**Web Traffic Forecasting**

**Report-1**

**Table of Contents**

1. Business Problem 2
2. Business Objective 2
3. Data Acquisition 2
4. Exploratory Data Analysis (EDA) 2

4.1 Missing Values Treatment 2

4.2 DataFrame Transforming 3

4.3 Selected Columns Exploration 3

1. Machine Learning Analysis 6

5.1 Data Preparation 6

5.2 Model Evaluation 7

5.3 Hyperparameter Tuning 8

1. Conclusions 10
2. **Business Problem**

The field of time series encapsulates many different problems, ranging from analysis and inference to classification and forecast. There are real-world problems out there corresponding to sequential or temporal observations. By utilizing time series data, we do can help predict future views and answer questions from financial market, pandemic spreading, climate change, et al. Here, we want to forecast the future web traffic. Sites could utilize the forecasting to determine if their sites are popular and if there are any apparent trends, such as one specific page being viewed mostly by people in a particular country. The web traffic forecasting is able to help structure sites, highlight security problems or indicate a potential lack of bandwidth.

1. **Business Objective**

The primary objective of this project is to build a machine learning model to predict the future web traffic. Relying on the previous records, time series analysis will be performed to deliver the most accurate machine learning model. It will ensure that sites are capable to apply the prediction to determine if their sites are popular and if there are any apparent trends, in the meantime to adjust their marketing strategies.

1. **Data Acquisition**

There are 6 datasets from the original data source: the first two train\_1 and train\_2 are the two main datasets with approximately 145k time series. Each of these time series represent a number of daily views of a different Wikipedia article, starting from July 1st, 2015 up until December 31st, 2016. Each row of these csv files corresponds to a particular article and each column correspond to a particular date. Key\_1 and key\_2 dataset give the mapping between the page names and the shortened ID column. And sample\_1\_submission and sample\_2\_submission are submission files showing the correct format.

The original data is available on <https://www.kaggle.com/c/web-traffic-time-series-forecasting>

1. **Exploratory Data Analysis (EDA)**

In this section, the main dataset will be analyzed with multiple visualization methods and their main characteristics will be summarized.

* 1. *Missing Values Treatment*

One of the most common problems faced in Data Cleaning/Exploratory Analysis is handling the missing values. There are several causes of missing values: sometimes values are missing because they do not exist, or because of improper collection of data or poor data entry. In that case, various filling strategies are required to operate for different situations. Here, since there is impossible for web traffic to be 0, unless unexpected situations, median number of each specific website is used to filling in missing values in this project.

* 1. *Dataframe Transformation*

In order to meet the requirement of Prophet model for machine learning, dataframe is melt to keep only two columns: ‘ds’ for date and ‘y’ for visits.

* 1. *Selected Columns Exploration*

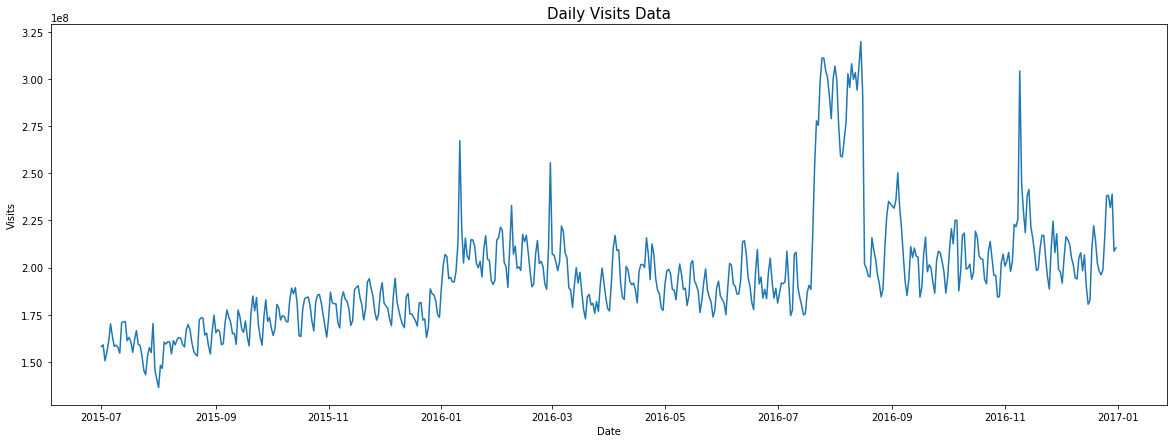


Fig 1. Daily Visits Data

It is clear to observe from figure 1 that there is huge spike around 2016-08, and several medium spikes during year 2016.

