Forecasting Web Traffic

Diana Tsai 20567329

December 10, 2018

Abstract

As the Internet becomes widely used, websites have become an important method of reaching out to potential audiences. Looking at trends of web traffic has become popular amongst marketers, content creators, businesses and hosting companies. Many of them use web analystics to track web traffic volume. The next step is to forecast web traffic based on historical data collected. To do this, we will introduce the idea of using the AR, MA and ARMA models and attempt to forecast the number of visits for the next 10 days. The results of the model selection then implies that the data[1] is best modelled by ARMA(1, 1), which has the lowest AIC, meaning it only uses the prior day's traffic volume. We then use this model to forecast the next 10 days.
Keywords: Time Series, Forecasting, Web Traffic, AR, MA, ARMA
JEL Code: C53, Z00

1 Introduction

The Internet has become an essential component to people's lives, inducing more poeple use web analytics to collect data and understand how to optimize web usage. One metric of interest is the amount of we traffic a site receives over time. The forecasted web traffic is then of great interest to hosting companies, content creators and market researchers. With historical data on how well websites are doing based on web traffic, we can forecast the amount of traffic it will have in the future. Hosting companies are interested in web traffic forecasting so they are able to anticipate how much load they may encounter. Content creators or market researchers would be able to see how well their campaigns or content is doing so they can gauge popularity trends in order to improve. Another use for forecasting web traffic would be to help advertisers selecting which web page gives them the most views.

To perform our forecasting, we use time series models, which have been used in numerous real-world problems such as forecasting the performance of stocks in the financial markets and weather. Their ability to predict trends over time allows it to be effectively applied for web traffic forecasting.

[1]Credit to Kaggle for providing the data on web traffic for Wikipedia articles. This is the link: https://www.kaggle.com/c/web-traffic-time-series-forecasting.

1.1 Literature Review

There have been several other studies involving the use of time series modelling to web analytics. The paper written by Beatriz Plaza (1) studied how long the user stays on a page and about the origins of the web traffic whether it was from Google searches for their website. They tested that their time series data was stationary and used the ARMA model for prediction.

The Recurrent Neural Networks (RNN) seq2seq model is also used to forecast web traffic (2). RNNs are time-series neural network model, which predicts the conditional probability of the next time step given the full history of the time series. It was trained on historical pageviews using a sliding window method.

In this paper, we aim to forecast web traffic volume data instead of the origins of the visitors. The ARMA model is used instead of the RNN model for ease of interpretation.

2 The Models

2.1 Autoregressive Moving Average Process (ARMA)

ARMA was popularized by Box and Jenkins in 1976 (3) but was first described by Peter Whittle in 1951 (4). It combines two types of models: 1) The AR model and 2) and MA model. Its AR component involves regressing its values on its own past values while the MA component regresses over the error terms. It is helpful for understanding and predicting stationary time series data which can be used for forecasting.

2.1.1 Autoregressive Process (AR)

Autoregressive models are used to describe certain time-varying random processes. It shows that the future values are linearly dependent on its past values. The general AR(p) follow this form:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \epsilon_t \tag{1}$$

Where $\varphi_0 \ldots \varphi_p$ denotes the coefficients, p denotes the order which means the number of past observations needed to forecast the present and $\epsilon_t \sim WN(0, \sigma^2)$ denotes the error term. It can also model non-stationary processes, such as ones with a linear trend.

2.1.2 Moving Average Process (MA)

The Moving Average model depicts how current white noise or random shock is linearly dependent on past shock. It is always stationary. A general MA(q) follows this form:

$$y_t = \epsilon_0 + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \tag{2}$$

Where $\theta_0 \ldots \theta_p$ denotes the coefficients and $\epsilon_t + \epsilon_0 \ldots \epsilon_q$ denotes the errors.

2.1.3 ARMA

Combining the AR and MA models, the general form of the ARMA(p,q) model is:

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \tag{3}$$

Where $\varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p}$ is the AR(p) process and $\theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$ is the MA(q) process. The AR component models the overall trend of the data, while the MA component models the serial correlation of the error terms.

3 Empirical Data and Results

3.1 The Data

The time series used is the number of visits to the Facebook Wikipedia article from July 1, 2015 to December 31, 2016. The time series for the web traffic is displayed in Figure 1.

