Web traffic predictions

Sailaja Karra

Business case

With increasing availability of data for example items sold in a particular store or page views on wikipedia or trends in social media. It is always beneficial to predict these using prior data.

But with increased amount of data and model complexity i.e. trends, seasonality (weekly, monthly, yearly etc) we need to be able to model these independently and it's crucial to be able to run this in parallel.

Dataset and EDA

Initial Dataset:

145,000 rows
550 Columns

145k wiki page views
from 1st July 2015
to 31st December 2016

Challenge:

Given the huge amount of data, pandas is not able to handle the data. Since we cannot parallelize in pandas even on high spec computer EDA and modelling are not efficient.

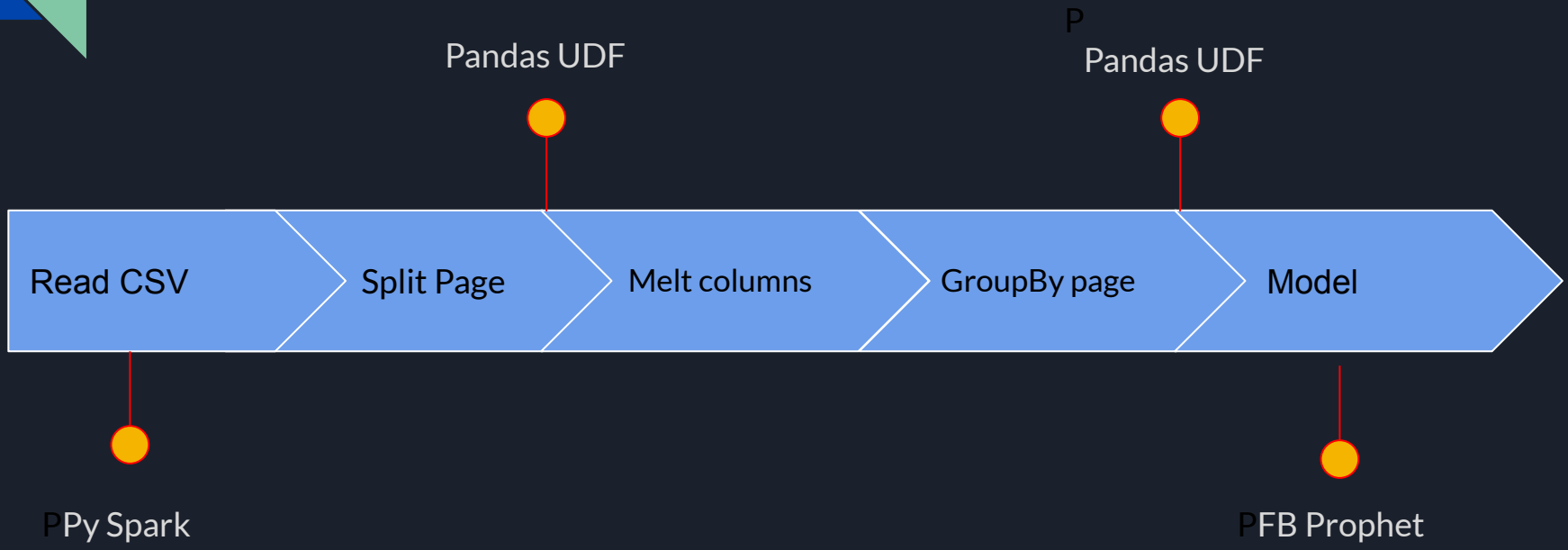
Cleaning Process:

Split page column into searchterm, language, access, agent.

Melted date columns.

