Web Traffic Forecasting Using ARIMA and LSTM

Dr.A.K.MARIAPPAN[1], ANJANA SRIRAM[2], R.NIVEDHA NANDHINI[3], S.V NAGESWARI[4]

[1]Professor, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. maris2612@yahoo.com
[2]Student, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. anjanasriram999@gmail.com
[3] Student, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. nivedhanandhini2000@gmail.com
[4] Student, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. coolnacha99@gmail.com

Abstract - In today's world web traffic is one of the serious issues faced by many. Web traffic tends to hinder the smooth user experience and it is also very challenging for the web service providers to maintain a smooth user-server interaction. We are looking to overcome this problem by building a prediction model to forecast the web traffic in advance to avoid all the problems faced. Our model thoroughly studies the previous web traffic data to efficiently predict the web traffic of a particular website at a given point in time.

Forecasting is one of the important goals of mining time-series databases. The efficacy of Time series forecasting has been proved while decision making in various domains. This method is vastly different from the other proposed methods for prediction and analysis. This paper proposes the use of ARIMA and LSTM algorithms to forecast web traffic.

Keywords – Web traffic, Time series, ARIMA, LSTM, forecasting, Multivariate.

I. INTRODUCTION

The increase in web traffic could pose a hindrance to the workflow and also creates a lot of issues thus an organization is forced to find a way to manage the web traffic efficiently to be successful. Many would have come across extraordinarily slow websites that take an extended time to load once the traffic on their specific website is high, for instance, we may have gone over a few internet businesses sites that may crash when the number surpasses the expected limit. which causes a lot of weight for the customers and on account of that it could reduce the customer's evaluations of that website and the customers will prefer and use more stable websites which might result in the decline in one's trade. Appropriately, a traffic forecasting system is required to plan prior to avoid such unnecessary conditions. Starting in the relatively recent past, the need to predict the traffic did not exist, since the workload was relatively lesser due to the low demand from the customers and accessibility were scarce and uncommon. With less complexity in traffic specialist could handle the traffic allocations without the aid of a prediction system. But since this generation has access to every application the traffic to each website has multiplied and allocations cannot be computed in short period of time by humans since the data is massive.

Time series is a series of real time data that is plotted with respect to time which is uniformly placed. The required data are collected over a period of time. Example of these data could be high-low tide in the ocean, demand for seasonal products, vehicle traffic of a route etc. from such data we can infer meaningful stats that will help us foresee future patterns. From the data graph is generated from which one can understand the trends of the data based on demand, season and requirement. We can understand the characteristics of the data. This information further helps in forecasting future patterns beforehand. This could be a hot exploration point for its various usage in account, development costs expectation, and some different fields.

Ordinarily, we have two well-known sorts of worldly information, first and the acclaimed one is time arrangement information, and the second is information with time focuses. Time arrangement information is a significant class of worldly information protests, an assortment of perceptions, which are in sequential request made. A period arrangement is a succession of noticed information, generally requested on schedule. Inside the composing various techniques are presented for the expecting traffic in the website. These techniques can be separated into two based on the, data being sequential and data having no proper relation between the variables, by which they can be categorized into linear or nonlinear models. The linear models are AR model and MA model. The auto regressive uses the previously plotted hits whereas moving average model uses errors perceived from prior predicts. ARIMA consolidates AR and MA models together, both these models are integrated to for the series to remain constant.

Our proposed approach utilizes Long Short-Term Memory. Adding a chunk of current data to RNN changes the current data by appending a capacity. Subsequently, the entire data is refreshed, i.e., there is no regard for 'significant' data and 'not all that significant' data as a rule. Thus, the RNNs have their present sheet of input circles. RNN licenses them to hold data and information in disk additional time. In any case, it very well may be difficult and hard to prepare standard RNNs to take care of issues requiring long-haul fleeting conditions to comprehend. This is because the misfortune work slope rots dramatically extra time it is known as the issue of the disappearing inclination. LSTM networks are similar to RNN the use not only the standard units but also the uncommon units. LSTM contains a memory cell which can hold information for a long time. The design plays a major role long term. GRU’s are similar to LST. They use a series of doors for data but fail to use memory cells and futile doorways. By using LSTM RNN, the memory is increased when compared to traditional RNN

II. RELATED WORKS

Rodrigo N. Calheiros et al. [1], they present a model which is based on cloud for the prediction of SaaS Suppliers using ARIMA. Their model estimates the accuracy of future work by using traces of real requests of the web servers. The impact of obtained accuracy with respect to the efficiency in utilisation of resources (Qos)is also calculated. It can be found from the cumulative results that their model is able to achieve an accuracy of 91%, which implies that there is a minimum effect on the quality of service and also guarantees the efficiency of resource usage.

