Data Assignment – Amazon.com's Wikipedia Website Visits

Adriana Choy Sampei

MA611

Abstract

The aim for this analysis is to understand the patterns of the number of Wikipedia website visits for Amazon.com using time series models. The dataset used for this project has a 2-year time frame that goes from July of 2015 to July of 2017. Additive decomposition identified a weekly seasonality in the data, which helped determine the time series model and variables to be explored for this research: Holt-Winters, SARIMA, and Linear Regression. An NNAR model was also implemented for comparison purposes. The criteria to determine which model is best for the time series included checking accuracy (RMSE, MAE) and plotting the residuals and their correlograms to verify if they are White Noise. The NNAR model had the lowest RMSE as well as White Noise residuals, but for the sake of interpretability the Holt-Winters model is also considered a good candidate for understanding the patterns of the time series data.

Introduction

In the past decade, Amazon.com has become one of the biggest e-commerce marketplaces, even becoming one of the US' most trusted brands in 2022.[1] How does that apply when it comes to Internet users researching about the company's profile and history? The number of Wikipedia visits is a direct reflection of the public's interest on the company's brand and their strategic movements along the years. This research aims to understand time series of the number of Wikipedia page visitors over a 2-year period in order to determine what model is best at interpreting the data and could be used for future Wikipedia visitor forecasting. This would help understand how Internet users react and reflect their interest towards the brand via the use of their Wikipedia page.

The time series is plotted in Figure 1. There is no significant noticeable trend, but the level of the data seems to be going upwards towards the end of the time frame, starting from July 2016. This is more evident in Figure 2, which is an additive decomposition of the time frame, where the trend level has a slight upwards trend starting Week 60. The pattern also goes through very subtle cycles as slight up and down fluctuations are observed throughout the series. The very high peak towards the end of

[1] Quaker, D. (2023) *Amazon selling stats*. Amazon. Retrieved December 7th, 2023 on https://sell.amazon.com/blog/amazon-stats

the series, around June 2017, is around the time Amazon announced they would be buying Whole Foods, which sparked a lot of interest from the public at the time.[2]

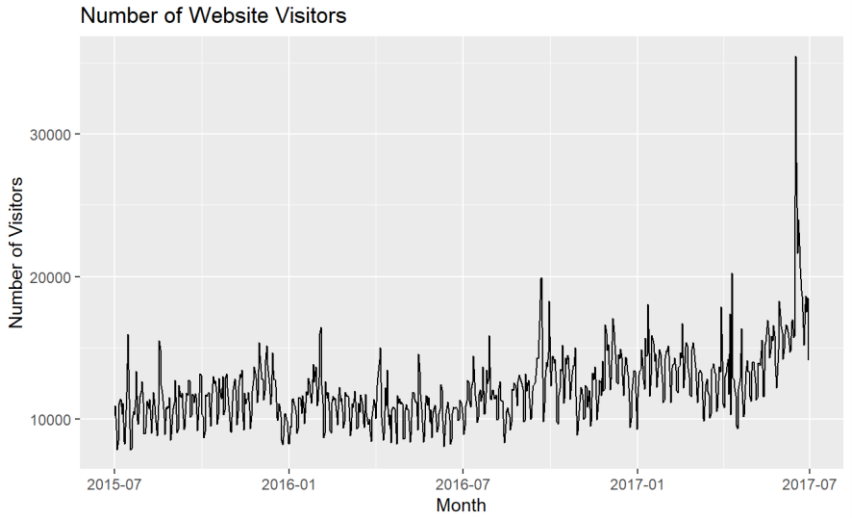

Figure 1. Plot of the Number of Wikipedia Visitors for Amazon.com from July, 2015 to July, 2017.

There also seems to be seasonality in the time series for which an additive decomposition and a seasonal plot were made in Figure 2. Wikipedia visits have a weekly seasonality where number of visitors is more likely to be higher during weekends and hit a low in the middle of the week (Wednesdays and Thursdays). These weekly fluctuations might be associated with users having more leisure time during those days to look up and research about the company.