Both need pyspark to be able to run this efficiently

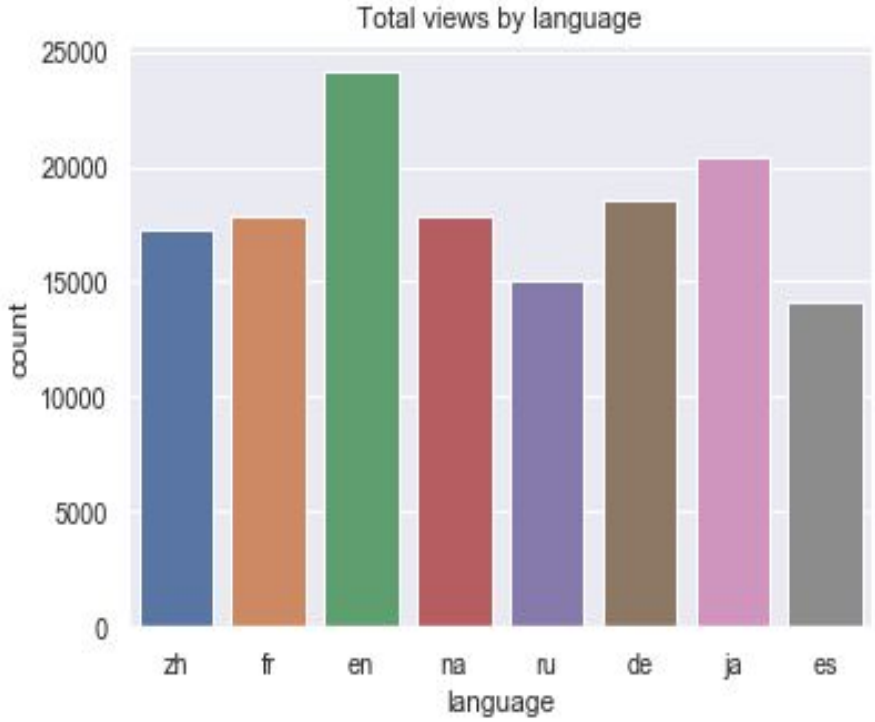
PROCESS FLOW CHART



Views by Language

This shows number of page views by language.

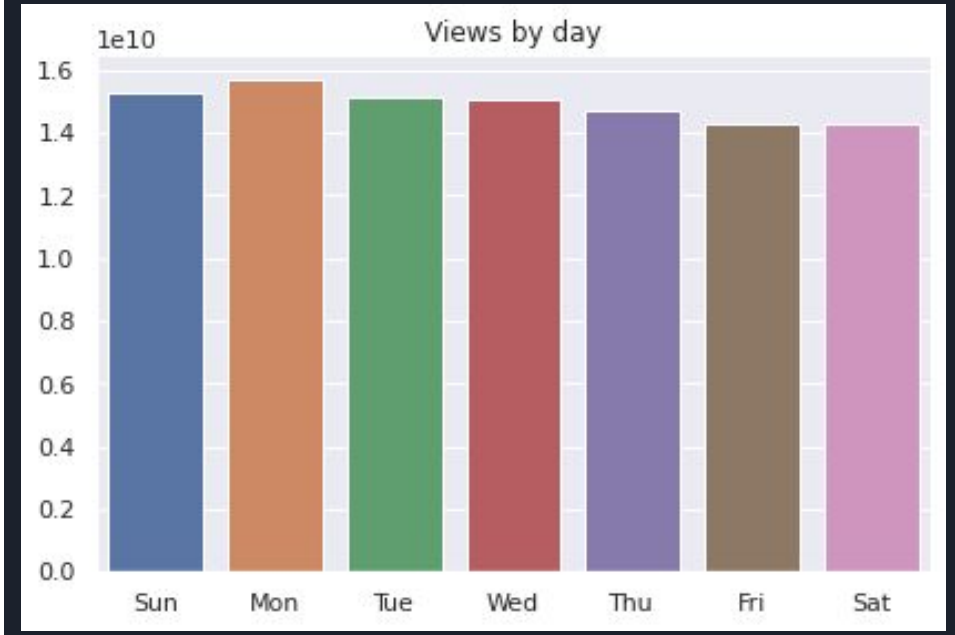
The views are high for english and are evenly distributed for rest of the languages.



Views by day

This shows number of page views by day of the week.