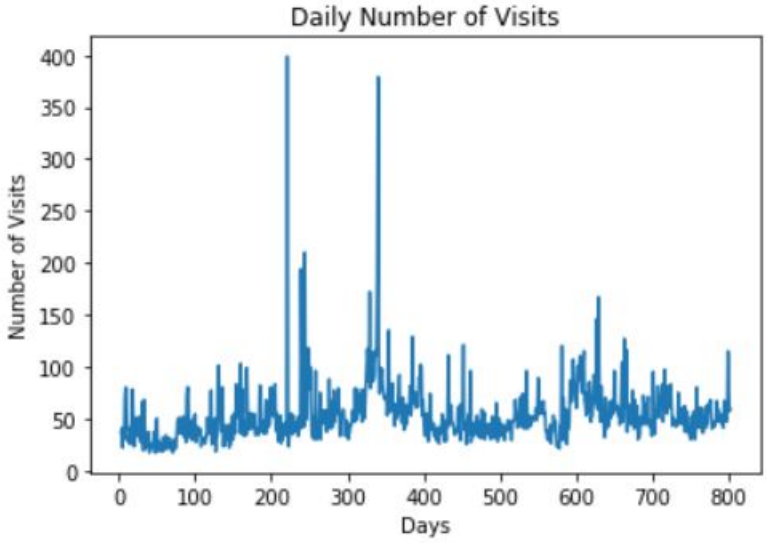

Figure 1: Number of Daily Visits in units to the Facebook Wikipedia article

The mean is 53.64, the variance is 822.44, skewness is 5 and the kurtosis is 48.55. There were two notable spikes in the data that may affect the regression such as the ones on February 2, 2016 and June 5, 2016. The massive spike on February 2, 2016 can be explained by Facebook's initial tests on the emoji reactions which was officially released on February 24 (7). The spike on June 5, 2016 was when Facebook was planning on making major changes to how ads are handled as well as a new Newsfeed look (8).

There are no consistent seasonal or weekly trends observed either in the original time series in Figure 1, and in the ACF plot in Figure 2. However, there is notable autocorrelation in the time series.

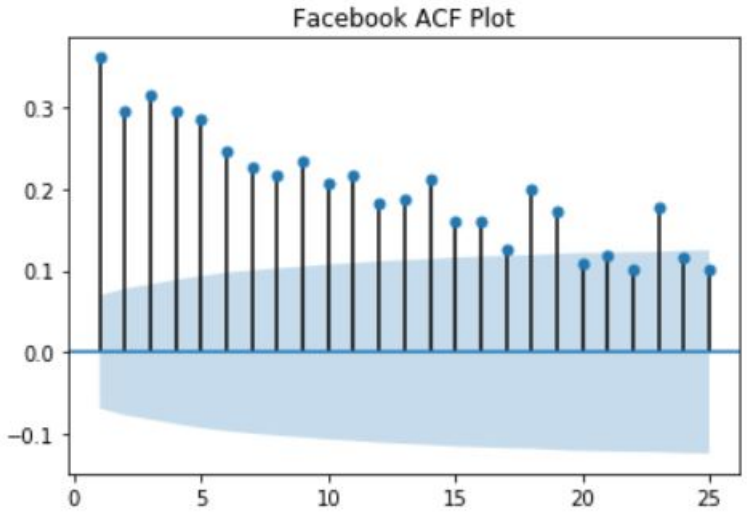

Figure 2: ACF Plot of Number of Daily Visits

When performing the Dickey-Fuller test (5) to see if the series is stationary, the result is that it is stationary with a value of -6.965 which is smaller than its 1% critical value which is -3.439 and a p-value of $8.810406134104259e^{-22}$.

3.2 Results

Since the data is stationary, we can use the AR, MA and ARMA model. When fitting different combinations of p and q into the AR, MA and ARMA models we use Akaike's Information Criterion(AIC) (6) to select the model that best fit the data. It measures the amount of information captured by the model, while taking into account the number of parameters that were used to capture that information. We concluded that the ARMA(1, 1) model was better as it had the lowest AIC out of all the models ranging from orders $p =$ (1, 2, 3, 4, 5) and $q =$ (1, 2, 3, 4, 5) (as shown in Table 1). Hence the model selected is

$$y_t = 0.9478y_{t-1} - 0.7784\epsilon_{t-1} + \epsilon_t \tag{4}$$

Removing the outliers aforementioned did not significantly affect the model parameters.

