Stanza Data Challenge

CPM Detection System

Yufei Wang

USC Business Analytics

CONTENTS

Executive Summary

Project Goal:

Build a notification system to identify daily abnormal CPM performance for 173 sites

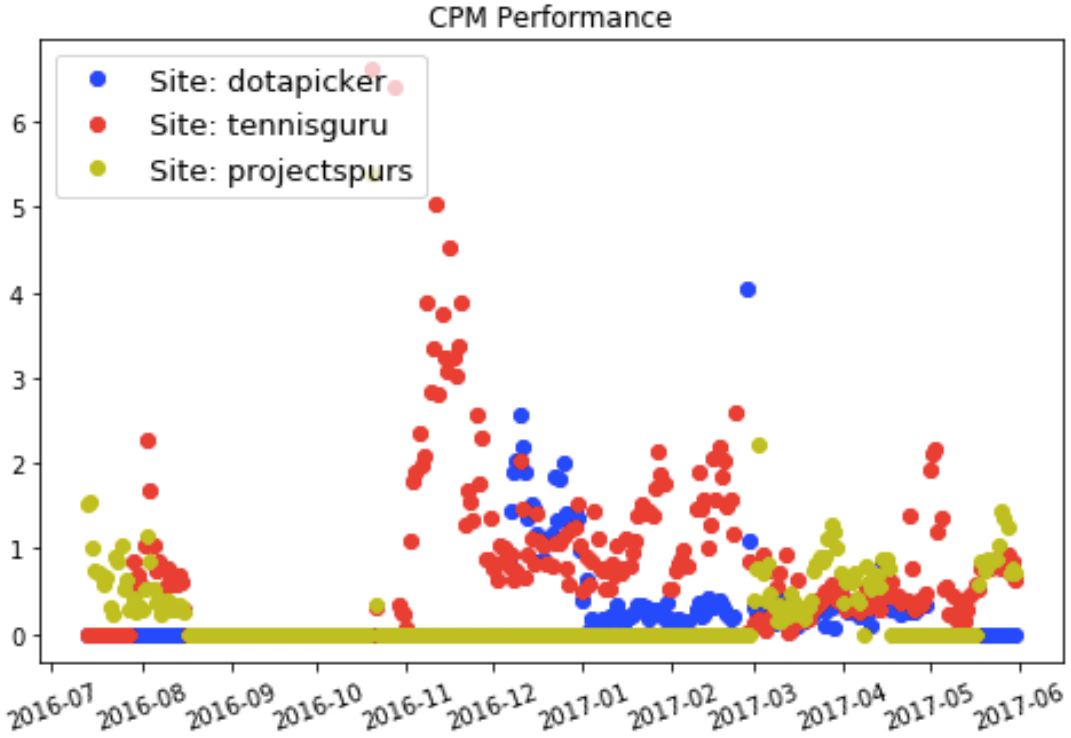
Analysis Results:

- Implemented 5 time series models to forecast CPM on given date
- Compare prediction with real data and offer evaluations of daily performance

Next Steps:

Every site has its special seasonality and trend, in order to increase the accuracy of prediction, we would better to build customized model for each site.