G. P., Zhang et al. [2], There model is a hybrid model consisting of both ARIMA and neural networks models. This model takes the advantages of both models hence combining the unique characteristics of both ARIMA and neural networks in linear and non -linear modelling. Sampling, Uncertainty of model, Variation and Change in Structure are some of the factors considered by this model. Testing of this model shows that this model is effective and accuracy can be accomplished by higher than the both models individually.

Tejas Shelatkar et al. [3], this model uses ARIMA and LSTM RNN to predict the web traffic. This model foretells the number of users who might visit the website in the future. As more user data is fed, the more accurate the results will be for this model. The advantages of including LSTM RNN to this model is the increased accuracy to the system; it also effectively records patterns which is the cause for increased efficiency.

Saman Feghhi et al. [4], introduced Associate in Nursing attack on the encrypted net traffic that utilizes solely the packet temporal order information on the transmission. This attack is thus impenetrable to existing packet artifact defences. Likewise, in distinction to existing approaches, this timing-only attack doesn't like the knowledge on the beginning or finish of the net fetches and then is effective against traffic streams. we have a tendency to exhibit the effectiveness of the attack against the wired and wireless traffic, accomplishing average success rates of ninetieth. Likewise, this timing-only attack serves to stress deficiencies within the already gift defences and additionally to the areas wherever it might be helpful for virtual non-public network (VPN) designers to concentrate their additional attention.

Rishabh Madan et al. [5], they have proposed a system which predicts web traffic based on its history. This model uses many forecasting algorithms like ARIMA and is suitable for linear time series datasets. In contrast algorithms like RNN are equipped to work with datasets which are non-linear. This presented model uses DWT (Discrete Wavelet Transform), high pass filter and a low pass filter thus being able to work with both linear and non-linear datasets thereby proving to be more efficient than RNN and ARIMA when considered individually.

Navyasree Petluri et al. [6] they have proposed a system to predict the web traffic of wikipedia. They do this by considering the already existing traffic data of the wikipedia page. By predicting the web traffic in advance, we enable effective load balancing and clear understanding of client behaviour. This model is built using RNN seq2seq. This uses SMAPE for the measurement of the accuracy of the built model. Finally, the predicted data is compared with the real data to determine the efficacy of the proposed model in predicting the future traffic of the wikipedia site.

Seyyed Meysam et al. [7], This paper proposes a model to deal with the detection of DoS and DDoS attacks. This model uses two features, one is the IP address of the source and the other is the number of packets that are calculated per minute. Box-Cox transformation is used to create a time series based on the number of packets. Thus, by classifying the chaotic behaviour and by using Lyapunov exponents we can identify normal traffic from attack traffic. This system is proved to be 99.5 percent efficient.

Soheila Mehrmolaei et al. [8] this paper proposes a model that is based on the duration of the forecasting thereby dividing them into two groups. Average estimation of error for time series forecasting is presented in this model thus proving to be more efficient than the original ARIMA model.

III. PROPOSED SYSTEM

At the point when the quantity of hits increments past the limit of a site, it will in general crash consequently making a gigantic misfortune for an organization. To evade this, we have come up with a prediction model which predicts the web traffic ahead of time with the goal that the necessary server can be assigned well ahead of time subsequently forestalling the event of an accident. This model aids in overseeing and limiting accidents adequately which thus forestalls the deficiency of an Organization. Our Model is a half and half multivariate model as our model is assembled utilizing both ARIMA and LSTM which radically builds the proficiency of individual calculations. ARIMA is best with linear data and LSTM is best with non-linear data. Our model is subsequently acceptable with the two kinds of reports. The yield of the ARIMA is given as input to the LSTM consequently training the dataset twice and thus acquiring better outcomes.

