Time Series Analysis on “Netflix” Search Term Data

Anna Xu
Aleksa Basara

STSCI 4550 Applied Time Series
Cornell University
May 7, 2017

Time Series Analysis on “Netflix” Search Term Data

1. Abstract

In this paper, we used time series analysis to analyze data on how frequently “Netflix” is searched for in Google over time. Our findings would allow companies to make better business decisions with the knowledge of when people are likely looking to stay in and watch movies. After preliminary analysis, we found that the data is highly seasonal; hence, we fit a seasonal ARIMA model and determined the best performing model was ARIMA $(4,1,3)(1,1,0)_{52}$.

2. Data description

Our data was publically available at trends.google.com, Google’s method of tracking search word usage. Each data point represents how often “Netflix” was searched in Google for a given week. This number refers to how often “Netflix” was searched as a service or company, not how many times it has appeared in a Google search. Due to the diverse preferences and marketing strategies across the globe, we decided to limit the data to searches done within the United States. The data is weekly data that goes back 5 years, giving us 260 data points. We chose to look at the past 5 years to allow for analysis of a few cycles, but limit the effects of changing technology over time.

3. Data Preparation

As the dataset had no missing observations, we started by plotting the data over time to check for any immediate data transformations that needed to be done. Shown in **Figure 1**, there is a strong seasonal component with a large peak appearing around every December. We can see that the data is not-stationary because the variance decreases over time. This is most evident when we focus on the peaks in December that become less prominent over time.

Figure 1: This graph shows the untransformed Time Series of Netflix search volume over time

4. Preliminary Analysis

Because the data is non-stationary and is seasonal, we took a seasonal difference using lag 52 and again made a plot of the Netflix time series, shown below in **Figure 2**. In this new plot, we again observe a level of seasonality and nonstationarity, albeit a lesser one than before. We also show the ACF and PACF of the newly differenced data in **Figure 3.** The sinusoidal shape of the ACF plot suggests that further differencing must be done to achieve stationarity.

Figure 2: Time Series after seasonal difference was applied

Figure 3: ACF and PACF of Seasonally Differenced Time Series

After further differencing the data, we once again created a plot of the time series to check for seasonality and stationarity. As evidenced by **Figure 4,** the time series appears to be fairly stationary, with no significant changes in mean or variance over time. We also no longer observe the same seasonality that we saw in the original data. As shown in **Figure 5,** both the ACF and PACF show no significant pattern in correlations for different lags. Once we observed

these traits in our analysis process, we determined that twice differencing the time series was necessary when fitting our seasonal ARIMA model.

Figure 4: Time Series after second differencing

Figure 5: ACF and PACF of Time Series after second differencing

5. Model Identification and Diagnostic Checking

In this stage of the process, we wanted to fit an appropriate ARIMA model. Based on the PACF, the significant spike at lag 4 suggested an AR(4) component. From visual inspection, we know the data is highly seasonal. Therefore, we fit a ARIMA(4,1,0)$(0,1,0)_{52}$ model, indicating a first and seasonal difference, and a non-seasonal AR(4) component. We fit this model and a many variations on it and we computed the AIC values. **Table 1** shows some of the models we fit along with their AICs.

Table 1: AIC values of 5 best models

Model	AIC
ARIMA(4,1,0)$(1,1,0)_{52}$	1111.420
ARIMA(4,1,4)$(1,1,0)_{52}$	1111.406
ARIMA(4,1,3)$(1,1,1)_{52}$	1111.202
ARIMA(4,1,3)$(0,1,1)_{52}$	1110.467
ARIMA(4,1,3)$(1,1,0)_{52}$	1109.594

Of these models, the two best are ARIMA(4,1,3)$(1,1,0)_{52}$ and ARIMA(4,1,3)$(0,1,1)_{52}$ because they have the lowest AIC values. We then proceeded to perform some of the diagnostic checks on these two models. As can be seen in the ACF plots in **Figure 6** and **Figure 7**, both models have residuals that are uncorrelated for different lags. This suggests the residuals have nothing more to tell us about the data and our two best models do a sufficient job of describing the data. The non-significant values of the Ljung-Box test confirm this.