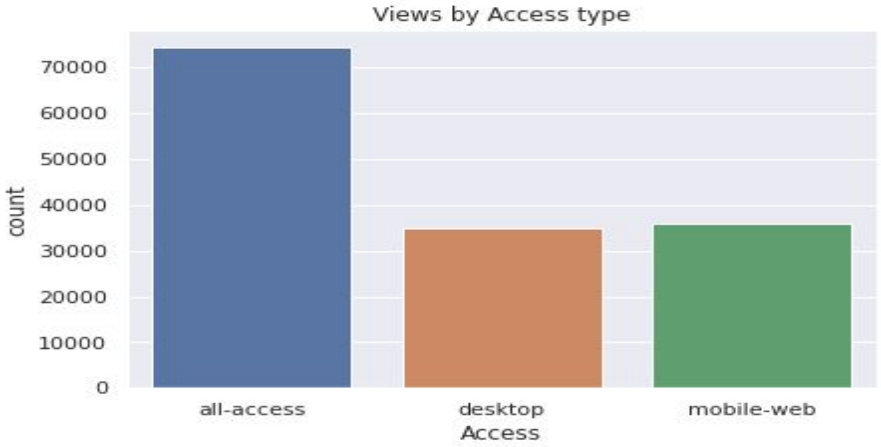
The views are high for Monday and are evenly distributed for rest of the days.

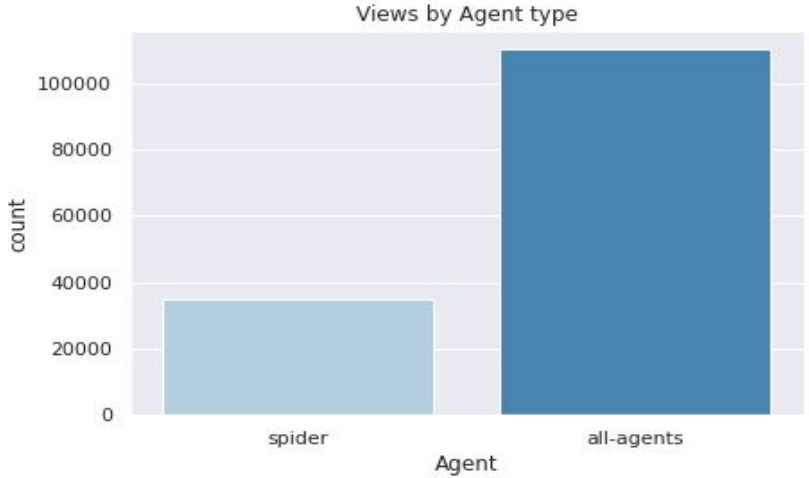


Views by Type

This shows number of page views by access type.

This shows number of page views by agent.



