GEORGIA STATE UNIVERSITY
ROBINSON COLLEGE OF BUISNESS

Time Series Forecasting Of Google Trends

Somesh Yadav (002632424)

Yash Wali (002654499)

Satvik Vandanam

March 27, 2022

Abstract

We analyse the Trends in Google searches over a particular period of Time for a particular search and try to forecast further results using time various time series analysis methods. The topic we selected was Home Depot because that search would have some seasonal trends. The time series models used were auto Arima, Seasonal ARIMA and Facebook Prophet. We further use additional variables such as Ad. spends to draw a correlation and further use that variable in our forecasting. The performance of multiple models were compared using MSPE and MAPE. We found out that SARIMA without additional variables provided the best accuracy.

1 Methodology

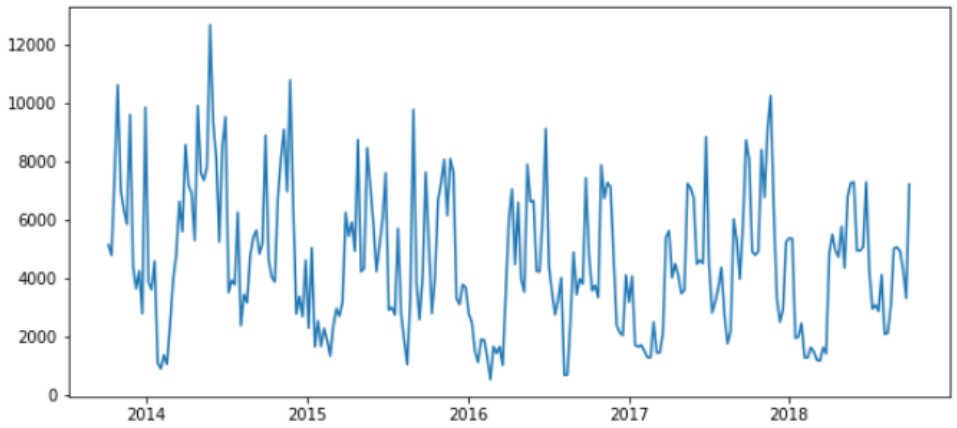
We begin with extracting the data and analyzing their Time Series Graphs to search for any visible trends. Our next step is to test for stationarity, using the ADF test. After confirming the stationarity of the data, we move on the model selection process for ARIMA. We do this using two methods; firstly, using the auto ARIMA function. Secondly, we visually inspect the ACF and PACF plots of both our target variable, and the difference of our target variable. Furthermore, we work on FB Prophet as another model to forecast our target variable. All of the above steps are done with multiple granualities of data, and we use the granuality which gives us the best results.

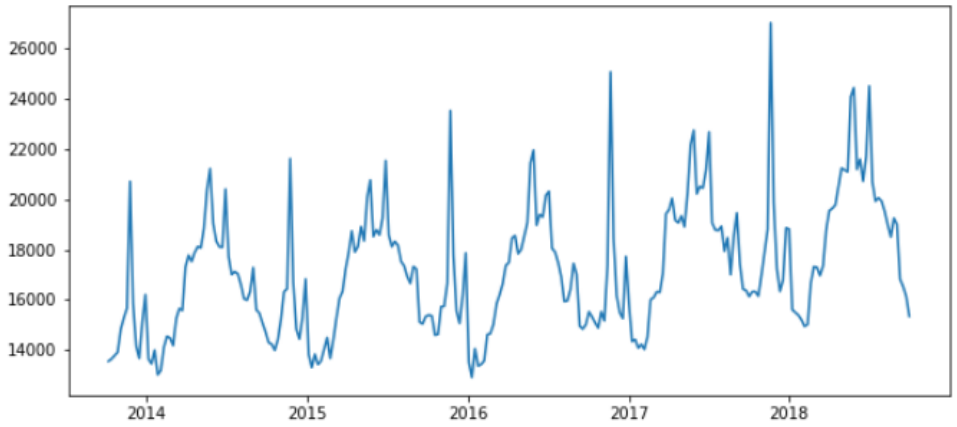
2 Data Exploration and Visualization

We have two data sets:
1) Home Depot Ad spend: This dataset contains the monetary amount spent on Ads by Home Depot through various sources such as Newspapers, Magazines, Network TV, Network Radio, etc.
2) Home Depot Google Trends: The number of google searches of Home Depot.