Data Exploration

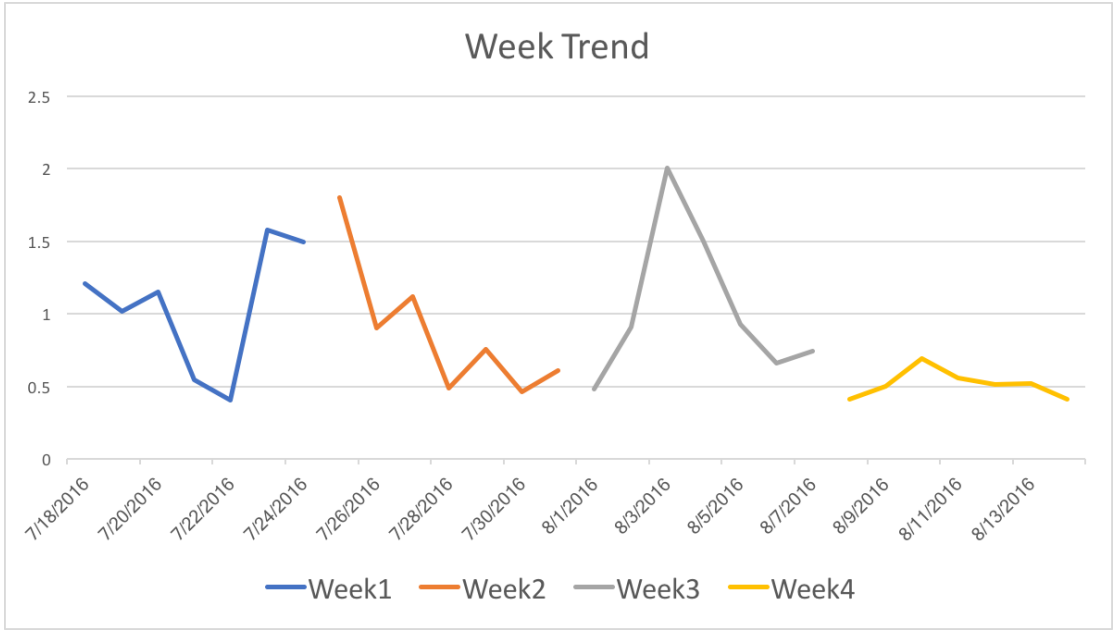


- The CPM performance of different sites are significantly different
- The CPM records between 2016-8-15 to 2016-10-15 are lost
- For site tennisguru, there is a peak of CPM in the second month of every quarter
- The monthly trend is not clear

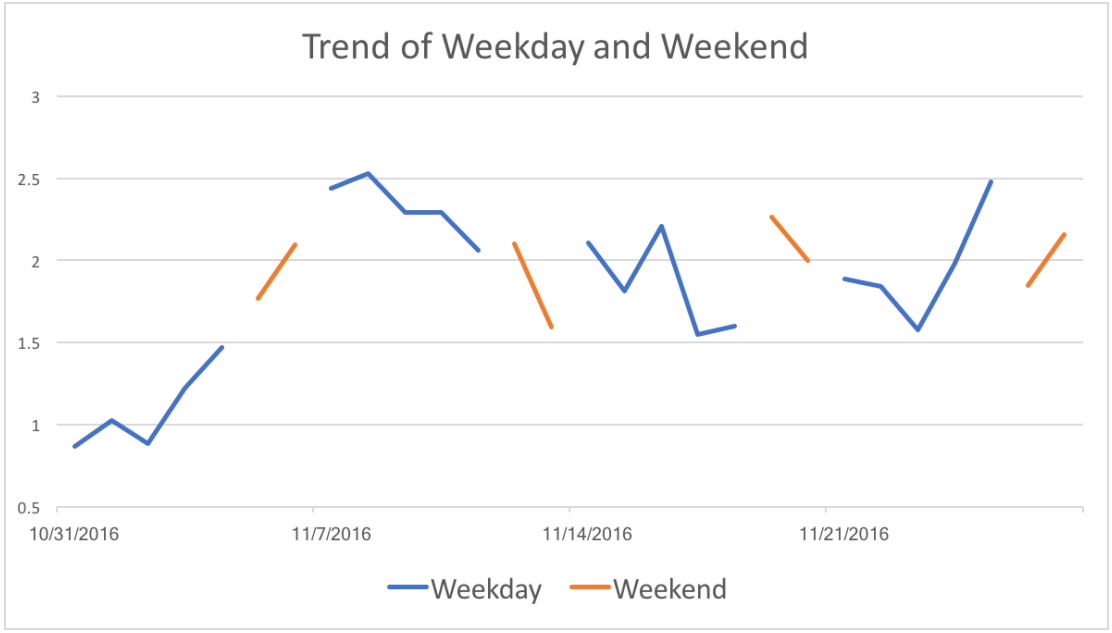
Conclusion:

Implement distinct models for each site

Data Exploration



There is no obvious week pattern of CPM performance of site therepublikofmancunia



The pattern between weekday's CPM and weekend's CPM is not stationary.

Conclusion: Try multiple seasonality like week(7 days), month(30 days), quarter(90 days) to fit models

Data Manipulation

- **Expand Data**

  Add missing daily CPM records for 173 sites from 2016-7-13 to 2017-5-31

- **Fill NA with 0**

  Assume the revenue and impressions are zero if the data point is not present.