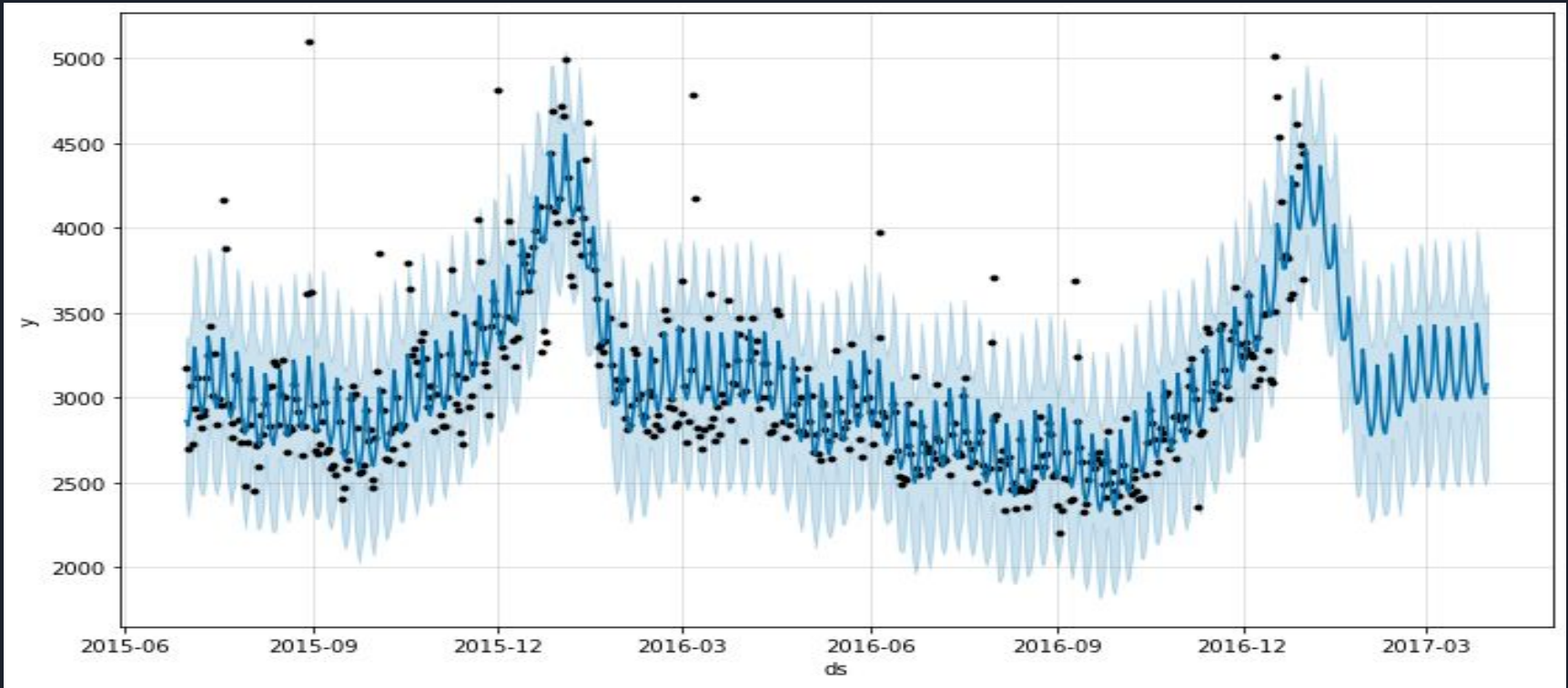


Fast and accurate

Fully automatic

Tunable forecasts

Lord of the rings wiki page prediction

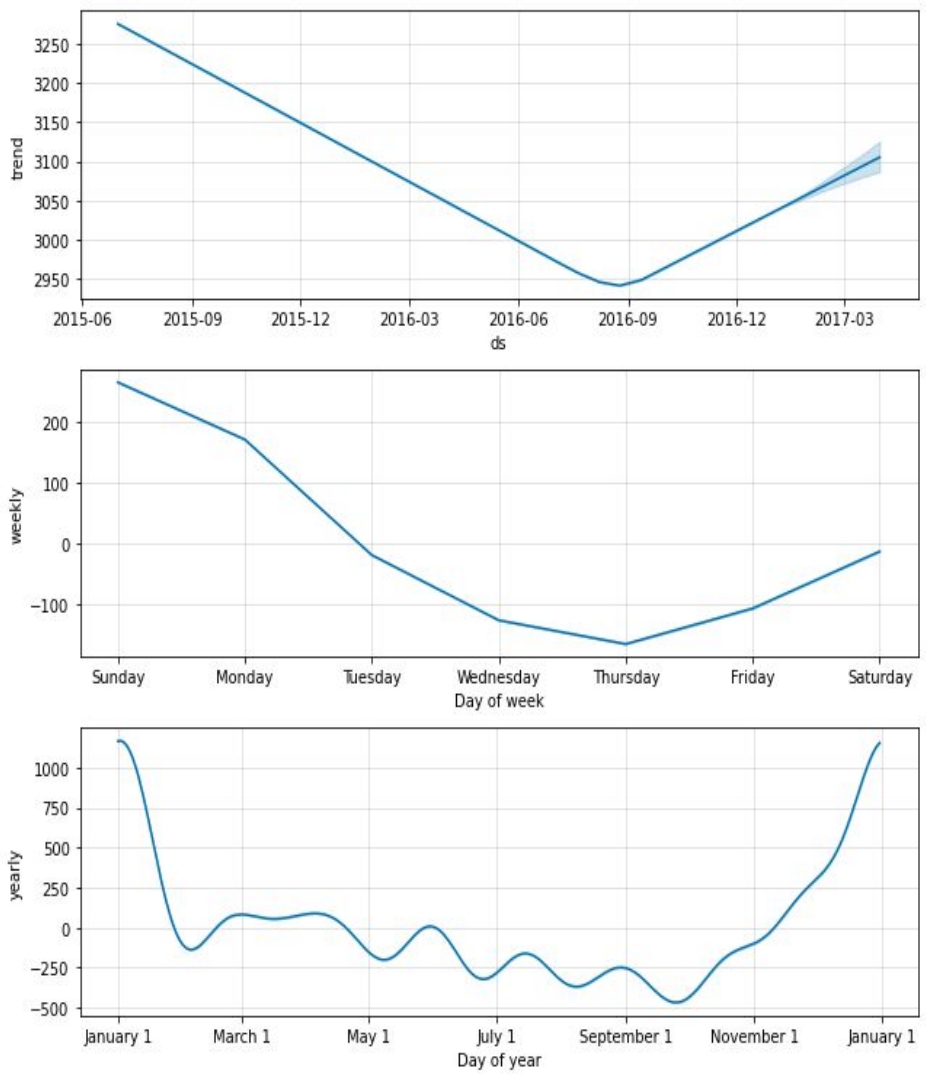


Prophet model prediction for Lord of the rings wiki page views time series.
It follows nicely and gets both the trends and seasonality correct.