2.1 Time Series Graph

We will now look at two graphs:
1) The Graph of our Time series of the total value spent on ads by Home depot on multiple sources and then we total them to get the total ad spends.
2) The second graph is the time series graph of the number of searches. This is our target variable.



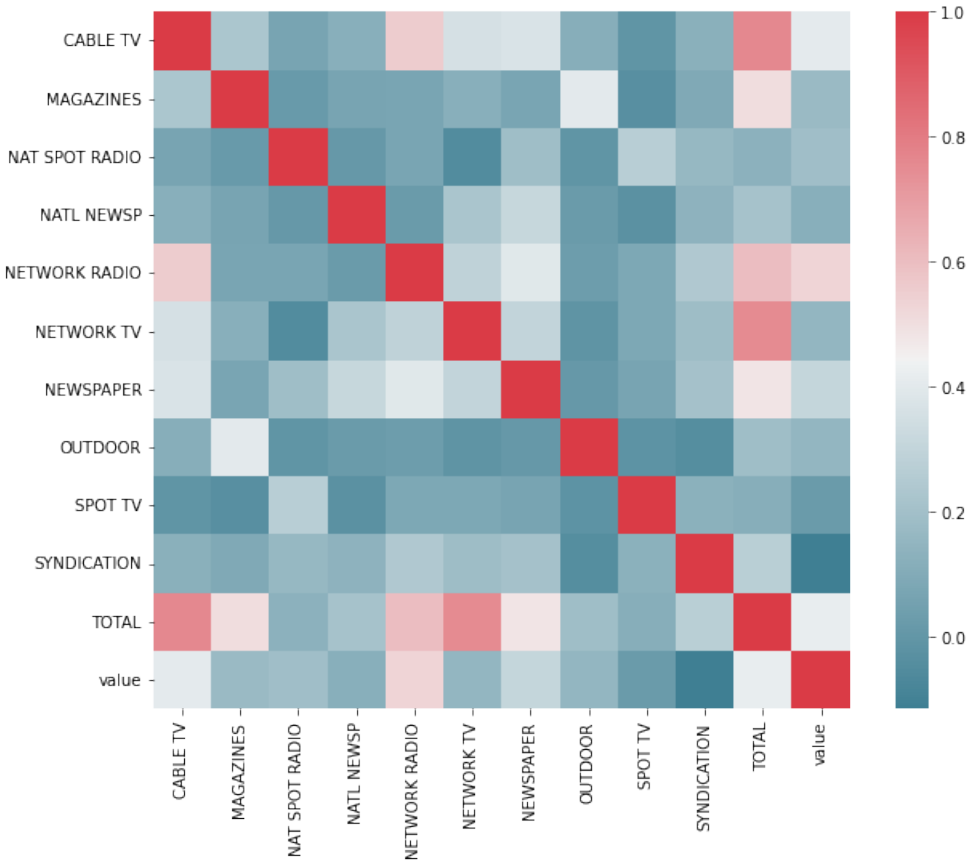


We have two interesting observations:

- We can see the peaks and troughs roughly lining up, showing a positive correlation. This makes sense, as Homedepot spends more money on adverts, they get more google searches.
- We can see average amount of ad spending reduce but the average searches increase. This is an interesting observation. Although we cannot say anything definite from just this, we can possibly conjecture that the Ad spending has been optimized for better results over time.

2.2 Correlation Graph

The second interesting graphic is of the correlation table of the Home Depot Ad Spend Data-Frame:



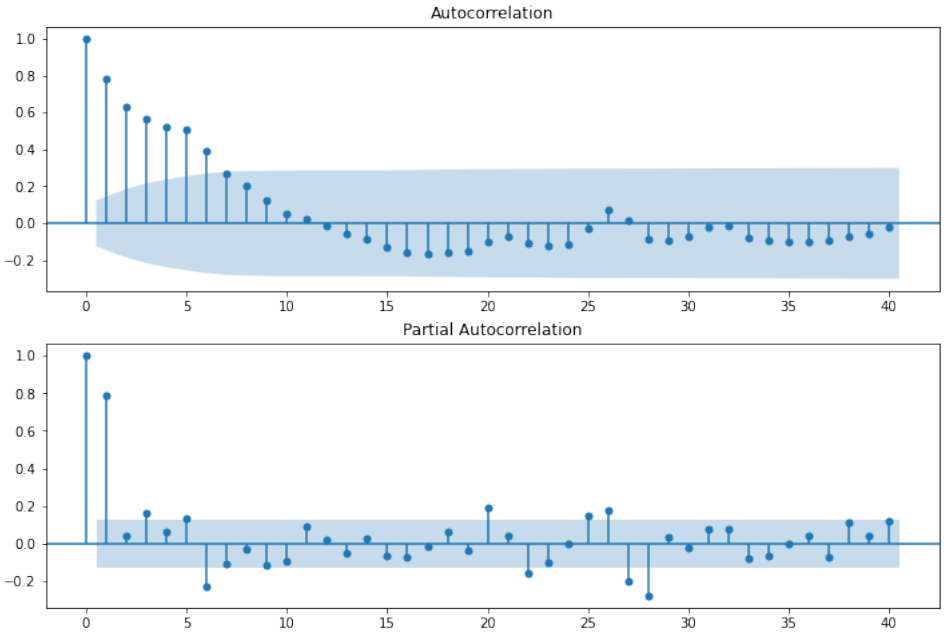
Here we try to find some significant correlation between our target variable i.e value and the ad spent through various channels. Network Radio spending has the highest correlation to our target variable which in fact is not very high either. In fact Total column which is the total spending of all channels combined is weakly correlated to our target variable. We can thus say that radio network ads drive the most traffic for searches.

3 Modelling And Forecasting

3.1 Seasonal Arima

Since we see seasonality in our time series graph we will be using seasonal ARIMA as the first method for forecasting. We have selected weekly searches as our forecast period. For this we took the weekly average of the value. We will use additional variables to see its effect on our forecast and accuracy. Further an ADF test was performed on the data to confirm if the data is stationary. In the case data is not stationary , there are methods like difference and log to make the data stationary.

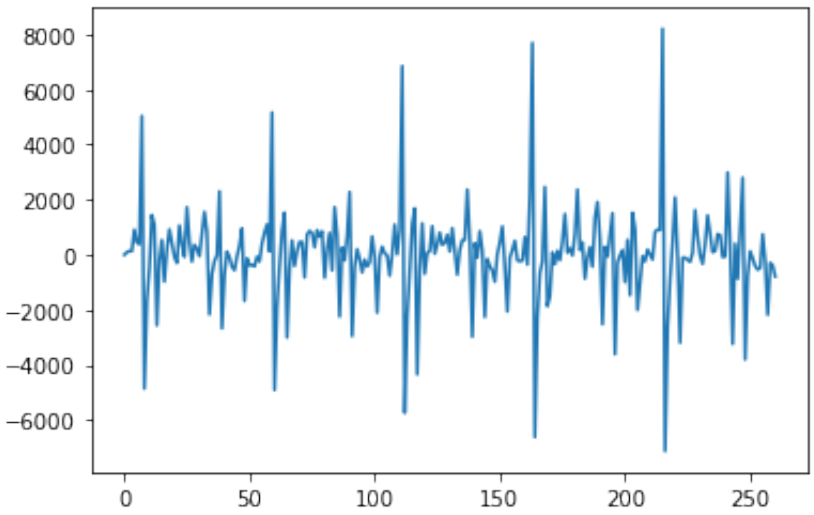
- For first step, we plot the ACF and PACF to decide the parameters of our model.



