Final Presentation

By Charles Yau

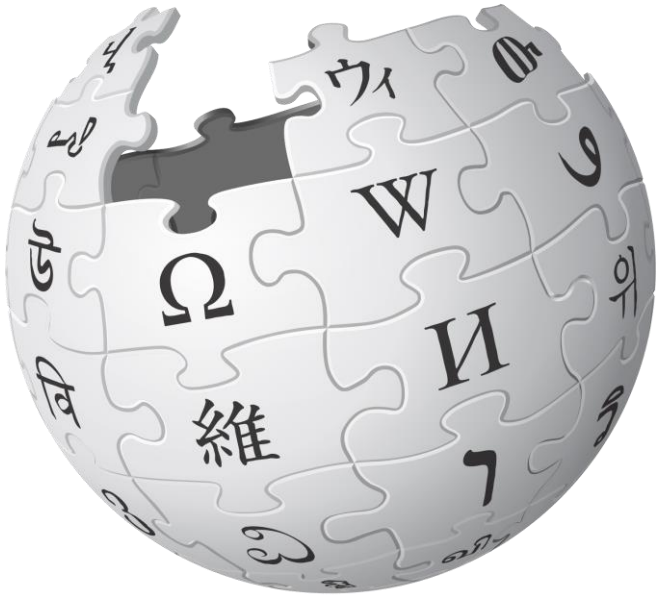
4th August 2017

DISCLAIMER
Change of Project

Project – Predicting web traffic to specific web pages
Problem Statement – How can we remain competitive in this digitized world?

- Digital Transformation
- Opened businesses to the world – SMEs and MNCs had an entirely new market/demand to tap into.
- Fully integrated stores with no brick and mortar store fronts
- Marginal costs and the “race to the bottom”
- Vendors with the lowest margin/largest economies of scale wins.
- USD 250 for 100 visitors a day for a month