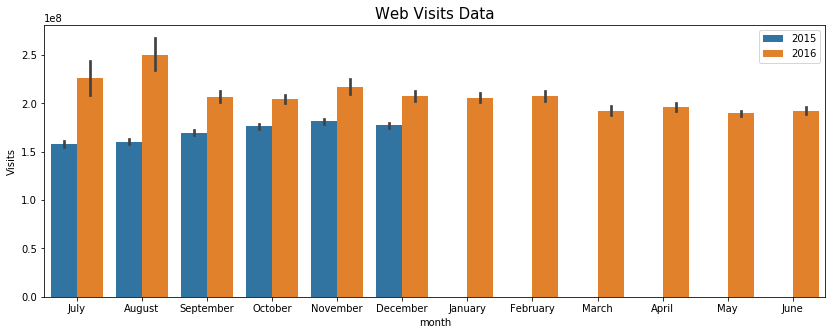


Fig 2. Web Visits Data Yearly Comparison

When look into the yearly comparison of web visits shown in Figure 2, there are spikes showing up in July and August in 2016; while in the latter half of 2015, there are more traffic in November. The whole year of 2016 has more visits than 2015.

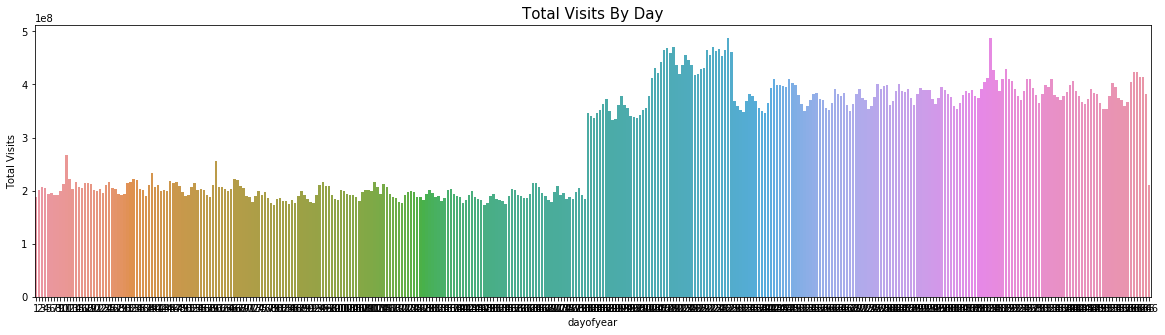


Fig 3. Daily Visits of Year

From figure 3, there is a higher plateau in year 2016 compared to year 2015 and more visits shown up in the middle of 2016.

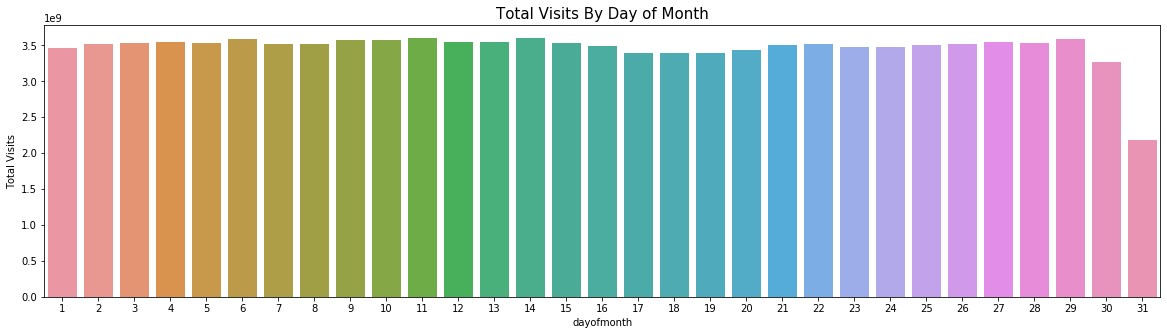


Fig 4. Day of Month Visits

The trend of visits on each day of month exhibited in figure 4. There is slightly more visits at the beginning of each month, while less visits displayed at the end of month.