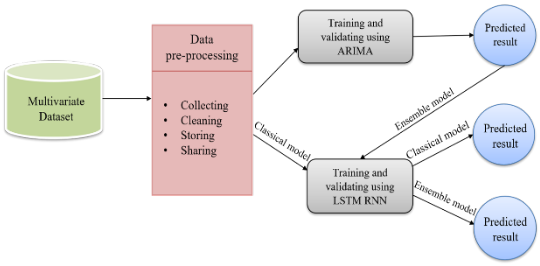


Fig (1) System design

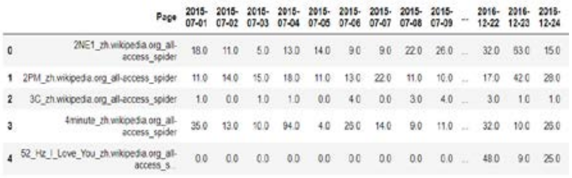
Module wise implementation:

- Pre-processing web traffic time series dataset.
- Training and validating with ARIMA
- Training and validating with LSTM-RNN
- Combining ARIMA and LSTM-RNN
- Deploying the forecasted model

Pre-processing web traffic time series dataset:

Data preprocessing is the way toward changing over raw data into a piece of machine-justifiable data. Raw data comprises of more insufficient information, and it contains additional missing values. Data preprocessing includes four steps, collecting, cleaning, storing and sharing.

As the first step, the dataset adopted for this project is daily views of Wikipedia articles provided by Kaggle comprising roughly 145,000 records. The dataset involves two fields, date and page. Page field comprises more than 1 lakh Wikipedia articles and the date field shows the number of hits day by day. The next step involves data cleaning, the gathered Wikipedia dataset may contain some missing values and this process fills the missing values with zero and organises the raw data for the following steps. When the dataset is cleaned and stacked, ensure that the dataset stored is right, prior to continuing further.



	Page	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	2015-07-08	2015-07-09	...	2016-12-22	2016-12-23	2016-12-24
0	2NE1_zh.wikipedia.org_all-access_spider	18.0	11.0	5.0	13.0	14.0	9.0	9.0	22.0	26.0	...	32.0	63.0	15.0
1	2PM_zh.wikipedia.org_all-access_spider	11.0	14.0	15.0	18.0	11.0	13.0	22.0	11.0	10.0	...	17.0	42.0	28.0
2	3C_zh.wikipedia.org_all-access_spider	1.0	0.0	1.0	1.0	0.0	4.0	0.0	3.0	4.0	...	3.0	1.0	1.0
3	4minute_zh.wikipedia.org_all-access_spider	35.0	13.0	10.0	94.0	4.0	26.0	14.0	9.0	11.0	...	32.0	10.0	26.0
4	52_Hz_I_Love_You_zh.wikipedia.org_all-access_s...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	48.0	9.0	25.0

Fig (2) Wikipedia Article

The refined dataset in its raw state cannot be supplied to the deep learning model which could lead to mis-leading prediction or an accuracy deficient model. To overcome these data-oriented challenges, the data set has to be segregated into a coarse-grained variant of itself which not only reduces the complexity of the dataset but also provides a clear and segregated view of data. This stage is called as data framing and in this stage, the dataset used to train the model is transposed from its existing state and classified based on the linguistic preferences of the user (hits based on user's language preferences). Whence the dataset is segregated, the model is fed with the framed structure which initiates the training process. The model plots the graphical representation of the segregated dataset which serves as the precursor for the whole prediction model which will be later put forth on the task of forecasting the future state of plausible user hits the target site / server could get at the predicted time frame.

Training and validating using ARIMA:

ARIMA stands for Auto-Regressive Integrated Moving Average is extraordinary compared to other time series models for the linear dataset. This model explains a time series dependent from its previous dataset, that, depends on its past lags and lagged errors. With this, the future values are anticipated. ARIMA is isolated into three sections:

- Autoregressive (AR) forecasts future outcome from the past value.
- Integrated (I), it has to do with the distinction in time arrangement.
- Moving average (MA) model doesn't utilize the previous estimates to anticipate the future qualities though it utilizes the blunders from the past result.

Before training the data frames ARIMA model goes through different steps and one of them is its plot autocorrelation function (ACF) and partial autocorrelation (PACF) to distinguish the potential MA and AR model.

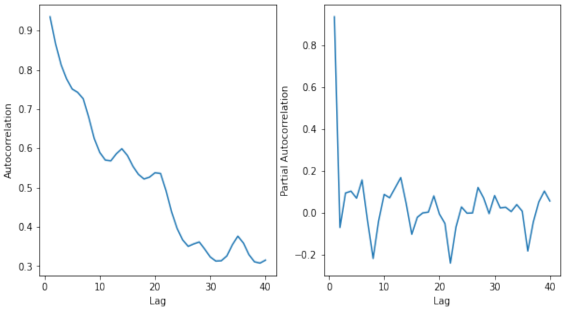


Fig (3) ACF and PACF

Based on the ACF and PACF values, the best ARIMA fir model is found for training and validating. In this manner, the forecast outcome is generated and plotted with the best model.

