

Forecasting Website Traffic Using Prophet Time Series Model

A. Subashini¹ K. Sandhiya^{2*}, S. Saranya², U. Harsha²

¹Assistant Professor, Department of Computer Science Engg, United Institute of Technology, Coimbatore, TN, India

²UG Student, Department of Computer Science Engg, United Institute of Technology, Coimbatore, TN, India

*Corresponding author E-Mail ID: sandhiya486@gmail.com, Mobile: +91 8940403464.

DOI: <https://doi.org/10.34256/irjmt1917>

ABSTRACT

Web traffic is the amount of data sent and received by visitors to a website and it has been the largest portion of Internet traffic. Internet traffic flow prediction heavily depends on historical and real-time traffic data collected from various internet flow monitoring sources. With the wide spread traditional traffic sensors and new emerging traffic sensor technologies, traffic data are exploding, and we have entered the era of big data internet traffic. Internet traffic management and control driven by big data is becoming a new trend. Although there have been already many internet traffic flow prediction systems and models, most of which use shallow traffic models and are still somewhat unsatisfying. This inspires us to reconsider the internet traffic flow prediction model based on deep architecture models with such rich amount of internet traffic data. ARIMA is a existing forecasting technique that predicts the future values of a series based entirely on its own inertia. Existing traffic flow prediction methods mainly use simple traffic prediction models and are still unsatisfying for many real-world applications. Now we proposed the prophet time series model to forecasting website traffic.

Keywords: Website Traffic, ARIMA, Prophet, Deep learning.

1. INTRODUCTION

Web sites must forecast Web page views in order to plan computer resource allocation and estimate upcoming revenue and advertising growth. This necessarily does not include the traffic generated by bots. Since the mid-1990s, web traffic has been the largest portion of Internet traffic. This is determined by the number of visitors and the number of pages they visit. Sites monitor the incoming and outgoing traffic to see which parts or pages of their site are popular and if there are any apparent trends, such as one specific page being viewed mostly by people in a particular country. There are many ways to monitor this traffic and the gathered data is used to help structure sites, highlight security problems or indicate a potential lack of bandwidth. Not all web traffic is welcomed. Some companies offer advertising schemes that, in return for increased web traffic (visitors), pay for screen space on the site. There is also "fake traffic", which is bot traffic generated by a third party. Common business analytics task is trying to forecast the future based on known historical data. Forecasting is a complicated topic and relies on an analyst knowing the ins and outs of the domain as well as knowledge of relatively complex mathematical theories. Because the mathematical concepts can be complex, a lot of business forecasting approaches are "solved" with a little linear regression and "intuition." More complex models would yield better results but are too difficult to implement.

2. LITERATURE REVIEW

A deep-learning neural-network based on Tensor Flow is suggested for the prediction traffic flow conditions, using real-time traffic data. The suggested supervised model is trained by a deep learning algorithm, which uses real traffic data aggregated every five minutes. Results demonstrate that the model's accuracy rate is around 99%. Research shows the potential for Tensor Flow deep learning models for the accurate analysis of real-time traffic data, and precise estimation of traffic flow conditions. There are, however, several limitations in this research; due to memory capacity limitations for example, only 1 percent of traffic data for a given day could be used. In order to increase accuracy of estimation and improve the DNN architecture, it is now necessary to redefine the TPI according to transportation engineering knowledge. [1] Internet traffic analysis, models simulation and prediction play a very important part in the network management and design. Here present the modified LSL algorithm, which can modify the parameters with each new input data update and has the property of fast convergence and high accuracy. Time series models are often applied to simulate network traffic today, but they are seldom related with the adaptive ability of the model. By far, most of models have not been considered to possess the adaptive ability. The common method is using the input data to calculate the model coefficient, while the model coefficient cannot be updated dynamically. The goal is to improve the model adaptive ability and make the model parameters update dynamically. Consequently, we bring the new adaptive thought in network traffic model and prediction, and propose a new algorithm to modify the model coefficient continuously. CERNET is a large network, if we want to solve the network management and network design problems, we have to do some work about building network traffic model and prediction. [2]. The accurate short-term traffic flow forecasting is fundamental to both theoretical and empirical aspects of intelligent transportation systems deployment. In order to play the ARIMA model with good linear fitting ability and artificial neural network model with strong nonlinear relation mapping ability, this study aimed to develop a simple and effective hybrid model for forecasting traffic volume that combines the Autoregressive Integrated Moving Average (ARIMA) and the Radial Basis Function Artificial Neural Networks (RBF-ANN) models. By combining different models, different aspects of the underlying patterns of traffic flow could be captured. [3]. Accurate and timely internet traffic information is important for many applications, such as bandwidth allocation, anomaly detection, congestion control and admission control. Over the last few years, internet flow data have been exploding, and we have truly entered the era of big data. Existing traffic flow prediction methods mainly use simple traffic prediction models and are still unsatisfying for many real-world applications. This situation inspires us to rethink the internet traffic flow prediction problem based on deep architecture models with big traffic data. Here they propose a novel deep-learning based internet traffic flow prediction method, which is called SDAPM and also effective. [4]. Fast and accurate methods for predicting traffic properties and trend are essential for dynamic network resource management and congestion control. With the aim of performing online and feasible prediction of network traffic, this paper proposes a novel time series model, named adaptive autoregressive (AAR). This model is built upon an adaptive memory-shortening technique and an adaptive-order selection method originally developed by this study. Compared to the conventional one-step ahead prediction using traditional Box-Jenkins time series models (e.g. AR, MA, ARMA, ARIMA and ARFIMA), performance results obtained from actual Internet traffic traces have demonstrated that the proposed AAR model is able to support online prediction of dynamic network traffic with reasonable accuracy and relatively low computation complexity.[5]