Fig 5. Weekday Visits

According to weekday visits trend, there are similar visits during a week. Monday is slightly higher than the rest.

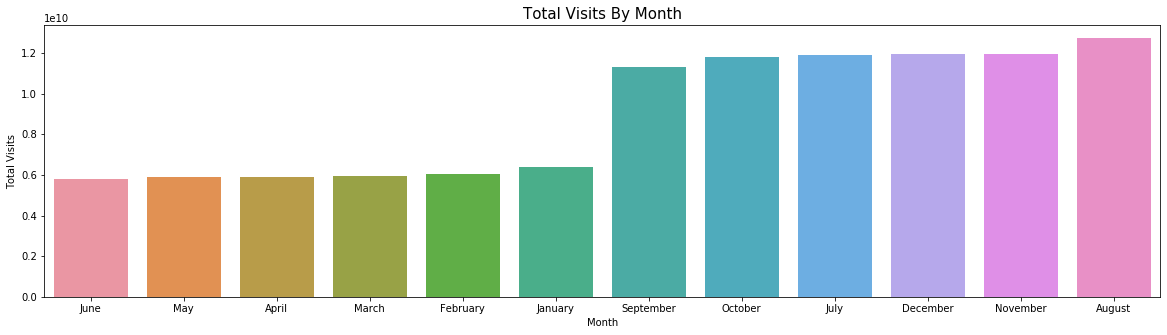


Fig 6. Monthly Visits

In month comparison figure 6, August has the most visits; and the rest months of latter half year gain more visits compared to the first half year.

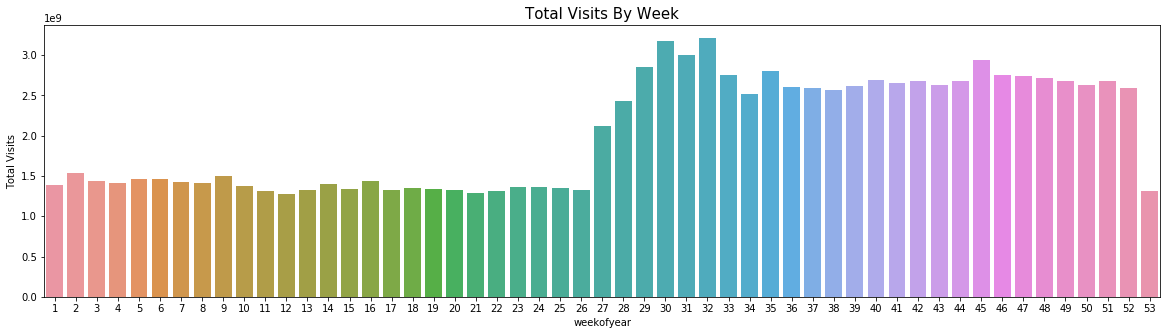


Fig 7. Week of Year Visits

Base on the week of the year visits, the latter half of year attracts more visits compared to the first half of the year.



Fig 8. Quarter Visits

It is clear to see that more visits happened in quarter 3 and 4 from figure 8.

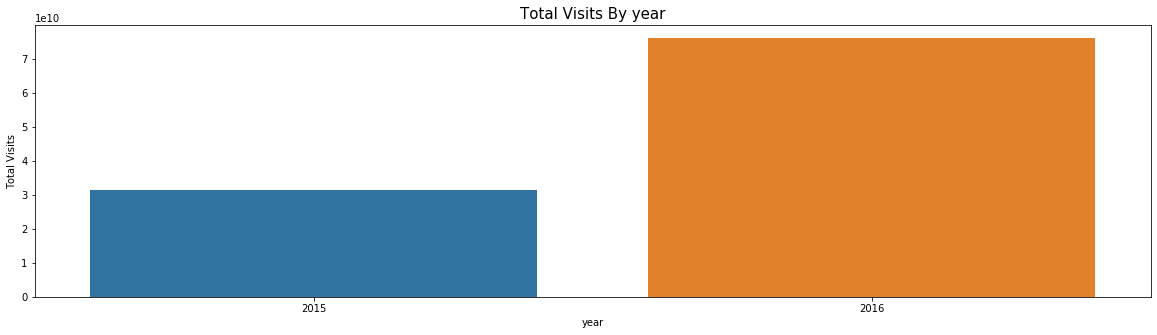


Fig 9. Yearly Visits

In figure 6, there is a dramatic visits increasing in year 2016.