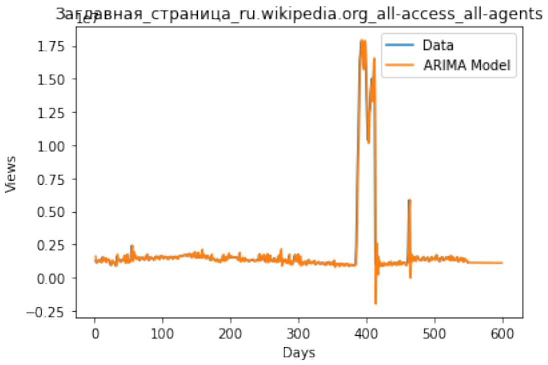


Fig (4) ARIMA model

ARIMA model at times can anticipate the week-by-week base of the sign, which is acceptable. In different cases, it appears to simply give a direct fit. This is conceivably valuable.

Be that as it may, on the off chance that we just aimlessly apply the ARIMA model to the entire dataset, the outcomes are not close to the same as utilizing the basic models. It actually appears to make them interesting properties, so perhaps we can consolidate this with another model to improve results.

The performance of ARIMA is evaluated by Mean Absolute Percentage Error (MAPE) is 15.704892.

Training and validating using LSTM:

LSTM represents Long short-term memory is a self-supervised learning method, it is appropriate for both univariate and multivariate dataset. In this project, the multivariate dataset is utilized. A multivariate dataset implies where there is more than one field to forecast.

Subsequent to pre-processing, the Wikipedia article is divided for validating and training. Converting the separated Wikipedia article into NumPy array and reshaped the array (3D) to which the LSTM model accepts. At that point construct the LSTM design. Fabricated model train and test the dataset for evaluating the performance.

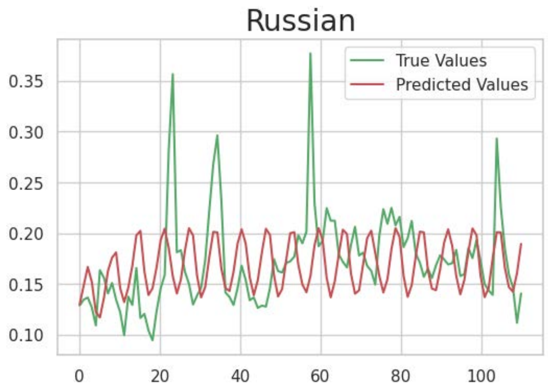


Fig (5) LSTM model

The performance of LSTM is evaluated by Mean Absolute Percentage Error (MAPE) is 35.686939.

Combining ARIMA and LSTM:

The over two models are fitted depending on their best. The ARIMA model suits well for the linear dataset and LSTM works out positively for the non-linear dataset. As an individual outcome, both show their anticipated worth with some flaws. For better precision and result, the two models are ensemble together.

As the blend of both the models, from the outset ARIMA model's yield is given as the contribution of the LSTM model. Along these lines the dataset is trained twice. With this mix, the precision level expanded and the rate blunder decreased. MAPE score for the ensembled model is 7.862425, lower than ARIMA (15.704892) and LSTM MAPE score (35.686939) exclusively.

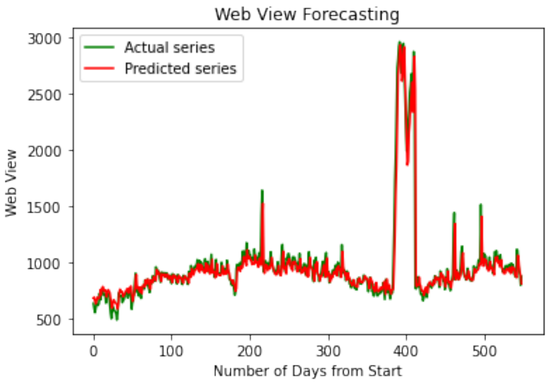
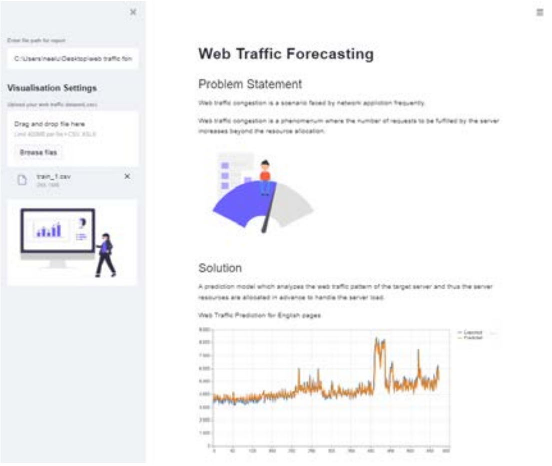