3. METHODOLOGY

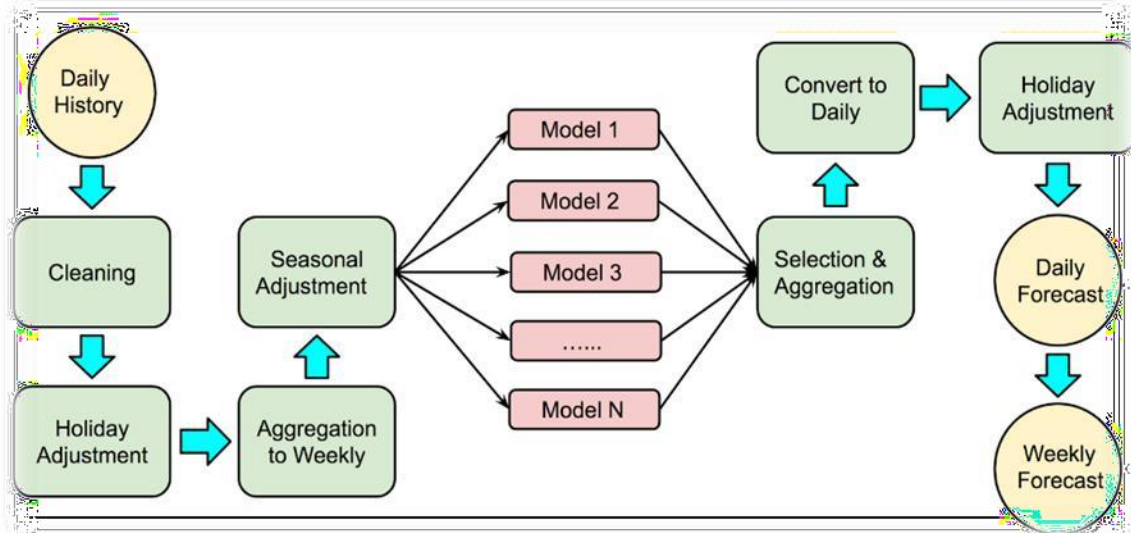


Fig 1. Process in Forecasting Website Traffic

A common business analytics task is trying to forecast the future based on known historical data. Forecasting is a complicated topic and relies on an analyst knowing the ins and outs of the domain as well as knowledge of relatively complex mathematical theories. Because the mathematical concepts can be complex, a lot of business forecasting approaches are “solved” with a little linear regression and “intuition.” More complex models would yield better results but are too difficult to implement. In this article, I’ll introduce prophet time series model and show how to use it to predict the volume of traffic in the next year in any business analytics.