- **Get the Site with Best CPM Performance**

  Sort sites by valid data points that CPM > 0, then use the data of site therepublikofmancunia to build model

- **Training / Testing Dataset**

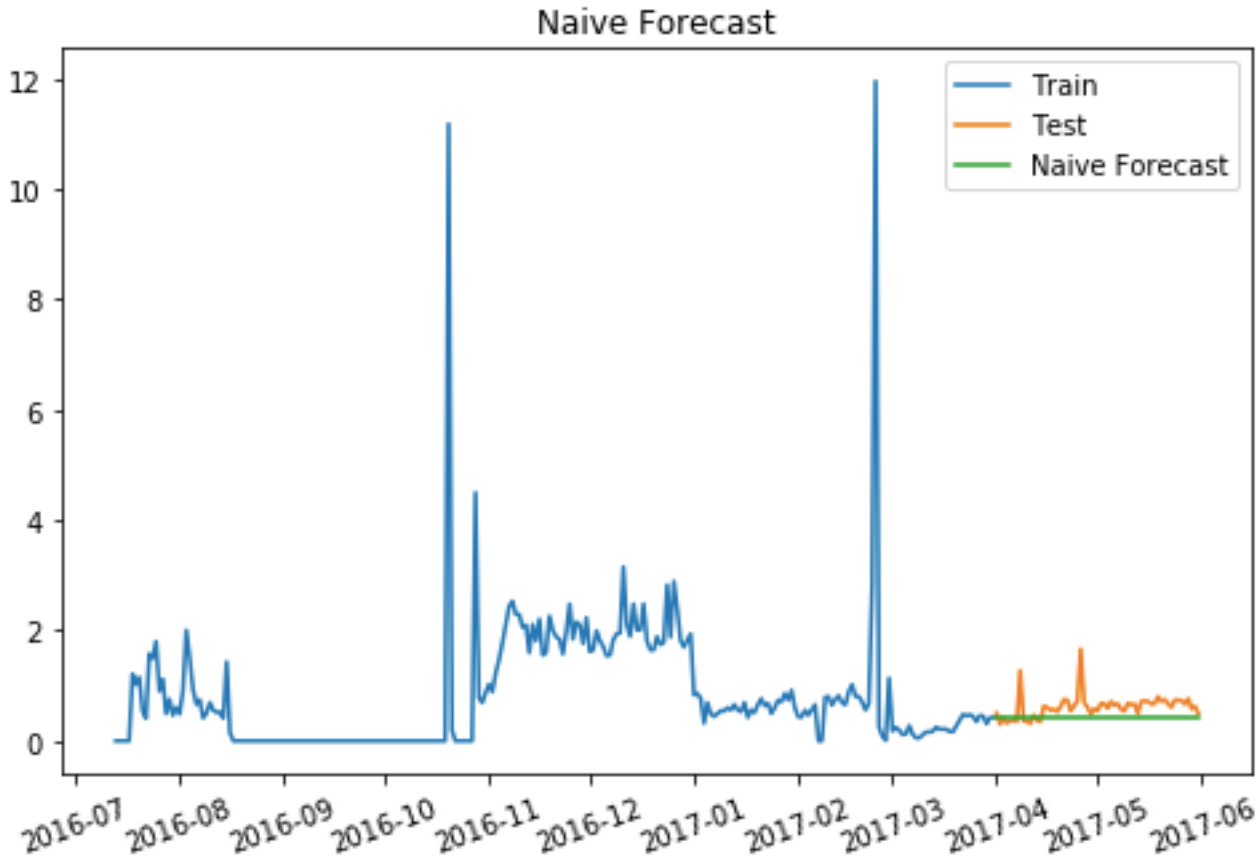
  Training Data: 262 Records from 2016-7-13 to 2017-3-31