Fig (6) Ensemble Model

Deploying the forecasted model:

When the models are prepared and approved with ARIMA and LSTM, the forecasted model is incorporated with a website. The application is created with streamlit which is an open-source structure for deploying ML models.

The customer can discover their site traffic by uploading the past hits (ensure it is a multivariate dataset). The customer can see their traffic in graphical portrayal and can download the traffic record as a text file.



Enter file path for report

C:\Users\neelu\Desktop\web traffic fon

Visualisation Settings

Upload your web traffic dataset(.csv)

Drag and drop file here

Limit 400MB per file • CSV, XSLX

Browse files

train_1.csv

Web Traffic Forecasting

Problem Statement

Web traffic congestion is a scenario faced by network application frequently.

Web traffic congestion is a phenomenum where the number of requests to be fulfilled by the server increases beyond the resource allocation.

Solution

A prediction model which analyses the web traffic pattern of the target server and thus the server resources are allocated in advance to handle the server load.

Web Traffic Prediction for English pages

Fig (7) Website

IV. EXPERIMENTAL RESULT

The performance evaluation for the time series forecasting model is calculated with the underneath methods,

a) Mean Forecast Error (MFE)
b) Mean Squared Error (MSE)
c) Mean Absolute Error (MAE)
d) Root Mean Squared Error (RMSE)
e) Mean Absolute Percentage Error (MAPE)

Methods	ARIMA	LSTM	Hybrid
MFE	0.000204	0.01451	0.00563
MAE	0.188291	0.09685	0.08114
MSE	0.054247	0.02006	0.02240
RMSE	0.232909	0.14164	0.11493
MAPE	15.704892	35.6869	7.86242

TABLE I: Error Metrics

V. CONCLUSION

Web traffic is a major issue these days. It makes sites crash impeding the smooth client experience, subsequently making a difficult issue for the organization. To take care of this issue, A web traffic forecasting model is fabricated utilizing ARIMA and LSTM, which proficiently predicts the web traffic in advance, and thereby the server can be allocated based on the requirement and numerous issues identified with web traffic can be addressed. We have created a website whereby uploading the previously obtained traffic as a CSV file, the website predicts the web traffic.

VI. FUTURE WORKS

We have created a website whereby uploading the previously obtained traffic as a CSV file, the website predicts the web traffic. Future work would be to implement this as a plug-in, by using it, the service provider can get the predicted traffic in an instant.

References:

1.“Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications’ QoS” by Rodrigo N. Calheiros, Enayat Masoumi, Rajiv Ranjan, Rajkumar Buyya, 2014.

2."Time series forecasting using a hybrid ARIMA and neural network model", by G. P. Zhang, 2003.

3. “Web Traffic Time Series Forecasting using ARIMA and LSTM RNN” by Tejas Shelatkar, Stephen Tondale, Swaraj Yadav, Sheetal Ahir, 2020.

4. “A Web Traffic Analysis Attack Using Only Timing Information” by Saman Feghhi and Douglas J. Leith, August 2016.

5. “Predicting Computer Network Traffic: A Time Series Forecasting Approach Using DWT, ARIMA and RNN” by Rishabh Madan, Partha Sarathi Mangipudi.

6. “Web Traffic Prediction of Wikipedia Pages” by Navyasree Petluri, Eyhab Al-Masri, 2019.

7. “A Novel DoS and DDoS Attacks Detection Algorithm Using ARIMA Time Series Model and Chaotic System in Computer Networks” by Seyyed Meysam Tabatabaie Nezhad, Mahboubeh Nazariy, and Ebrahim A. Gharavo, 2015.

8. “Time series forecasting using improved ARIMA” by Soheila Mehrmolaei,2016.