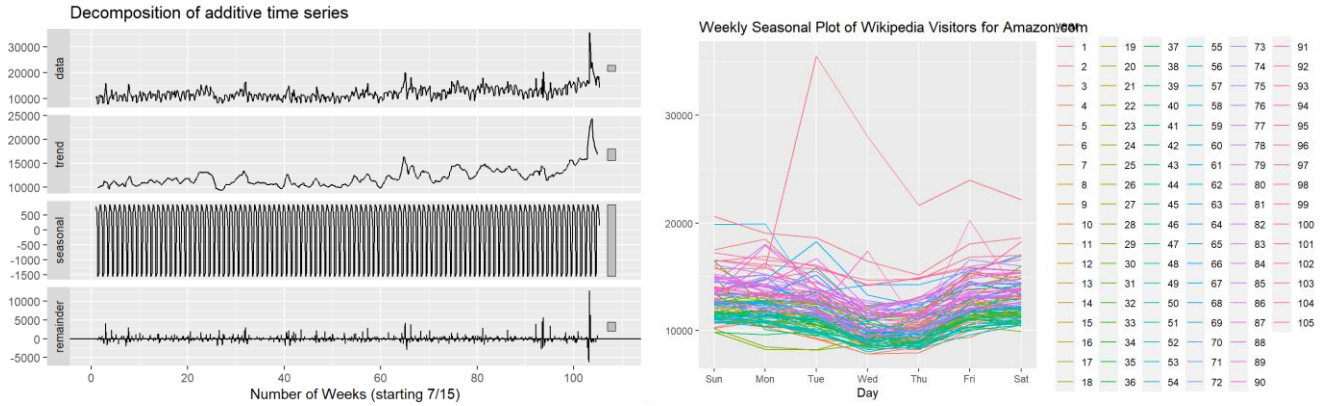

Figure 2. Additive decomposition (left) and Weekly seasonal plot of the Number of Visitors for Amazon.com's Wikipedia page.

[2] La Monica, P. & Isidore, C. (2017) Amazon is buying Whole Foods for $13.7 billion. CNN Business. CNN. Retrieved December 7th, 2023 on https://money.cnn.com/2017/06/16/investing/amazon-buying-whole-foods/index.html

Methodology

The presence of seasonality and a slight upwards trend in the time series helped determine which models would be most appropriate to explore for this research: Holt-Winters, SARIMA, and Linear Regression using trend and season predictors. NNAR was also looked into due to its flexibility with time series models.

The criteria used to compare these different models was to compare accuracy measures such as RMSE and MAE, as they represent the amount of error that the models have fitting the data. Therefore, models with smaller RMSE or MAE would be best for this research.

Additionally, the model's residuals and its correlograms were plotted in order to verify that they looked like Gaussian White Noise and had no autocorrelation issues. The residuals looking like White Noise signify that the model has appropriately captured any trend or seasonality patterns that were in the time series. If they did not look like White Noise or had lags with high autocorrelation, the residuals were put into an ARMA model and plotted again to check, once again, if they looked like White Noise.

The assumption for our research question is that the models that fit best with the data: 1) minimize error, and 2) properly capture trend and seasonality patterns. This would signify that we have found time series models that best describe the fluctuations in number of visitors throughout the time frame of the data. It is also important to consider the interpretability of these models for future use.