1. **Machine Learning Analysis**

In this section, machine learning is utilized to build the model to predict the future web traffic.

* 1. *Data Preparation*

There are total 145063 rows representing each specific website and 551 columns representing website visits number, starting from July 1, 2015 to Dec 31, 2016, in the main dataset “train\_1.csv”. Some of the missing values are occupied the whole row, which are deleted since only a small portion. As mentioned early, the strategy of filling in missing values with median is applied in this project.

According to the requirement of FB Prophet model, only two columns are kept for the further evaluation, ‘ds’ representing date and ‘y’ representing visits. The dataset has been transformed to meet the requirement. In order to achieve the most accurate prediction, hyperparameter tuning is also performed, MAPE (mean absolute percentage error) number is calculated for each approach and compared to get the best parameter combination for model tuning.

Dataset is split into two parts for model evaluation, 70% for training and 30% for test.

* 1. *Model Evaluation*

FB Prophet model is applied for prediction. Below is the plot including current data and next 180 days prediction.

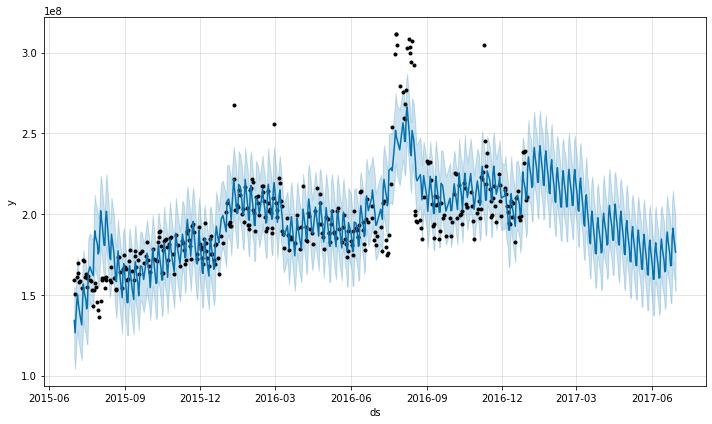


Fig 10. Prophet Model Prediction

As shown in figure 10, black dots represent the real data, blue line and light blue background represent the prediction number range. The whole prediction almost match the current real data and show the reasonable future trend. rmse (root mean square error) value is 16764868. MAPE for the parameter combination is 39.6.

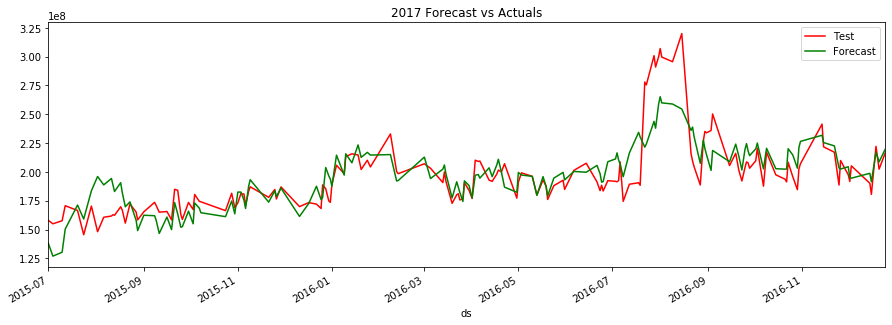


Fig 11. Comparison of Test and Forecast Data

Above are plots (figure 11) exhibiting the trend comparison between current data from July 1st 2015 to Dec 31st 2016 and prediction data based on those. It is clear to notice there are similar trend for both plots, which means the prediction is reliable and only tuning is required in order to achieve the highest accuracy.

* 1. *Hyperparameter Tuning*

In this section, hyperparameter tuning/optimization are applied, a set of optimal hyperparameters for a machine learning algorithm will tuned, and rmse number is the evaluation criteria here, the smaller rmse, the better prediction accuracy.