Project Dataset – Wikipedia pages



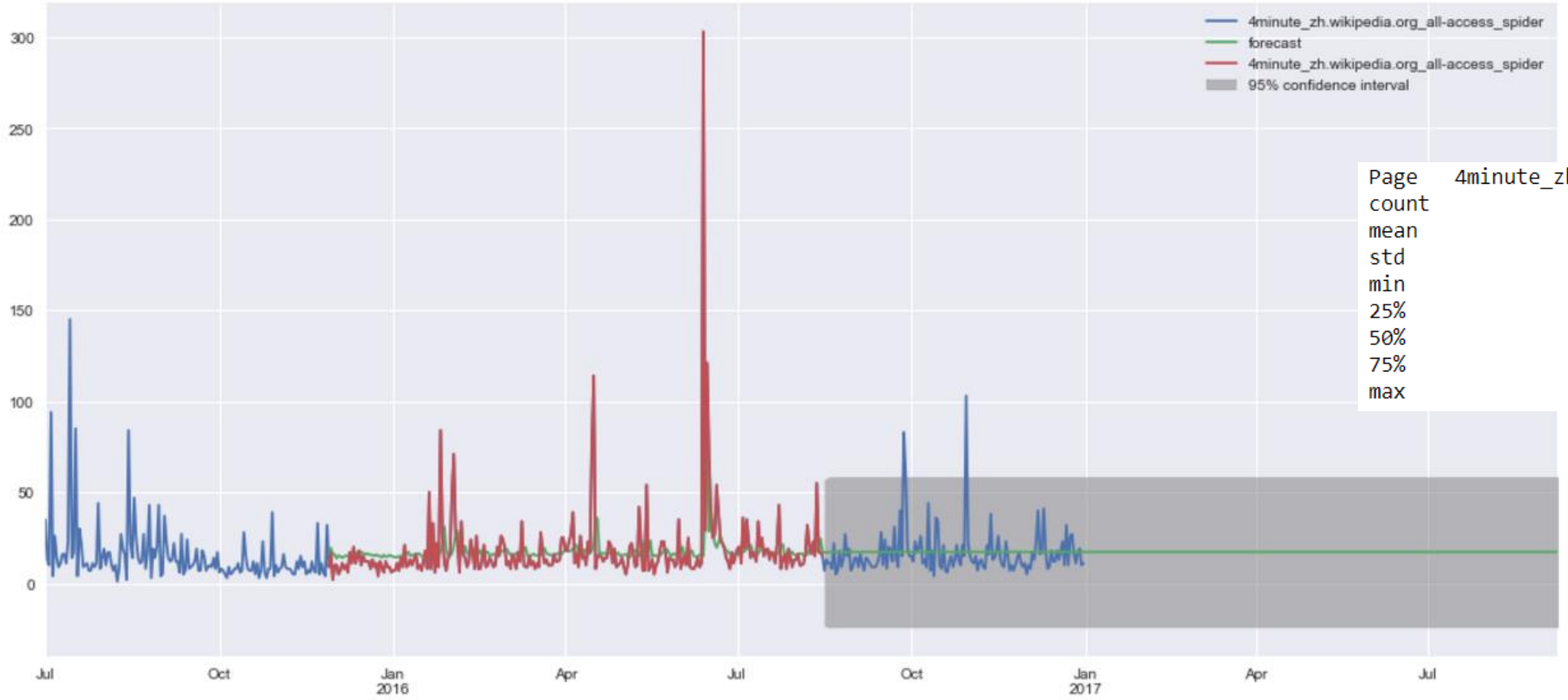
WIKIPEDIA
The Free Encyclopedia

- Dataset gathered from Google
- Data is straightforward and has minimal empty data entries.
- There are seven different languages but are not taken into consideration
- Attempting to gain a one size fit all model which can provide a close/accurate forecast on web traffic using historical data
- It went through exploration phase, testing out a variety of models, predictions and testing predicted against actual

Introduction to the Data

	Page	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	2015-07-08	2015-07-09	...	2016-12-22	2016-12-23	2016-12-24	2016-12-25	2016-12-26	2016-12-27	2(1:
0	2NE1_zh.wikipedia.org_all-access_spider	18.0	11.0	5.0	13.0	14.0	9.0	9.0	22.0	26.0	...	32.0	63.0	15.0	26.0	14.0	20.0	2:
1	2PM_zh.wikipedia.org_all-access_spider	11.0	14.0	15.0	18.0	11.0	13.0	22.0	11.0	10.0	...	17.0	42.0	28.0	15.0	9.0	30.0	5:
2	3C_zh.wikipedia.org_all-access_spider	1.0	0.0	1.0	1.0	0.0	4.0	0.0	3.0	4.0	...	3.0	1.0	1.0	7.0	4.0	4.0	6.
3	4minute_zh.wikipedia.org_all-access_spider	35.0	13.0	10.0	94.0	4.0	26.0	14.0	9.0	11.0	...	32.0	10.0	26.0	27.0	16.0	11.0	1:

The Model – Trial 1

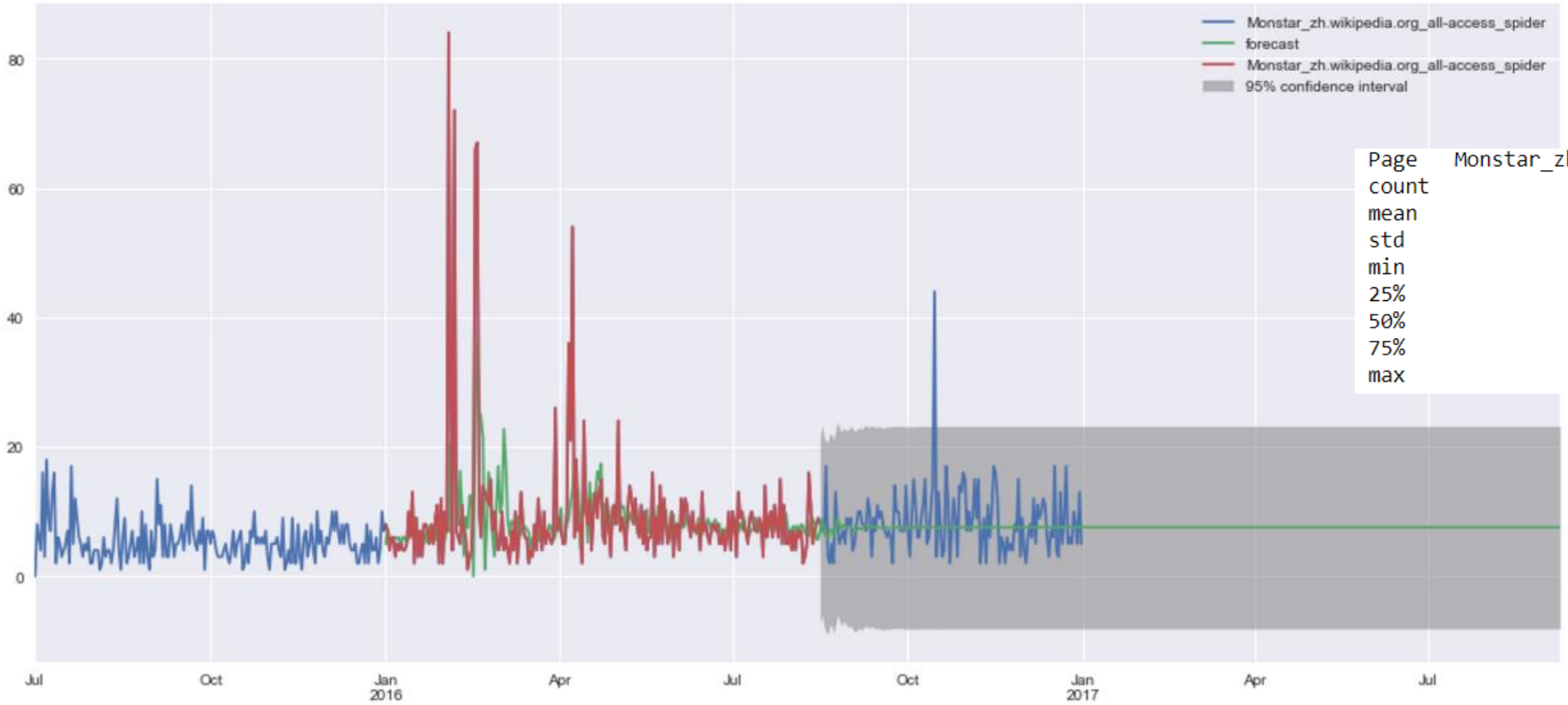


Page	4minute_zh.wikipedia.org_all-access_spider
count	550.000000
mean	17.125455
std	19.337118
min	1.000000
25%	9.000000
50%	13.000000
75%	18.750000
max	303.000000

Caveat:

- Doesn’t take into account seasonality or spikes

The Model – Trial 2



Page	Monstar_zh.wikipedia.org_all-access_spider
count	550.000000
mean	7.769091
std	7.370704
min	0.000000
25%	4.000000
50%	7.000000
75%	9.000000
max	84.000000

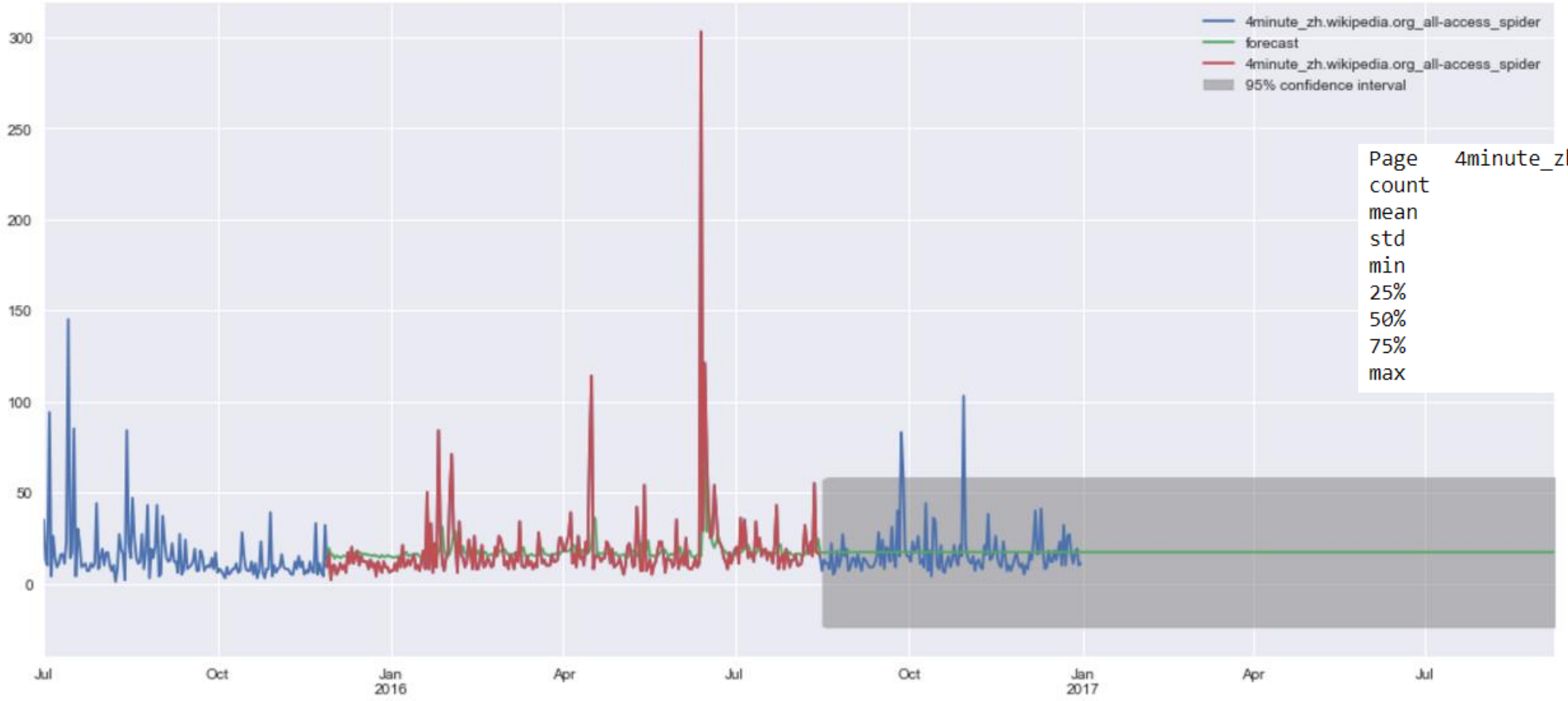
Caveat:

- Doesn't take into account seasonality or spikes

Utilized ARIMA model to generate the model

- Attempted Prophet, ARIMA, RMSE, Autocorrelation and by hand
- Prophet could not provide a enough granular insights
- ARIMA provided a 95% confidence interval and is far more modular
- ARIMA provided a good start point for future predictions

Trial 1 results

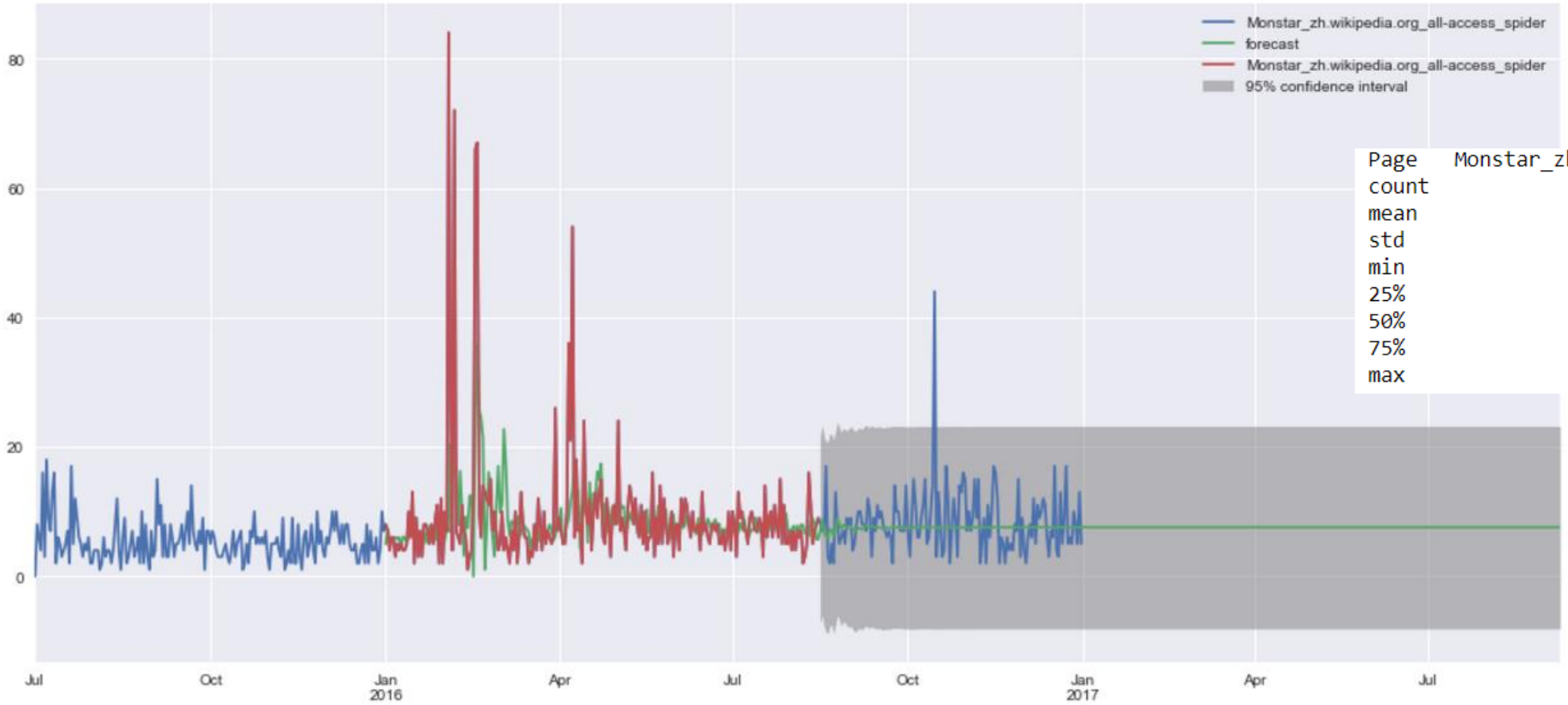


Page	4minute_zh.wikipedia.org_all-access_spider
count	550.000000
mean	17.125455
std	19.337118
min	1.000000
25%	9.000000
50%	13.000000
75%	18.750000
max	303.000000

RMSE = 15

Difference between forecast and actual

Trial 2 results



Page	Monstar_zh.wikipedia.org_all-access_spider
count	550.000000
mean	7.769091
std	7.370704
min	0.000000
25%	4.000000
50%	7.000000
75%	9.000000
max	84.000000

RMSE = 5

Difference between forecast and actual

Overall not a bad result

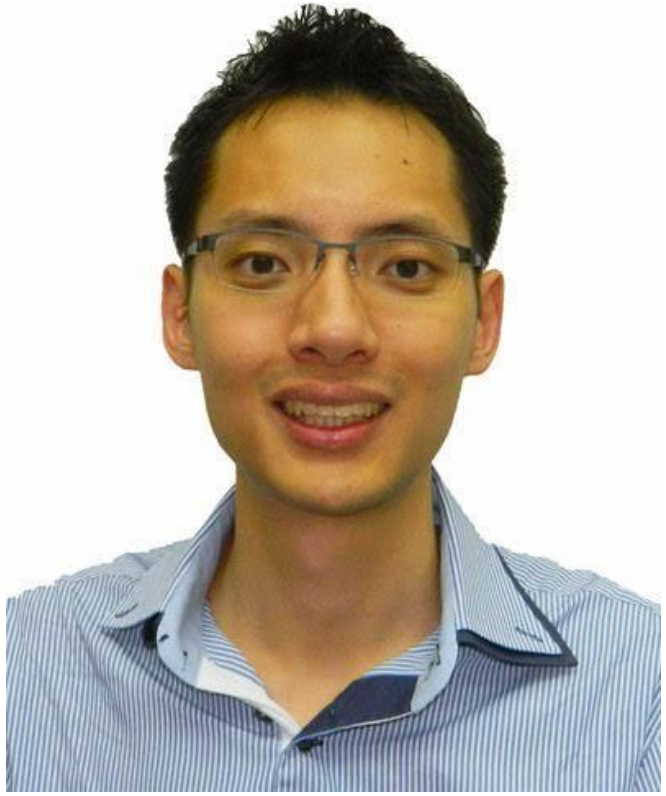
- Average but satisfactory results
- Spikes and seasonality threw off the predictions particularly at high anomalies like in Trial 1
- I can confirm that it is possible to gain a understanding of the threshold that is needed for bandwidth allocation
- I suggest removing spikes in the future to make the predictions slightly more accurate

Next Steps

- Tweaking and refining of model is still needed
- Other models can be used to see if there are better ways of predicting future web traffic
- XGBoost, heatmaps for general analysis

SPECIAL THANKS

Jocelyn Ong



Kwan Chong Tan

Thanks for listening!

Q&A