Results

After looking into all of the models mentioned, the accuracy measures and whether the residuals look like White Noise or not has been summarized in the following table:

	Holt-Winters	**SARIMA**	**Linear Regression**	**NNAR**
RMSE	1381.471	1485.191	1754.53	373.9925
MAE	806.7204	915.5807	1186.82	233.7013
Do the residuals look like White Noise?	Yes	Not ideal, autocorrelation	Yes	Not ideal, autocorrelation

Although the NNAR has very low RMSE and MAE in relation to the other models presented in this research, the residuals have a couple of lags with autocorrelation outside of the boundaries. In addition to that, there is a complexity in this model in terms of interpretability that make the Holt-Winters

model more suitable despite its lower accuracy to fit the data. A figure of a 2-week forecast using Holt-Winters is presented in Figure 3. The Holt-Winters' smoothing parameters are the following:

- Level smoothing parameter – $\alpha = 0.5536$
- Slope smoothing parameter – $\beta = 0.0008$
- Seasonality smoothing parameter – $\gamma = 0.0001$

This means that for the level/intercept, there is a bigger emphasis on more recent data. However, when it comes to the slope and the seasonality, there is a significantly higher weight in past observations. The parameters set for this model leave a better understanding of how the time series patterns are best adjusted throughout time which might help understand the forecasting process of future observations.

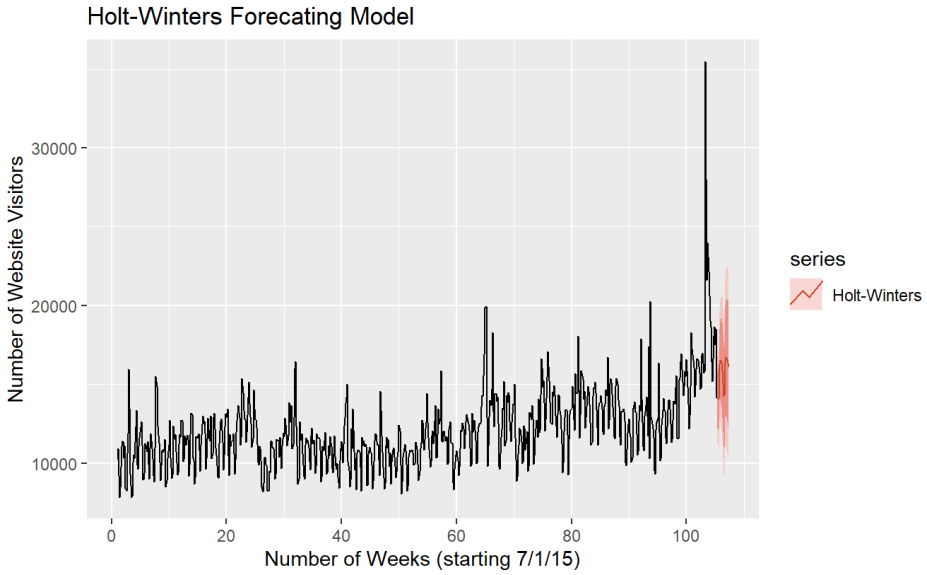

Figure 3. 2-Week Holt-Winters Forecasting for Number of Wikipedia Visitors for Amazon.com

Conclusion

The number of Amazon.com's Wikipedia visits from July 2015 to July 2017 has a weekly seasonality where it peaks during the weekdays and has a slight upward trend towards the second half of the time frame of the dataset. Because of these circumstances, a Holt-Winter model, a SARIMA model, a Linear regression using trend and season variables, and NNAR model were most appropriate to explore which one fit the time series best. Accuracy measures, residual plots, and correlograms were compared in order to determine which model would be best at minimizing error while capturing the trend and seasonality of the data.

After this process, the Holt-Winters model was recommended over the NNAR because of autocorrelation and interpretability issues. In the Holt-Winters model, level smoothing parameters put more weight on more recent data while slope and seasonality put a significantly higher weight on past observations. Finally, while Holt-Winters might seem to be best adjusted on this analysis, there might be other forecasting/time series models that have not been evaluated in the scope of this report.

Sources

- Quaker, D. (2023) *Amazon selling stats*. Amazon. Retrieved December 7th, 2023 on https://sell.amazon.com/blog/amazon-stats
- La Monica, P. & Isidore, C. (2017) Amazon is buying Whole Foods for $13.7 billion. CNN Business. CNN. Retrieved December 7th, 2023 on https://money.cnn.com/2017/06/16/investing/amazon-buying-whole-foods/index.html