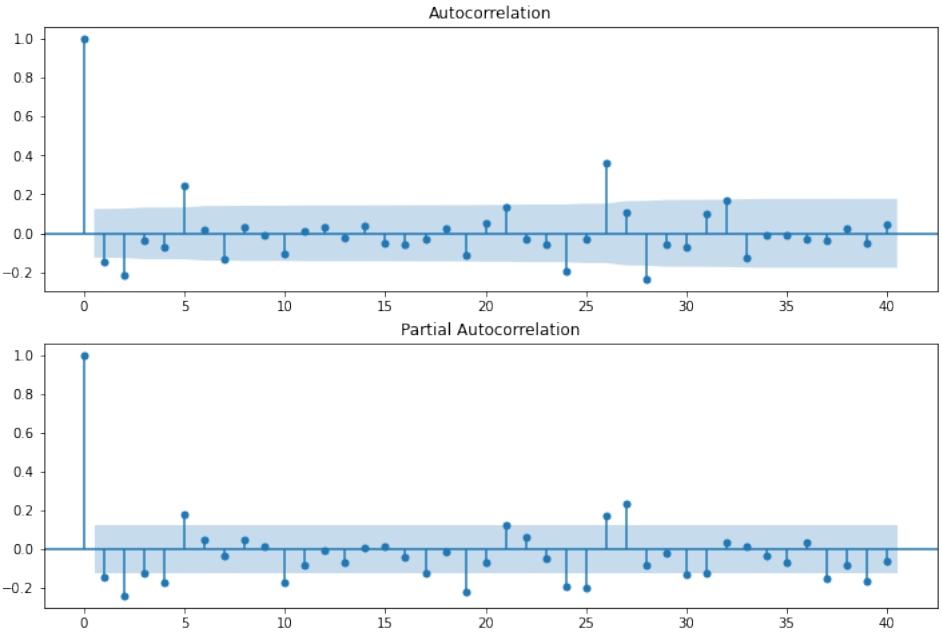
  - It was noticed that the ACF graph has significant values till the lag of 5 or 6. The ACF graph helps us in deciding the parameter for the MA part of our ARIMA model. Thus MA 5 or MA 6 will be an appropriate parameter for our ARIMA model.
  - Coming to the PACF part notice the value decrease suddenly after the first lag.The PACF graph helps us decide the parameter for the AR part of our ARIMA model thus we conclude this to be an AR 1 model.

- The next step was to detect parameters for the seasonal part of our seasonal ARIMA model. Our approach to find seasonality will be to take the difference

of our readings, i.e the readings comprise of previous reading subtracted from next as our difference value. Plotting this.

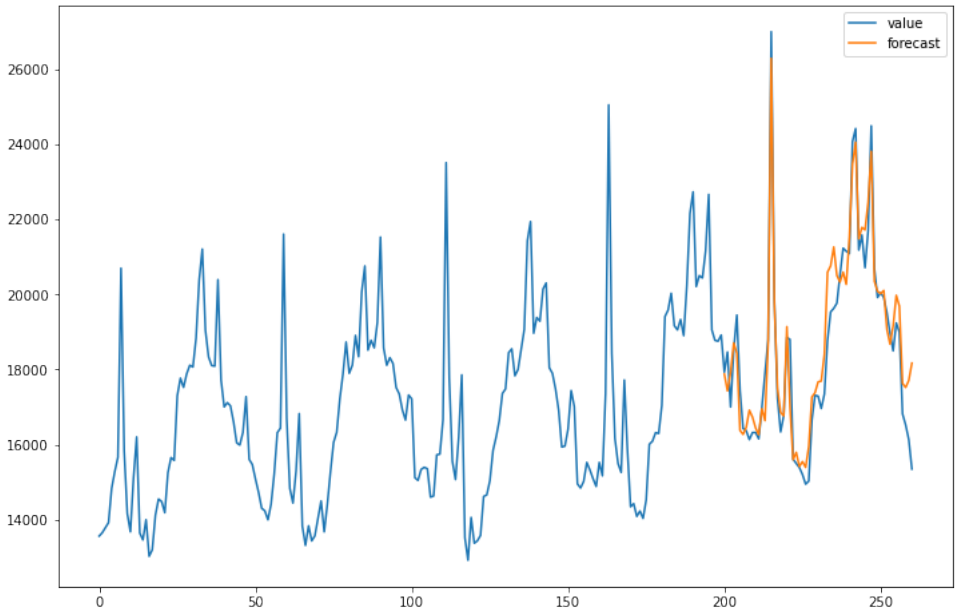


– From the above graph it is quite visible that our data follows a seasonality of 52 weeks,which is approximately 1 year. The next part will be to look at ACF and PACF and decided the seasonal parameters.