Decomposition of Time series

My model shows the time series decomposition of the trend, weekly and yearly seasonality.

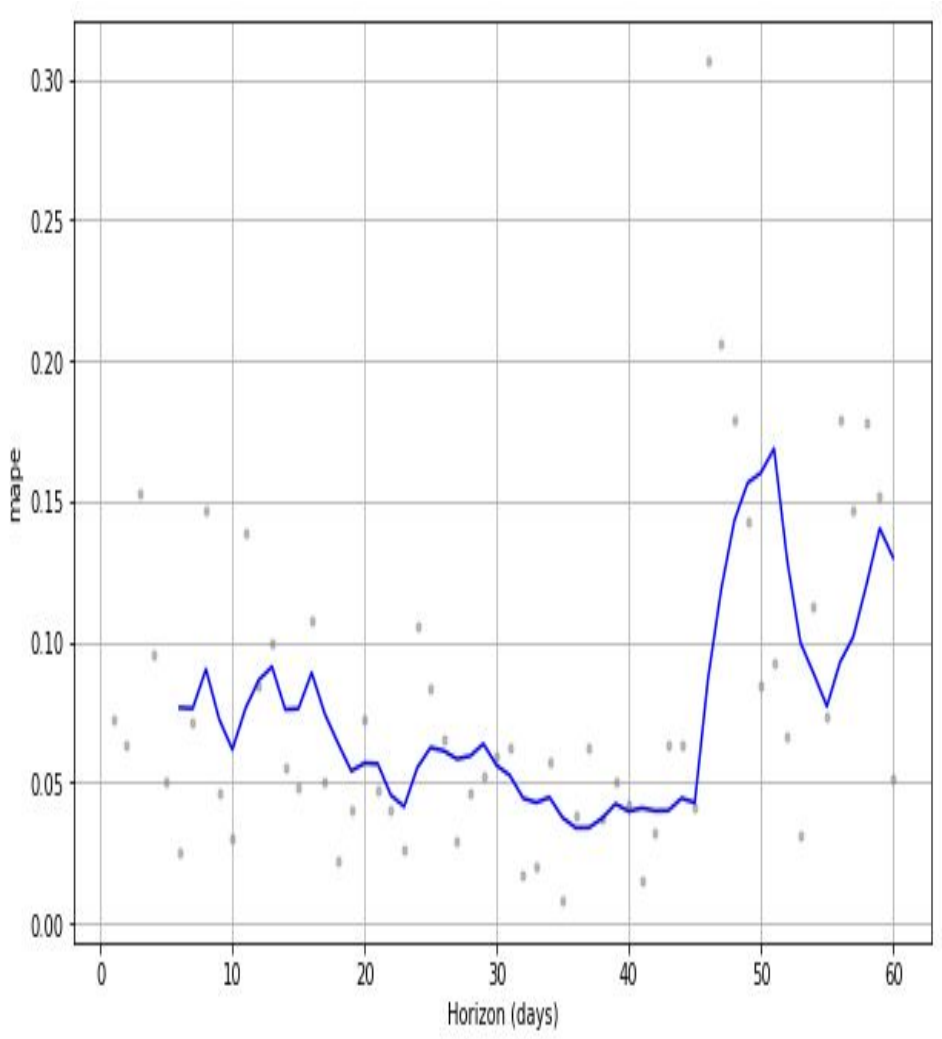
We can further fine tune the seasonality to specific events like NFL playoffs or spring break week etc.