4. ANALYSIS USING ARIMA MODEL

In an Existing system, ARIMA model have been used. It was introduced by Box and Jenkins (1976). ARIMA stands for Autoregressive Integrated Moving Average models. It is a forecasting technique that predicts the future values of a series based entirely on its own inertia. The essence of the model is that the non-stationary time series sequence by differential transforms method. There are few problems in this system. Some of the problems with traditional time series model are,

- Time interval between data has to be same throughout the data
- Day with NA is not allowed
- Seasonality with multiple periods(Week and Year) is hard to handle
- Parameter tuning by expert is necessary

5. PROPHET FORECASTING MODEL

When a forecasting model doesn’t run as planned, we want to be able to tune the parameters of the method with regards to the specific problem at hand. Tuning these methods requires a thorough understanding of how the underlying time series models work. The first input parameters to automated ARIMA, for instance, are the maximum orders of the differencing, the

auto-regressive components, and the moving average components. A typical analyst will not know how to adjust these orders to avoid the behavior and this is the type of expertise that is hard to acquire and scale.

The Prophet Forecasting provides intuitive parameters which are easy to tune. Even someone who lacks expertise in forecasting models can use this to make meaningful predictions for a variety of problems in a business scenario. It uses a decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- $g(t)$: piecewise linear or logistic growth curve for modeling non-periodic changes in time series.
- $s(t)$: periodic changes (e.g. weekly/yearly seasonality).
- $h(t)$: effects of holidays (user provided) with irregular schedules.
- ϵ_t : error term accounts for any unusual changes not accommodated by the model.

Using time as a regressor, Prophet is trying to fit several linear and non linear functions of time as components. Modeling seasonality as an additive component is the same approach taken by exponential smoothing in Holt-Winters technique .

5.1 TREND

Trend is modeled by fitting a piece wise linear curve over the trend or the non-periodic part of the time series. The linear fitting exercise ensures that it is least affected by spikes/missing data.

Table 1. Trend Parameters

Parameter	Description
Growth	linear' or 'logistic' to specify a linear or logistic trend
Changepoints	List of dates at which to include potential changepoints (automatic if not specified)
n_changepoints	If changepoints in not supplied, you may provide the number of changepoints to be automatically included
Changepoint_prior_scale	Parameter for changing flexibility of automatic change point selection

5.2 SATURATING GROWTH

Let's say we are trying to forecast number of downloads of an app in a region for the next 12 months. The maximum downloads is always capped by the total number of smartphone users in the region. The number of Smartphone users will also, however, increase with time. With domain knowledge at his/her disposal, an analyst can then define a varying capacity $C(t)$ for the time series forecasts he/she is trying to make.

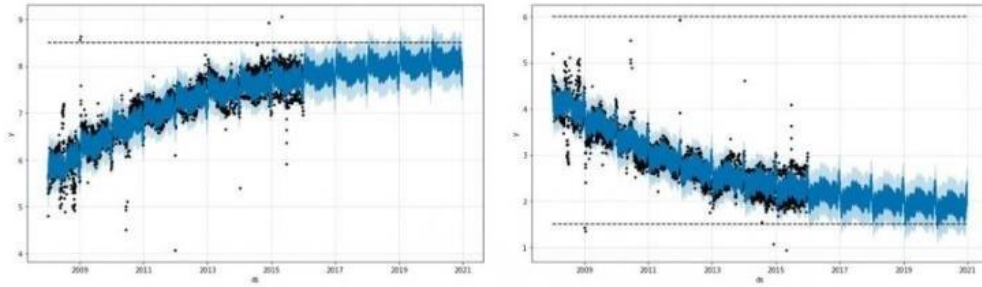


Fig 2. Saturation Growth in Periods

5.3 CHANGE POINTS

In a time series encounters any underlying changes in the phenomena e.g. a new product launch, unforeseen calamity etc. At such points, the growth rate is allowed to change. These changepoints are automatically selected. However, a user can also feed the changepoints manually if it is required. In the below plot, the dotted lines represent the changepoints for the given time series.