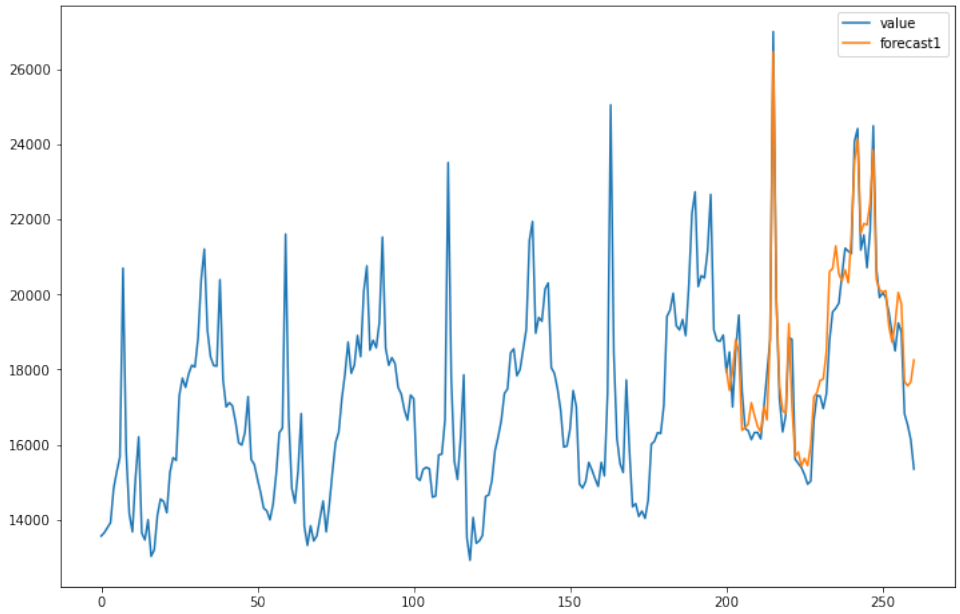
– From the above graph it was decided that either (0,1,1) or(0,1,0) will be appropriate seasonal parameters for the ARIMA model.

- Now most of the parameters have been estimated graphically and next step was to try auto arima to generate the best parameters using AIC values. The results obtained gave us a model of (1,1,1)X(0,1,0)52 as the parameters with least AIC value.The next step is to use the parameters to fit the value and prediction for which we split the data into train and test. 80% of the data is considered as training the rest as testing data.
- Training the model using both the parameters.
- Model Evaluation
  - Model with (1,1,5)X(1,1,0)X 52 Gave the best result.



  - Error:
    - MSPE is 0.21%
    - MAPE is 3.34%

* Model with (1,1,5)X(1,1,0)X 52 with additional variable.