Figure 6: Diagnostic check of ARIMA(4,1,3)(1,1,0) model
Figure 7: Diagnostic check of ARIMA(4,1,3)(0,1,1,) model

From the graphs of the standardized residuals, we observe a slight sinusoidal shape. However, we decide due to the uncorrelated nature of the residuals and lack of significant spikes at any lag, that we may proceed with these two models. As the difference in AIC was very small, we decided to further compare the ARIMA(4,1,3)$(1,1,0)_{52}$ and ARIMA(4,1,3)$(0,1,1)_{52}$ models to make a more educated selection.

In this stage of the analysis, we decided to implement a method of cross-validation that splits the time series up into a training set and a test set. The training set would contain the first 210 observations, about 80%, and the test set would include the rest of the data. Then, for both models, we predicted 5 weeks ahead and compared the forecasted values to the observed values in the test set and computed various measures of the error. As evidenced by **Table 2** below, all measures of the error were lower for the ARIMA(4,1,3)$(1,1,0)_{52}$ model. This, in addition to the slightly smaller AIC value, led us to conclude this was the better of the two models and the one we would use to model our time series of Netflix search volume.

Table 2: Errors from Cross-Validation Method

	ME	RMSE	MAE	MPE	MAPE
ARIMA(4,1,3)$(1,1,0)_{52}$	2.359	6.062	4.701	3.615	7.950
ARIMA(4,1,3)$(0,1,1)_{52}$	2.613	6.199	4.807	4.059	8.119

6. Forecasting

Using the ARIMA(4,1,3)$(1,1,0)_{52}$ and ARIMA(4,1,3)$(0,1,1)_{52}$ models, we forecasted 4 weeks ahead to see how their predictions and their corresponding standard errors would compare.

Table 3: 4-Step ahead predictions for ARIMA(4,1,3)$(1,1,0)_{52}$ model

	1st Week	2nd Week	3rd Week	4th Week
Predicted Value	50.542	53.156	53.838	56.026
S.E.	3.331	3.665	3.849	3.969

Table 4: 4-Step ahead predictions for ARIMA(4,1,3)$(0,1,1)_{52}$ model

	1st Week	2nd Week	3rd Week	4th Week

Predicted Value	50.423	52.901	53.376	55.892
S.E.	3.347	3.696	3.892	4.021

As can be seen above in **Table 3** and **Table 4,** there are a few noticeable differences between the predictions of the two models. The predictions of the ARIMA(4,1,3)$(1,1,0)_{52}$ model tended to be slightly higher and were associated with lower standard errors for all four predictions. Due to the smaller standard errors, we would expect that this first model yields narrower confidence intervals and hence more precise predictions. If this were true, it would correspond to the lower prediction error values we saw in the first model. We were curious to see whether these two features would hold out when predicting one year ahead into the future.

Figure 8 and **Figure 9** show the one year ahead predictions of the two models. Both sets of predictions include the characteristic spike around December that was observed in the collected data from the past 5 years, suggesting that both models capture the seasonality of the time series well. As mentioned before, we noticed slight differences in the one year ahead predictions. The ARIMA(4,1,3)$(1,1,0)_{52}$ model yielded higher search volume predictions than the ARIMA(4,1,3)$(0,1,1)_{52}$ model and had narrower confidence intervals for its predictions.

Figure 8: One year ahead predictions for ARIMA(4,1,3)$(1,1,0)_{52}$ model

Figure 9: One year ahead predictions of ARIMA(4,1,3)$(0,1,1)_{52}$ model

7. Results

At the end of our analysis of the Netflix time series and our model selection process, we have arrived at the model shown in **Equation 1**:

Equation 1: Best model

Conclusion

In this project, we considered the weekly volume of the search term “Netflix” in Google for the past 5 years. We initially ran into some issues with stationarity, so we removed the seasonality and trend from the data. We then fit an seasonal ARIMA model and selected ARIMA $(3,1,4)(1,1,0)_{52}$ based on how well it performed when we ran diagnostics, how well it forecasted, and the AIC value.