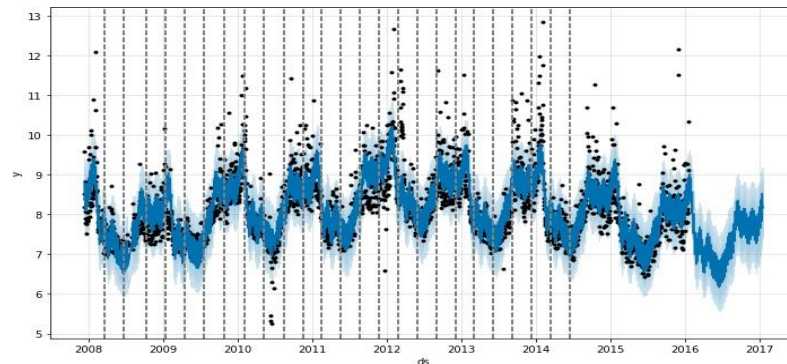


Fig 3. Change Points in Periods

As the number of changepoints allowed is increased the fit becomes more flexible. There are basically 2 problems an analyst might face while working with the trend component:

- Over fitting
- Under fitting

A parameter called `changepoint_prior_scale` could be used to adjust the trend flexibility and tackle the above 2 problems. Higher value will fit a more flexible curve to the time series.

5.4 SEASONALITY

To fit and forecast the effects of seasonality, prophet relies on Fourier series to provide a flexible model. Seasonal effects $s(t)$ are approximated by the following function:

$$s(t) = \sum_{n=1}^N \left(a_n \cos \left(\frac{2\pi nt}{P} \right) + b_n \sin \left(\frac{2\pi nt}{P} \right) \right)$$

P is the period (365.25 for yearly data and 7 for weekly data)

Parameters $[a_1, b_1, \dots, a_N, b_N]$ need to be estimated for a given N to model seasonality.

The Fourier order N that defines whether high frequency changes are allowed to be modeled is an important parameter to set here. For a time series, if the user believes the high frequency components are just noise and should not be considered for modeling, he/she could set the values of N from to a lower value. If not, N can be tuned to a higher value and set using the forecast accuracy.

5.5 HOLIDAYS AND EVENTS

Holidays and events incur predictable shocks to a time series. For instance, Diwali in India occurs on a different day each year and a large portion of the population buy a lot of new items during this period.

Prophet allows the analyst to provide a custom list of past and future events. A window around such days is considered separately and additional parameters are fitted to model the effect of holidays and events.

Table 2. Seasonality & Holiday Parameters

Parameter	Description
Yearly_seasonality	Fit yearly seasonality
Weekly_seasonality	Fit weekly seasonality
Daily_seasonality	Fit daily seasonality
Holidays	Feed dataframe containing holiday name and date
Seasonality_prior_scale	Parameter for changing strength of seasonality model
Holiday_prior_scale	Parameter for changing strength of holiday model

6. STARTING THE ANALYSIS

Training data sets can be downloaded from google analytics.

<https://raw.githubusercontent.com/jroakes/google-analytics/master/examples/holidays.csv>

```
import pandas as pd
import numpy as np
from fbprophet import Prophet

data_file = "All Web Site Data Audience Overview.xlsx"
df = pd.read_excel(data_file)
df.head()
```

The data set from 2012 to 2020 can be taken, to predict and forecasting website traffic.