  Testing Data: 61 Records from 2017-4-1 to 2017-5-31

Model Building

Naïve Model

RMSE: 0.281969



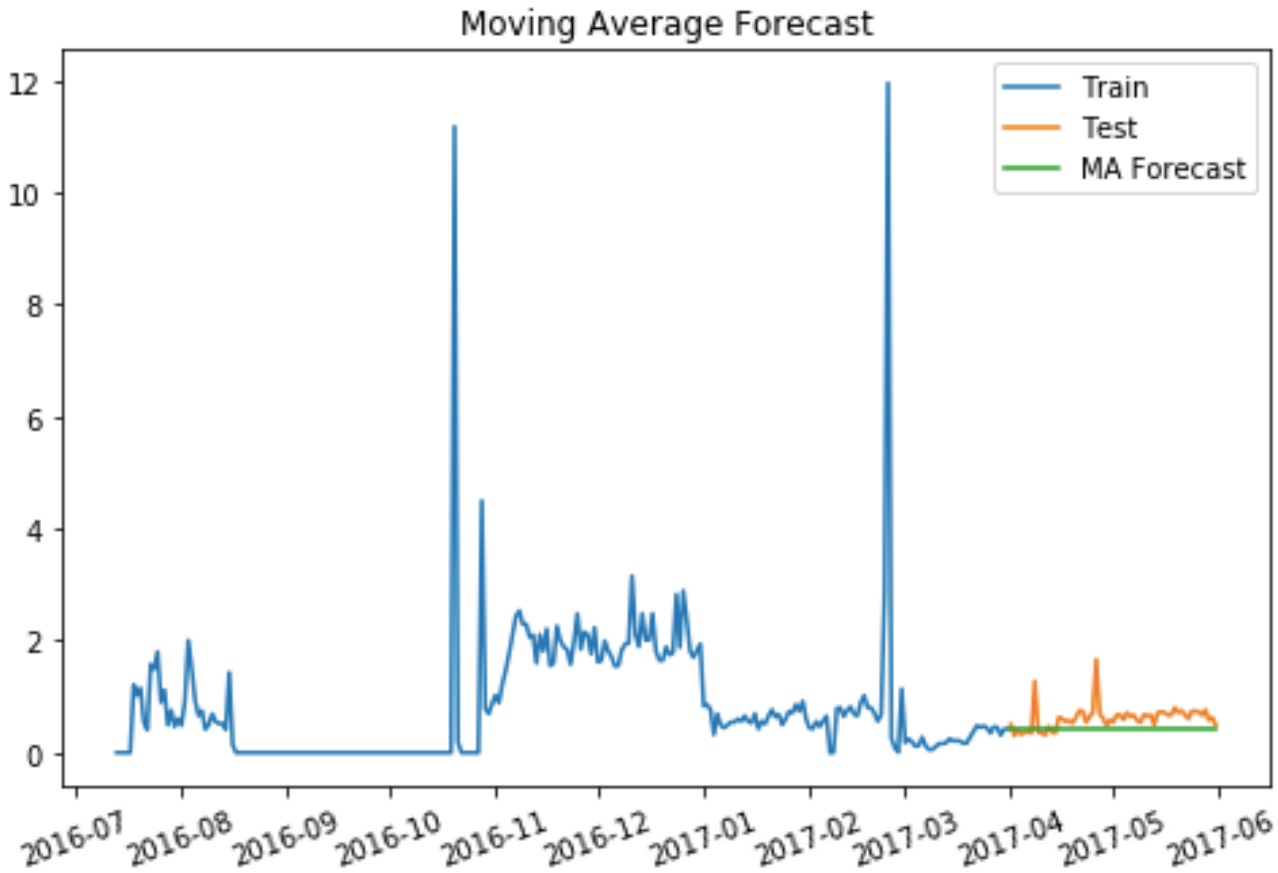
Model Building

Naïve Model

RMSE: 0.281969

Moving Average Model

RMSE: 0.282545



Model Building

Naïve Model

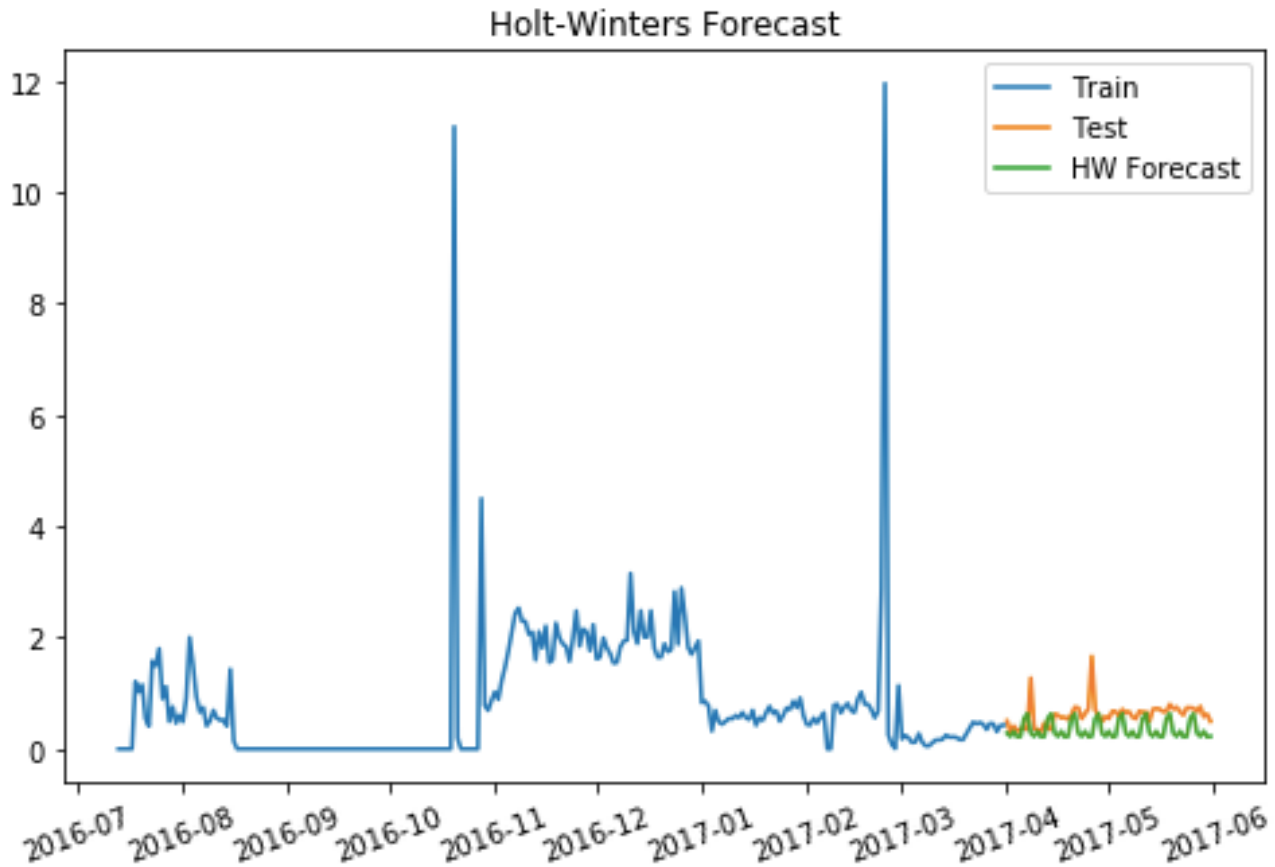
RMSE: 0.281969

Moving Average Model

RMSE: 0.282545

Holt Winters Model

RMSE: 0.379134



Model Building

Naïve Model

RMSE: 0.281969

Moving Average Model

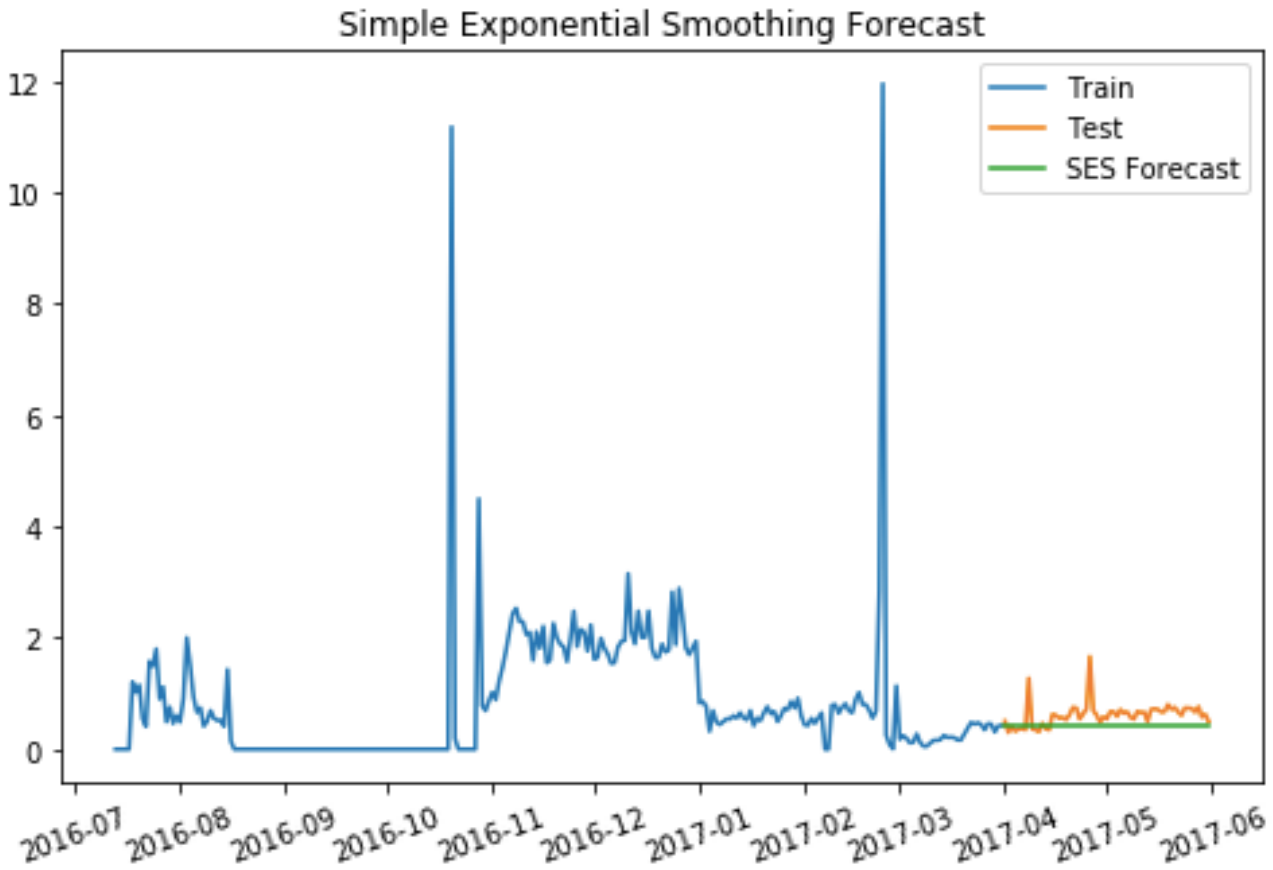
RMSE: 0.282545

Holt Winters Model

RMSE: 0.379134

SES Model

RMSE: 0.281969