Figure 3 shows the ACF plot of the residuals of the model. From this, we observe no serial correlation and trend, therefore, we can conclude that the residuals are independent of each other and that the ARMA model fits the data adequately.

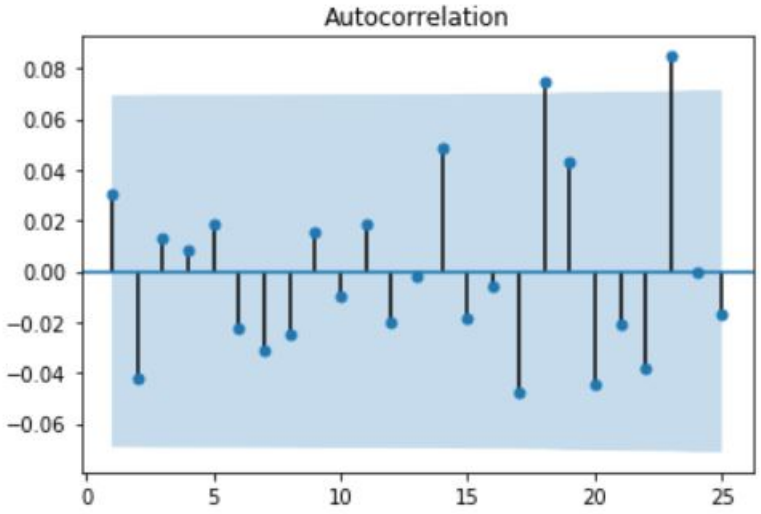

Figure 3: ACF Plot of Residuals

Since both the AR and MA components of the selected ARMA(1, 1) model only utilize one lag, this implies that today's number of visits conditionally depend only on yesterday's number of visits. The series is adjusted to the mean hence the AR part of the model states that after a spikes, the number of daily visits will slowly decrease overtime to the mean. With a negative MA coefficient, this can mean that the previous day's visits would have an opposite effect on the next day as seen in the jaggedness of the data.

In Figure 4, we use our ARMA(1, 1) model to forecast the web traffic for the next 10 days. The forecast will converge to the mean as we increase the horizon.

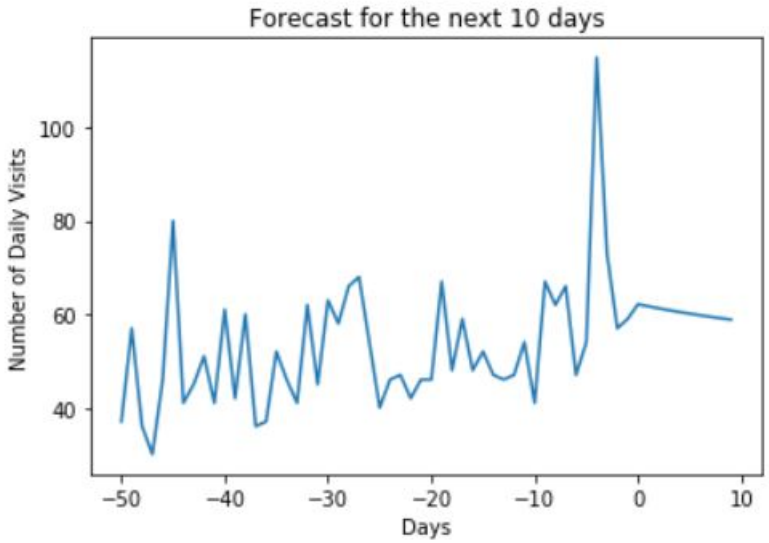

Figure 4: Number of daily visits forecasted up to 10 days in the future. Day 0 is the start of the forecast and the negative days are the historical data.

4 Conclusion and Future Research

In this paper, we model and forecast the amount of web traffic by using time series analysis. We tested AR, MA and ARMA models of varying orders p, q on the data with ARMA(1,1) being the best fit according to AIC. The ACF plot of its residuals suggests that it is a reasonable fit. Additionally, since p and q are both 1, the value for tomorrow.