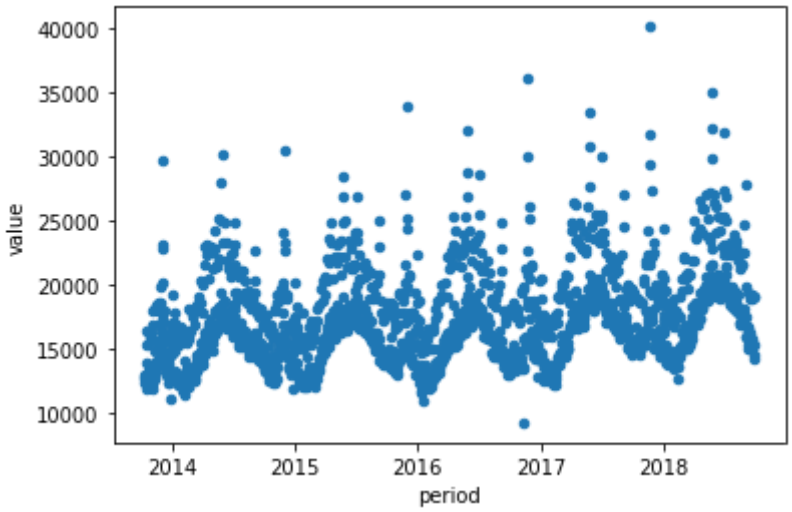


* Error:
  - MSPE is 0.22%
  - MAPE is 3.46%

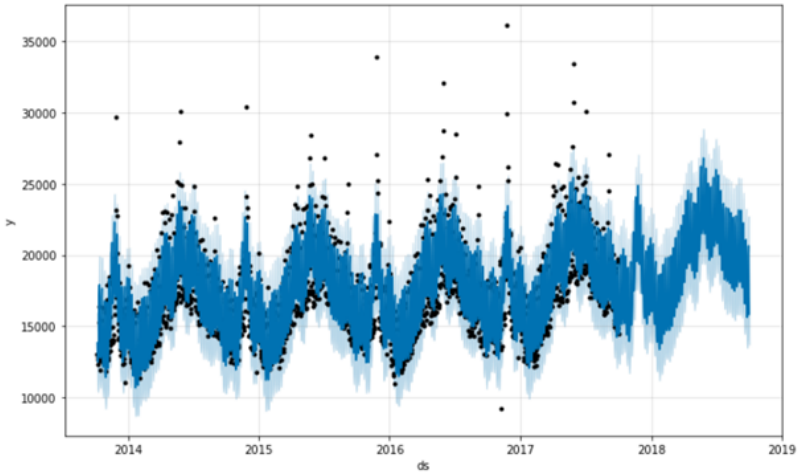
3.2 FB Prophet

FB prophet uses the date-time stamp, and the corresponding target variable for predictions. We prepare the data set accordingly. The date-time stamp column is renamed to 'ds', and the target variable column is renamed to 'y', as is required to run this model. Note that the resolution of the data is daily, unlike that of weekly in SARIMA. This is because the daily data gives the best result for this model. (More about this in the conclusion)

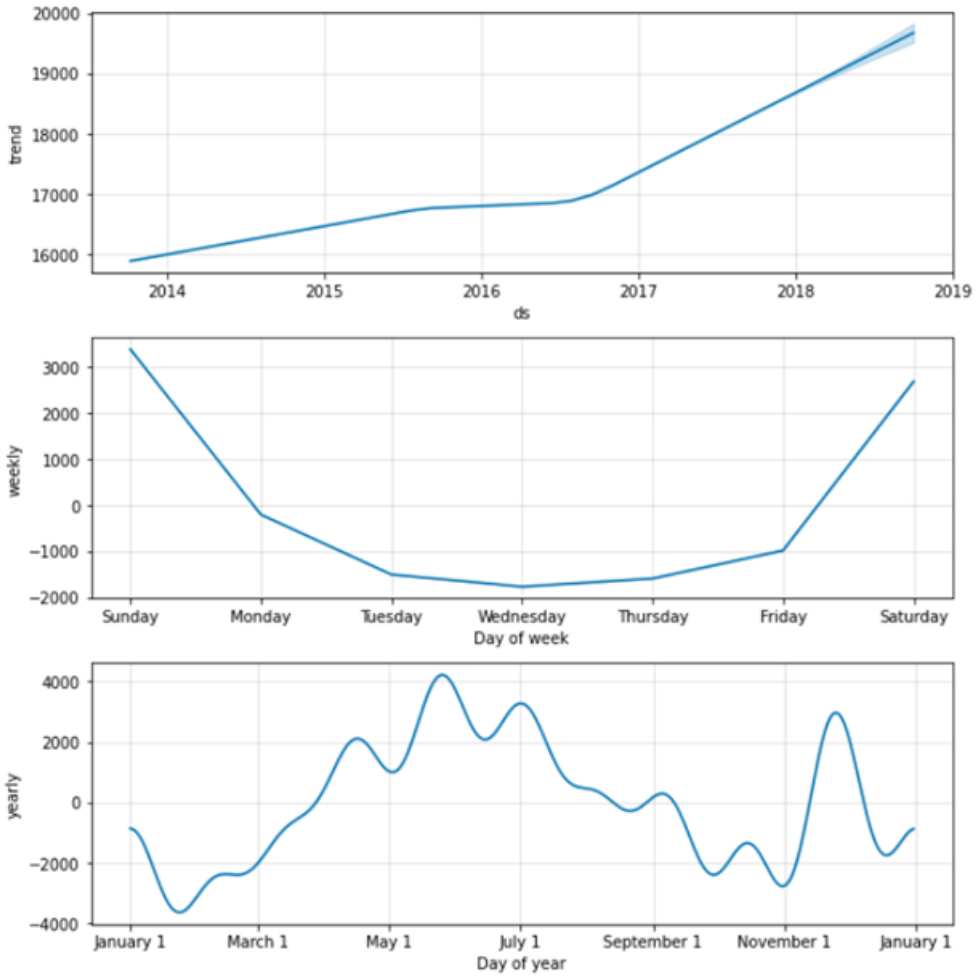
- Below is the scatter-plot of the data we use for this model.