Model Building

Naïve Model

RMSE: 0.281969

Moving Average Model

RMSE: 0.282545

Holt Winters Model

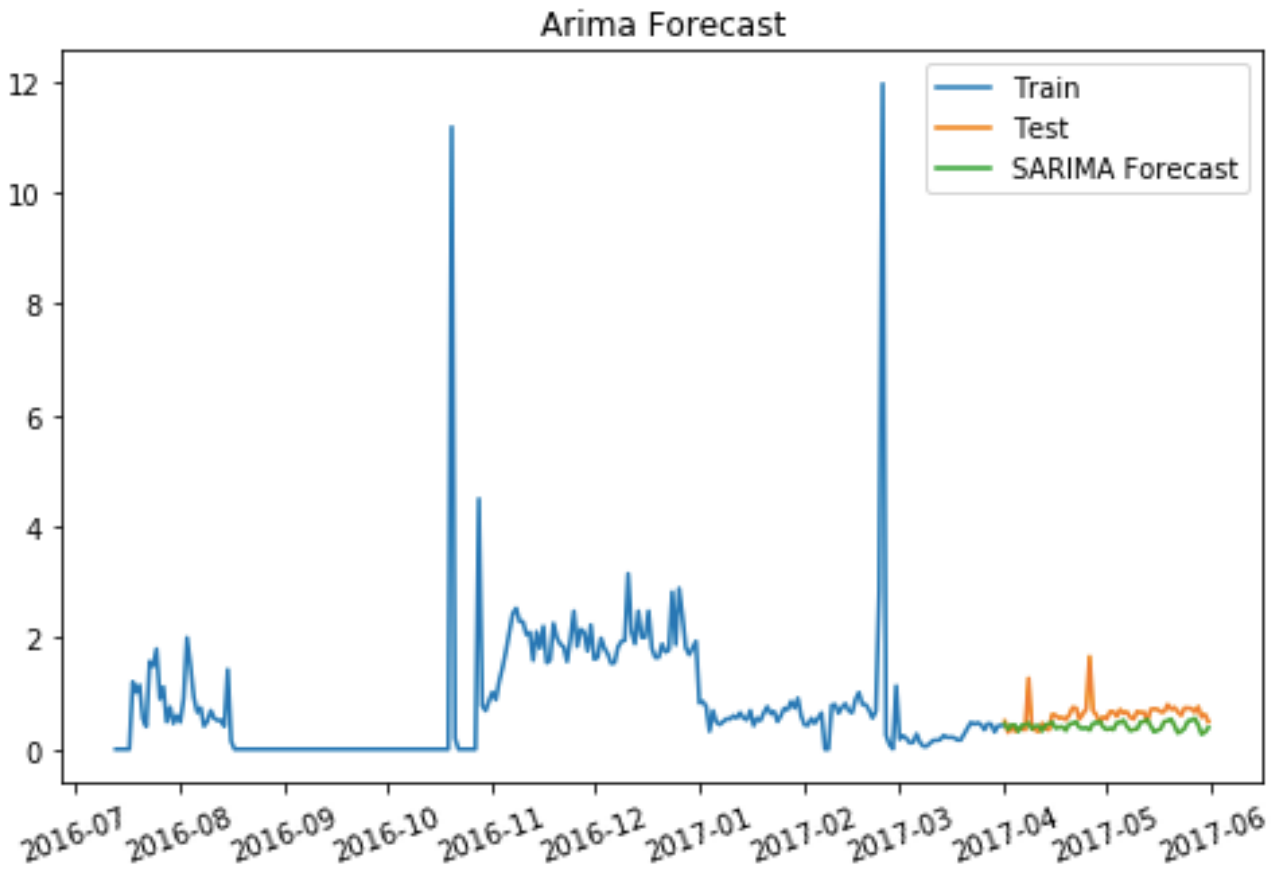
RMSE: 0.379134

SES Model

RMSE: 0.281969

Arima Model

RMSE: 0.294140



Structure of Detection System

- **Input**

Site Name, Date, Real CPM

- **Range of Irregularities**

CPM > 1.5*Prediction or CPM < 0.5*Prediction

- **Logistic Detail**

After getting the site name, the system will run those 5 models on previous training and testing data to get the best model with the lowest RMSE. Then it will rebuild the best model using all records before the-input-date as training data and perform prediction. Finally, it will compare the prediction with real CPM and give a conclusion.

Structure of Detection System

Example

- Input

CPM_detect("hoosierhuddle","2017-05-19",1.49)

- Output

The Best Model is Naive_model with 0.485 RMSE

Prediction is 0.842, real data is 1.49, ABNORMAL

The real CPM performance of site hoosierhuddle on 2017-5-19 is 1.49, while the prediction is about 0.84.

Based on forecasting result, the CPM is abnormal and over-performing on that day.

Extra Credit

Date	Site	Real Data	Prediction	Conclusion	Performance
5/15/2017	therepublikofmancunia	0.727	0.713	normal	overperforming
5/15/2017	snackmedia-claretandhugh	0.550	0.602	normal	underperforming
5/15/2017	presto-fswbucs	0.635	0.752	normal	underperforming
5/15/2017	cornellsun	1.807	0.871	ABNORMAL	overperforming
5/15/2017	gamurs-csgosquad	0.472	0.518	normal	underperforming
5/15/2017	overwatchtracker	0.335	0.360	normal	underperforming

I tested CPM performance for 6 sites on 2017-5-15. Cornellsun is the only site whose CPM is abnormal on 2017-5-15 but it is over-performing.