Due to the randomness of the users who visit the Facebook Wikipedia article, there can be potential for better methods to fit the data. A better way to forecast the web traffic could be to find and utilize other explanatory variables that could explain it such as whether the particular company or business has been mentioned in media. Such an explanatory variable would be able to help model the two spikes/outliers that we observed in our dataset. Another idea would be to see if there are any potential seasonality and weekly trends such as people going on the web more during the weekends and such for other websites.

References

[1] Beatriz Plaza, (2009) "Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series", Aslib Proceedings, Vol. 61 Issue: 5, pp.474-482, https://doi-org.proxy.lib.uwaterloo.ca/10.1108/00012530910989625

[2] Suilin, A. (2017). Web Traffic Time Series Forecasting. Retrieved from https://www.kaggle.com/c/web-traffic-time-series-forecasting/discussion/43795

[3] G. E. Box and G. M. Jenkins, Time series analysis: forecasting and control, revised ed. Holden-Day, 1976.

[4] Whittle, P. (1951). Hypothesis Testing in Time Series Analysis. Almquist and Wicksell. Whittle, P. (1963). Prediction and Regulation. English Universities Press. ISBN 0-8166-1147-5. Republished as: Whittle, P. (1983). Prediction and Regulation by Linear Least-Square Methods. University of Minnesota Press. ISBN 0-8166-1148-3.

[5] Dickey, D. A.; Fuller, W. A. (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". Journal of the American Statistical Association. 74 (366): 427431. doi:10.1080/01621459.1979.10482531. JSTOR 2286348.

[6] Akaike, H. (1974), "A new look at the statistical model identification", IEEE Transactions on Automatic Control, 19 (6): 716723, doi:10.1109/TAC.1974.1100705, MR 0423716.

[7] *Timeline of Facebook*, available at `https://en.wikipedia.org/wiki/Timeline_of_Facebook`.

[8] Karlson, K. (2016).TOP 5 Updates by Facebook That You Need to Know Now (July 2016 Edition). AdEspresso. Retrieved from https://adespresso.com/blog/top-5-updates-facebook-need-know-now-july-2016-edition/

5 Appendix

5.1 Appendix A: Table of AR, MA and ARMA model results

Model	AIC	BIC
ARMA(0,0)	7644.129	7653.498
MA(1)	7565.164	7579.218
MA(2)	7543.781	7562.781
MA(3)	7522.786	7546.209
MA(4)	7507.885	7535.993
MA(5)	7492.126	7524.919
AR(1)	7533.897	7547.951
AR(2)	7506.299	7525.037
AR(3)	7478.71	7502.133
AR(4)	7466.529	7494.637
AR(5)	7458.583	7491.375
ARMA(1,1)	7447.3	7466.039
ARMA(1,2)	7447.836	7471.259
ARMA(1,3)	7449.007	7477.114
ARMA(1,5)	7452.512	7489.989
ARMA(2,1)	7447.964	7471.387
ARMA(2,2)	7448.713	7476.821
ARMA(2,5)	7454.635	7496.796
ARMA(3,1)	7449.049	7477.157
ARMA(3,5)	7456.112	7502.958
ARMA(4,5)	7452.134	7503.665
ARMA(5,5)	7453.5	7509.716

Table 1: AIC and BIC values for various AR, MA and ARMA models. Models that failed to converge were excluded from the analysis.

5.2 Appendix B: ARMA(1,1) Summary

ARMA Model Results

Dep. Variable:	Facebook	No. Observations:	800
Model:	ARMA(1, 1)	Log Likelihood	-3719.650
Method:	mle	S.D. of innovations	25.287
Date:	Sun, 09 Dec 2018	AIC	7447.300
Time:	03:49:11	BIC	7466.039
Sample:	0	HQIC	7454.499

	coef	std err	z	P>|z|	[0.025	0.975]
const	53.6080	3.730	14.371	0.000	46.297	60.919
ar.L1.Facebook	0.9478	0.018	52.555	0.000	0.912	0.983
ma.L1.Facebook	-0.7784	0.037	-21.081	0.000	-0.851	-0.706

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0550	+0.0000j	1.0550	0.0000
MA.1	1.2847	+0.0000j	1.2847	0.0000

Figure 5: ARMA(1,1) Model Summary