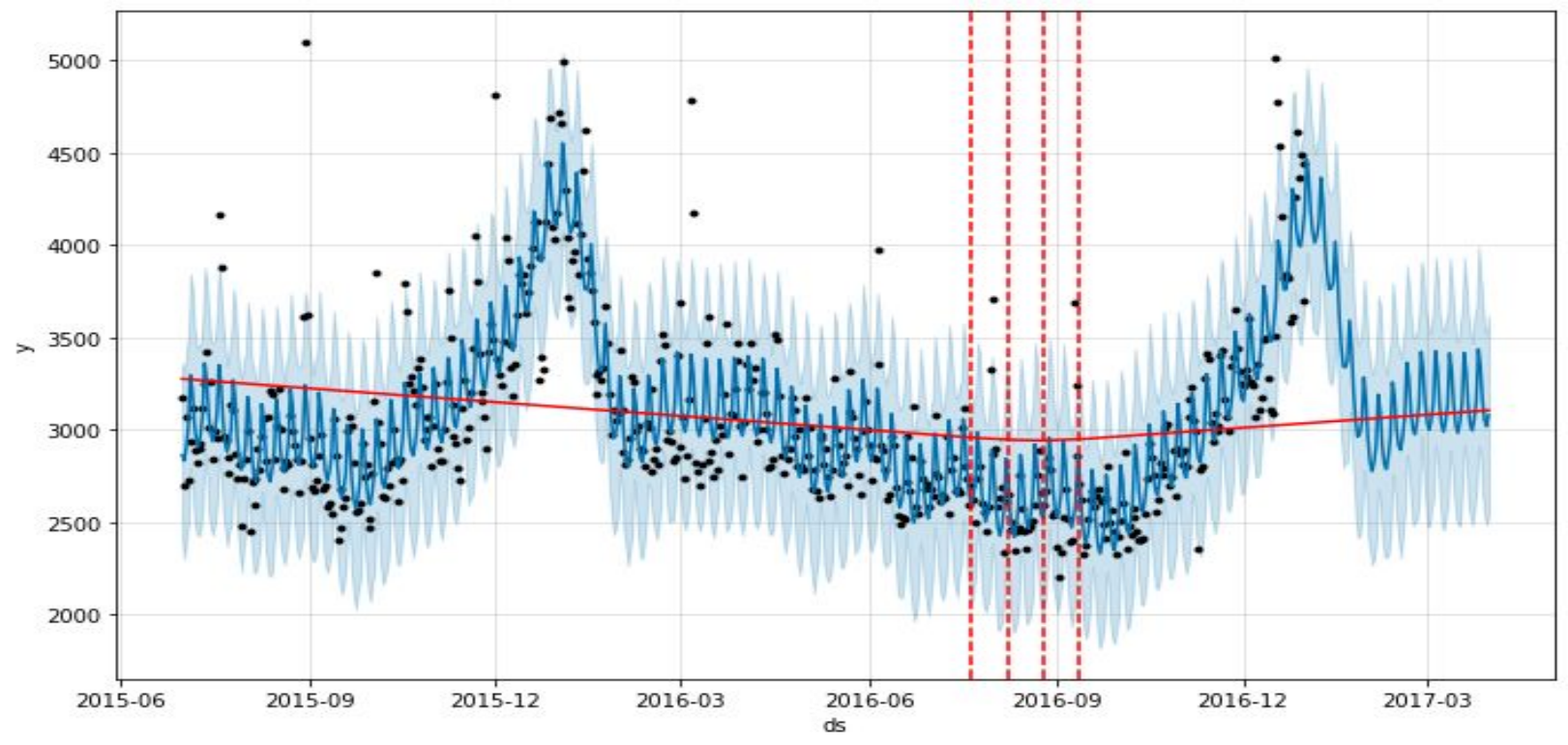
Model Metrics

MAPE: Mean absolute percentage error.

This shows in percentage how off the predictions are. As expected the error is small for few days of prediction but increases as we project for longer days.



Change Points



Prophet not only gives nice predictions and decomposition of the time series, it also shows where the **trend** has **changed**.

Further things to do

01 **Holidays:** We can improve our model prediction by additional country level holidays.

02 **Custom Seasonality:** We can also add specific custom seasonality like NFL final week, french open etc

03 **Deep Neural networks:** We can try to model this out as a deep neural network with memory units like LSTM or GRU and compare models

I hope you enjoyed this presentation.

Thank you!