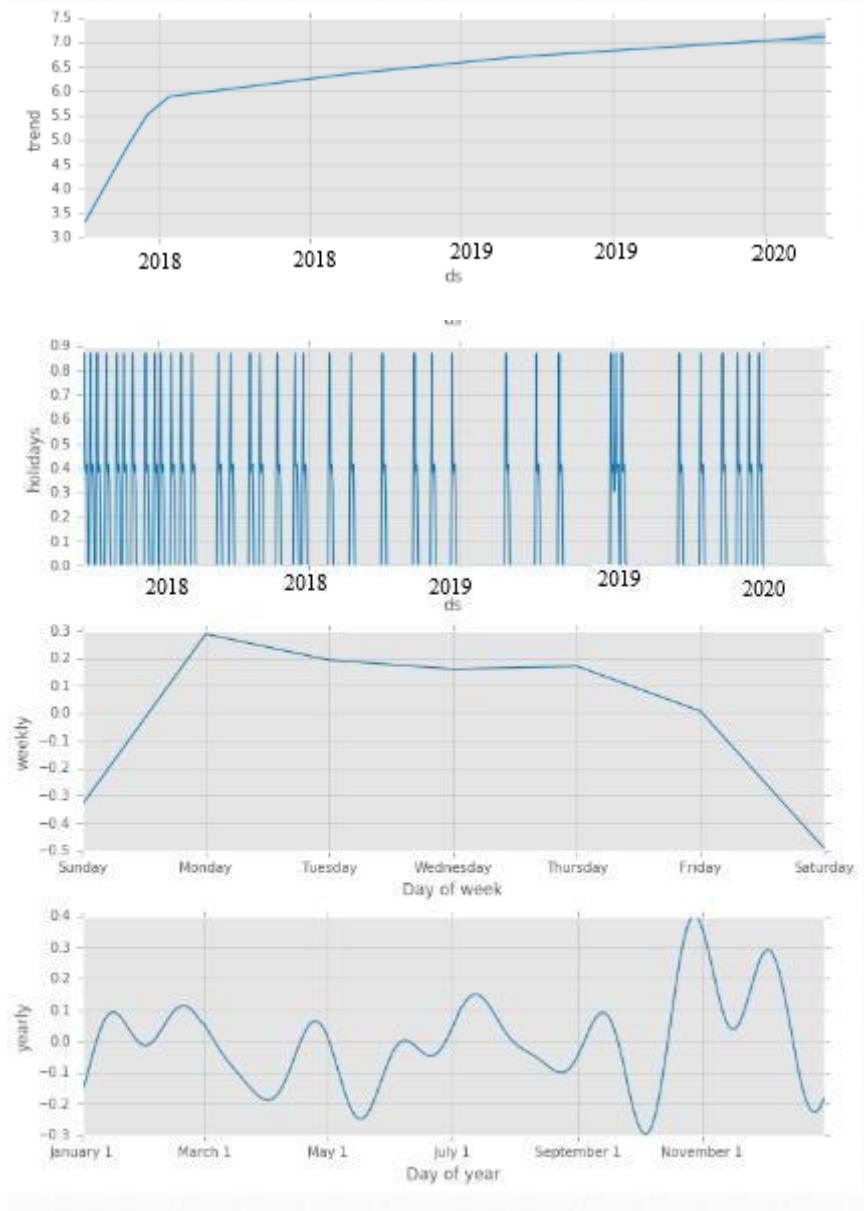


Fig 4. Implementation

7. CONCLUSION

Prophet certainly is a good choice for producing quick accurate forecasts. It has intuitive parameters that can be tweaked by someone who has good domain knowledge but lacks technical skills in forecasting models. One of the features that prophet supports is the concept of a “holiday.” The simplest way to think about this idea is the typical up-tick in store sales seen around the Thanksgiving and Christmas holidays these data points try to make better future predictions. The API is relatively simple and since it uses the standard panda’s dataframe and matplotlib for displaying the data, it fits very easily into the python data science workflow.

REFERENCES

- [1] Hongsuk Yi, HeeJin Jung, Sanghoon Bae “Deep Neural Networks for Traffic Flow Prediction” in IEEE 978-1-5090-3015-6/17.
- [2] Jun Lv ,Xing Li ,Tran Quang Anh ,Tong Li “A New Algorithm for Network Traffic Prediction” in Proceedings of the 11th IEEE Symposium on Computers and Communications, 0-7695-2588-1/06.

- [3] Weitao Wang, Yuebin Bai, Chao Yu, Yuhao Gu, Peng Feng, Xiaojing Wang, and Rui Wang, “A Network Traffic Flow Prediction with Deep Learning Approach for Large-scale Metropolitan Area Network” in IEEE, 978-1-5386-3416-5/18.
- [4] Hao Yin, Chuang Lin, Berton Sebastien , Network traffic prediction based on a new time series model in International Journal of communication systems , 2005-18:711–729.
- [5] Deep Learning Tutorials, <http://deeplearning.net/tutorial>.
- [6] Deep Learning Software, http://deeplearning.net/software_links.
- [7] V Paxson, S, Floyd. Wide-area traffic: the failure of poisson modeling [A], Proc of the ACM/SIGCOMM’94[A].1994.257-268.
- [8] W.E Legend, M.S.Taqqu, “On the self-similar nature of Ethernet traffic” IEEE/ACM transactions.
- [9] Van Jacobson. Congestion avoidance and control. ACM SIGCOMM ’88, Stanford, CA, August 1988; 314–329.
- [10] Tensorflow, <https://www.tensorflow.org>.
- [11] Paxson V, Floyd S. Wide area traffic: the failure of Poisson modeling. IEEE/ACM Transactions on Networking 1995;3(3):226–244.

Conflict of Interest

None of the authors have any conflicts of interest to declare.

About the License

The text of this article is licensed under a Creative Commons Attribution 4.0 International License