- We split our data into train and test sets with the ratio of 80:20. The train portion of the data is then fit to the model. Using this model, we then forecast future values. The test data-set is used later for model validation.
- The image output for the model and its forecast is shown below.



- The black dots are our “true data”. The deep blue lines are the predictions from the model. The light blue shaded area is the confidence interval of our predictions. As we can see, from a certain point onward, we have only the prediction without our "true test" data. That is the forecast predicted by this model.
- Before moving onto the model evaluation, we would like to go through the trends picked up by the Prophet model:

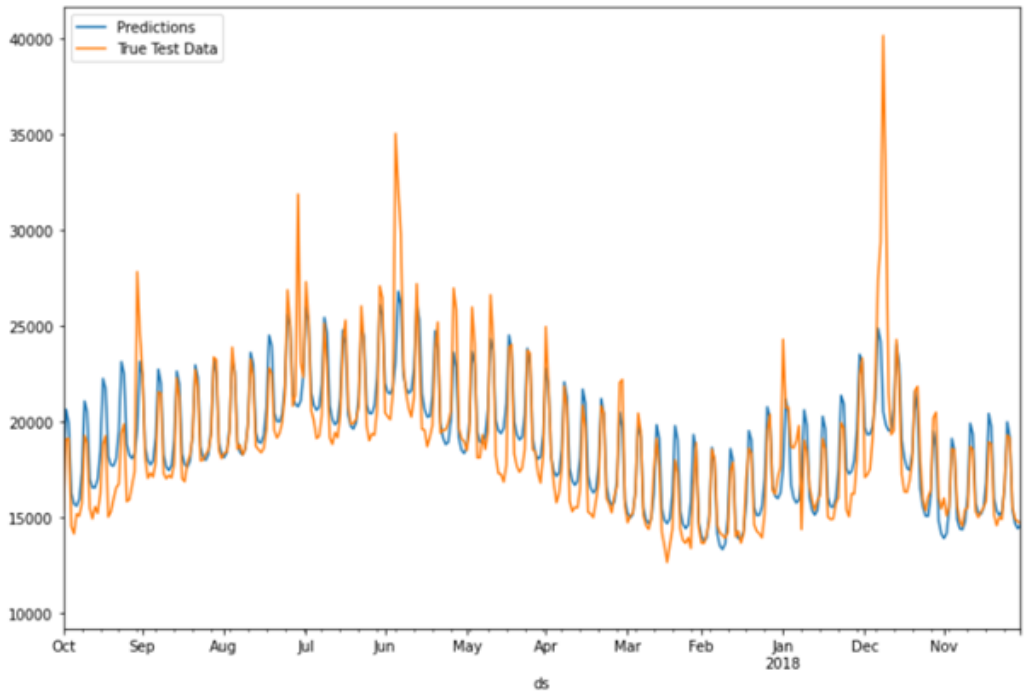


- Trends:
  - Overall: The overall trend is one of near linear increase in the search volume.
  - Weekly: The search volume is the maximum on the weekend and declines

steadily to the minimum on Wednesdays, after which it climbs back up when again as it nears the weekends. This trend is fairly obvious in a logical sense.

– Yearly: Clearly, searches jump up during the holiday seasons, including Christmas, and the Summer.

- Model Evaluation



- Error:

  – MSPE is 10.297184%

  – MAPE is 25.41556%

4 Conclusion

- Granularity/Resolution: For the FB Prophet Model, daily data gave the best results. Conversely, in our SARIMA models, weekly data gave us the best model.
- SARIMA model selection: The "best model" auto ARIMA presented to us was (1,1,1)x(1,1,0)x52. The model we thought would work best from analyzing the ACF/PACF plots was (1,1,5)x(1,1,0)x52. Surprisingly, the model which we analyzed from the ACF/PACF turned out to be better than the one found by auto ARIMA. (Criterion: MSPE and MAPE)
- Including the additional variables to the auto ARIMA model yielded a better result. Conversely, including the additional variables in the model analyzed from the graphs, yielded a worse result.(The difference in results for both models with additional variable was marginal, nearly identical)
- Overall, FB Prophet performed worse than SARIMA models.
- The best SARIMA model was the model we analyzed from the ACF/PACF graphs, with no additional variables, however only extremely marginally better than the one with additional variables. A surprising result.