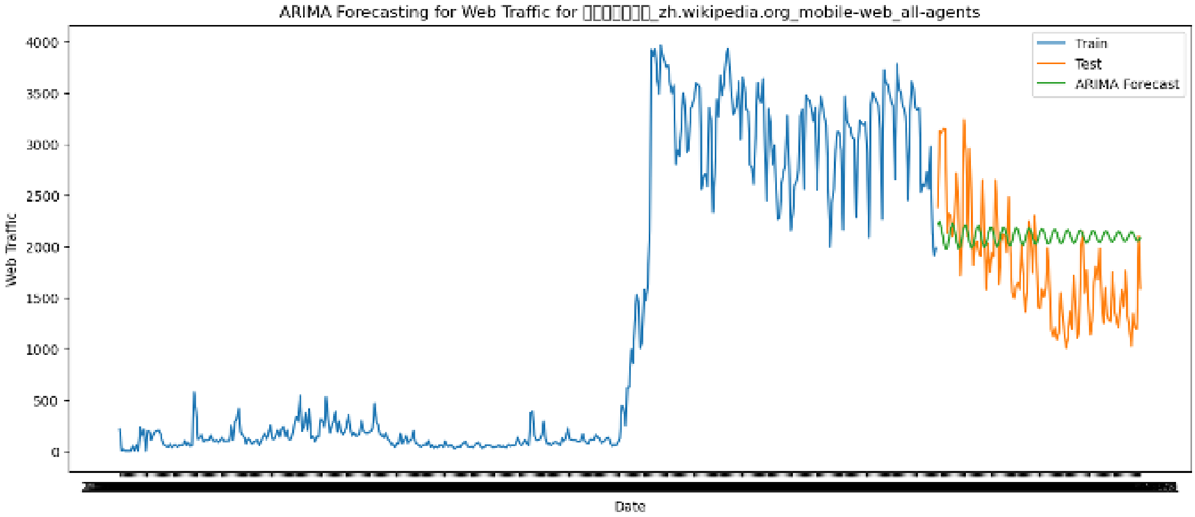
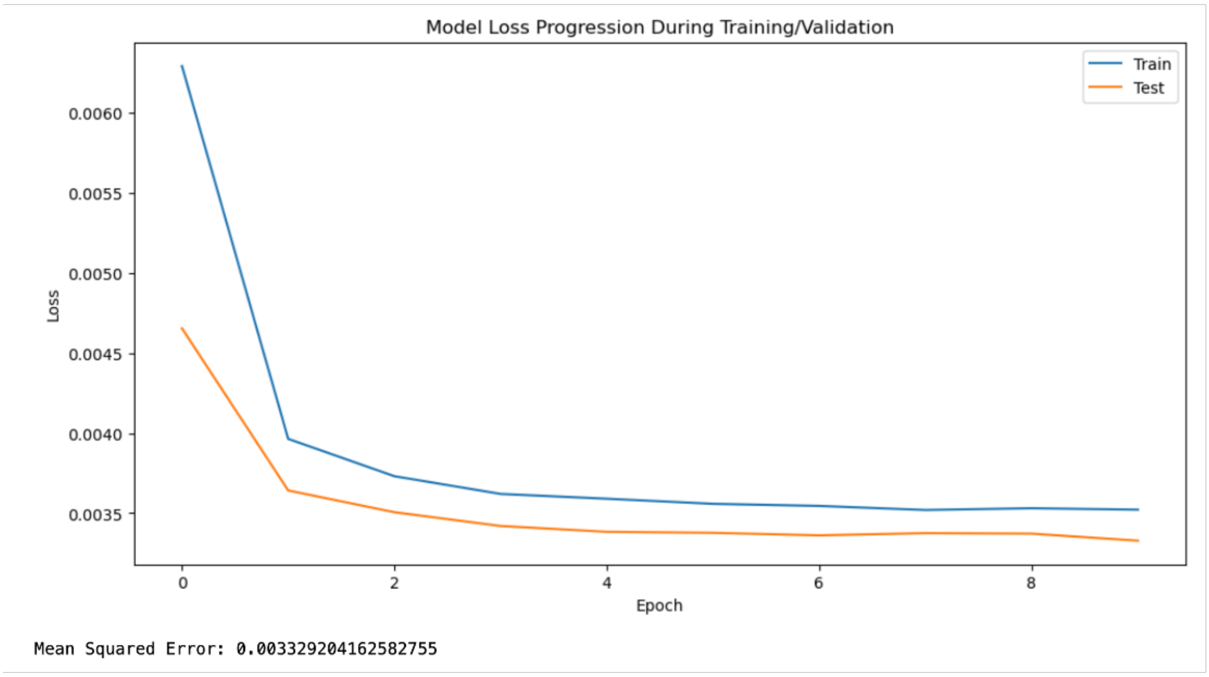


Figure 2: ARIMA forcasting on page 1



```
Mean Squared Error: 0.003329204162582755
```

Figure 3: LSTM loss progression

the forecasted traffic based on the model. The presence of spikes in the data indicates periods of unusually high traffic, which could be due to specific events or anomalies. The forecast shows how the model expects the series to continue beyond the known data, based on the patterns it has learned. Figure 3 shows the training progression of an LSTM (Long Short-Term Memory) model, which is a type of recurrent neural network suited to learning from sequences such as time series data. The graph plots the loss (a measure of how well the model's predictions match the actual data) over multiple epochs (full iterations over the training dataset). The blue line indicates the loss on the training dataset, while the orange line represents the loss on the test dataset. A decreasing loss over epochs suggests that the model is learning effectively. The reported Mean Squared Error (MSE) at the bottom provides a quantitative measure of the model's performance, with lower values indicating better predictive accuracy.

References

[1] Casado-Vara R, Martin del Rey A, Pérez-Palau D, de-la-Fuente-Valentín L, and Corchado JM. Web traffic time series forecasting using lstm neural networks with distributed asynchronous training. 2021.

[2] Clairvoyant Perspectives. Web traffic time series predictions using lstm arima models. 2021.

[3] Prerna. Web traffic time series forecasting — forecast future traffic for wikipedia pages. 2021.

[4] Shelatkar, Tejas, Tondale, Stephen, Yadav, Swaraj, and Ahir, Sheetal. Web traffic time series forecasting using arima and lstm rnn. *ITM Web Conf.*, 32:03017, 2020.