Here, several parameters, including ‘seasonality\_mode’, ‘changepoint\_prior\_scale’, ‘holidays\_prior\_scale’ and ‘n\_changepoints’ are searched via parameter grid strategy to obtain the best parameters, which shows the lowest MAPE value.

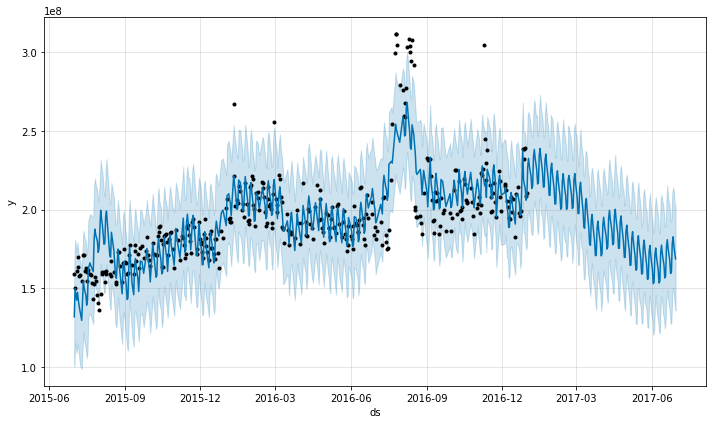


Fig 12. Prophet Model Prediction after Optimization

Figure 12 displays the Prophet model prediction plot after hyperparameters tuning. The rmse value for this trial is 16400257, which is better than the first trial.

Furthermore, besides the tuned parameters mentioned, seasonality parameters are also added into the grid search for best options.

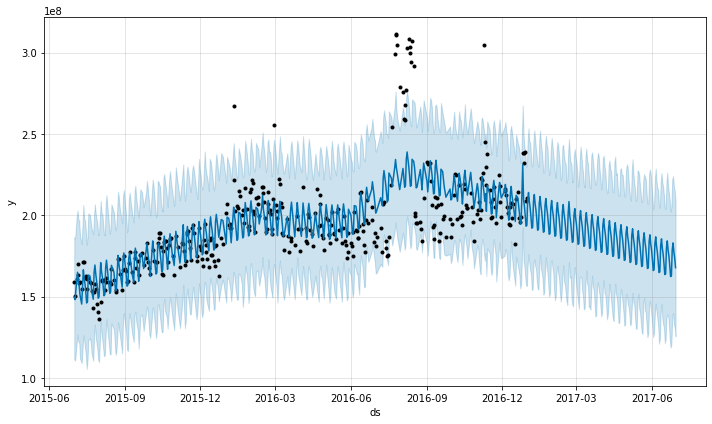


Fig 13. Prophet Model Prediction after Optimization

As shown in Figure 13, the plot displayed the trend after adding seasonality parameter tuning. The rmse value for this trial is 19477487, which is getting worse.

|  |  |  |
| --- | --- | --- |
|  | **rmse** | **MAPE** |
| ***Trial 1*** | 16764868 | 39.6 |
| ***Trial 2*** | 16400257 | 38.0 |
| ***Trial 3*** | 19477487 | 25.1 |

Table 1. Comparison of rmse and MAPE

The comparison of rmse values for each run of evaluation and MAPE value for parameter selection corresponding to each run is shown in Table 1. While Trial 3 delivers the lowest MAPE value for parameter choosing, trial 2 exhibits the lowest rmse value so that trial 2 model will be chose for the future web traffic forecasting with the highest accuracy.

1. **Conclusions**

Exploratory data analysis is performed and FB Prophet machine learning models is evaluated for time series analysis. Hyperparmeter tuning is applied to achieve the best parameters to get the highest accuracy of the prediction model.

rmse (root mean square error) value is the evaluation criteria for prediction and has been calculated for each run of model evaluation. MAPE (mean absolute percentage error) number is obtained and compared to achieve the best parameter combination from hyperparameter tuning. The best model displays rmse value 16400257, compared to original rmse 16764868.

The machine learning models can be used for prediction of the future web traffic. . It will ensure that sites could utilize the forecasting to determine if their sites are popular and if there are any apparent trends, such as one specific page being viewed mostly by people in a particular country. The web traffic forecasting is able to help structure sites, highlight security problems or indicate a potential lack of bandwidth.