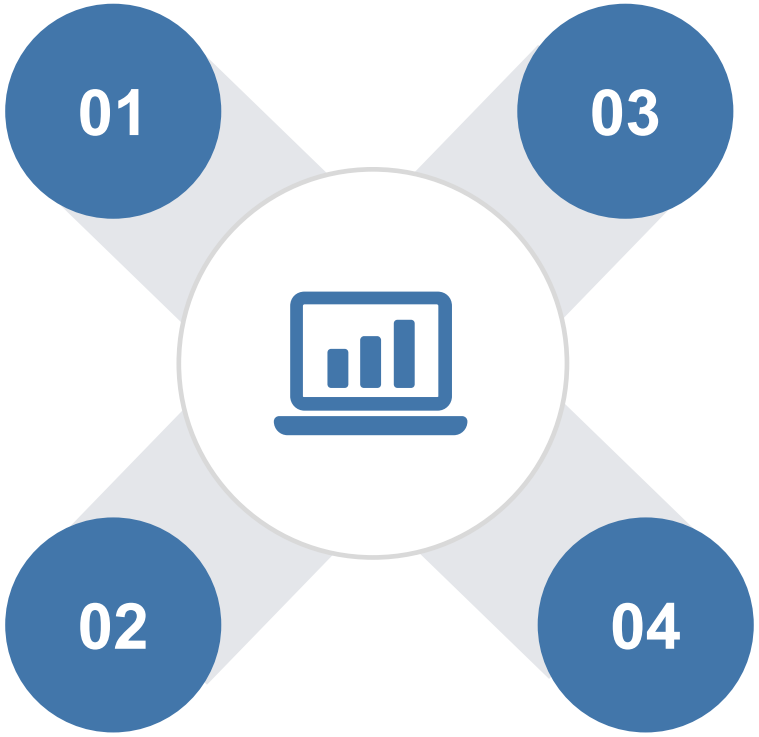
Next Steps

Model Customization

Build models for every site to fit their trend

More Attempts

Try more methods to find the best one

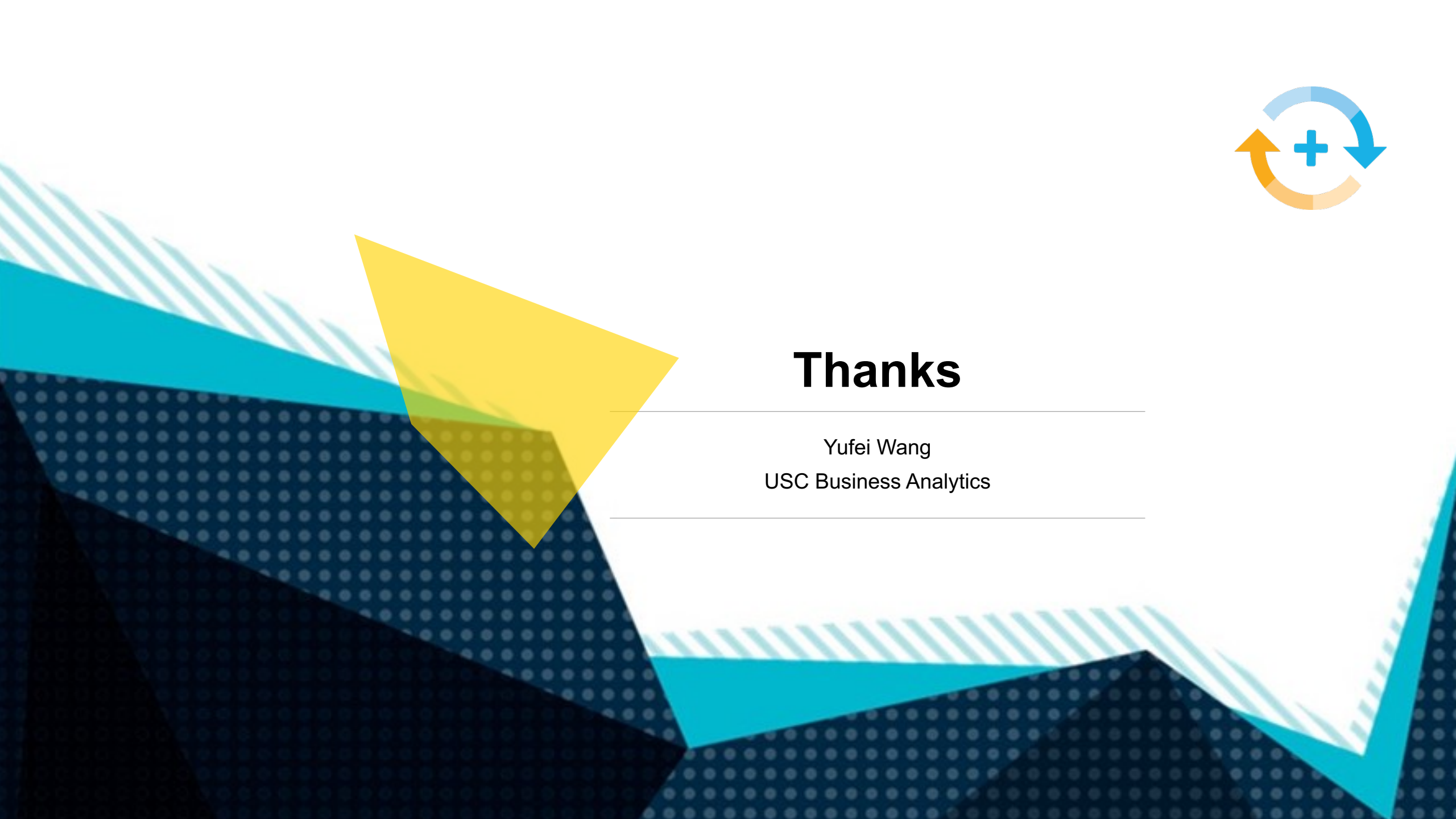


Data Integrity

The records between 2016-8-15 to 2016-10-15 are lost of many sites, data collection and data integrity is very important to build reliable models

Related Information

There are other factors should be included in prediction model like Ad content, ad price and the number of daily active user as they always have direct influence on CPM performance.



Thanks

Yufei Wang

USC Business Analytics