Web Traffic Forecast

Capstone Project for Springboard

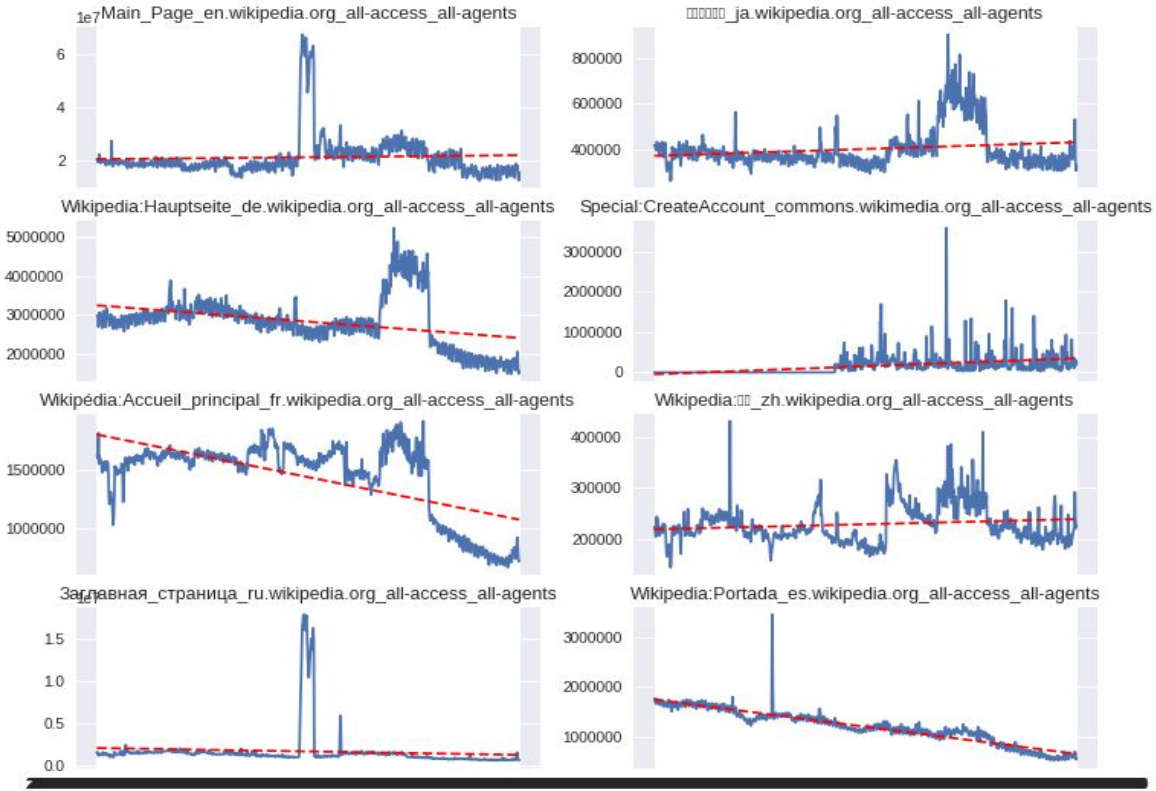
Xiao Zhang
12/15/2018

Background and Motivation

- Data: daily visit time series of approximately 145000 Wikipedia pages
- Goals of this project:
  a. explore the time series data and analyze trend, seasonality and stationarity
  b. get familiar with state-of-the-art time series forecast methods including ARIMA, Prophet and RNN

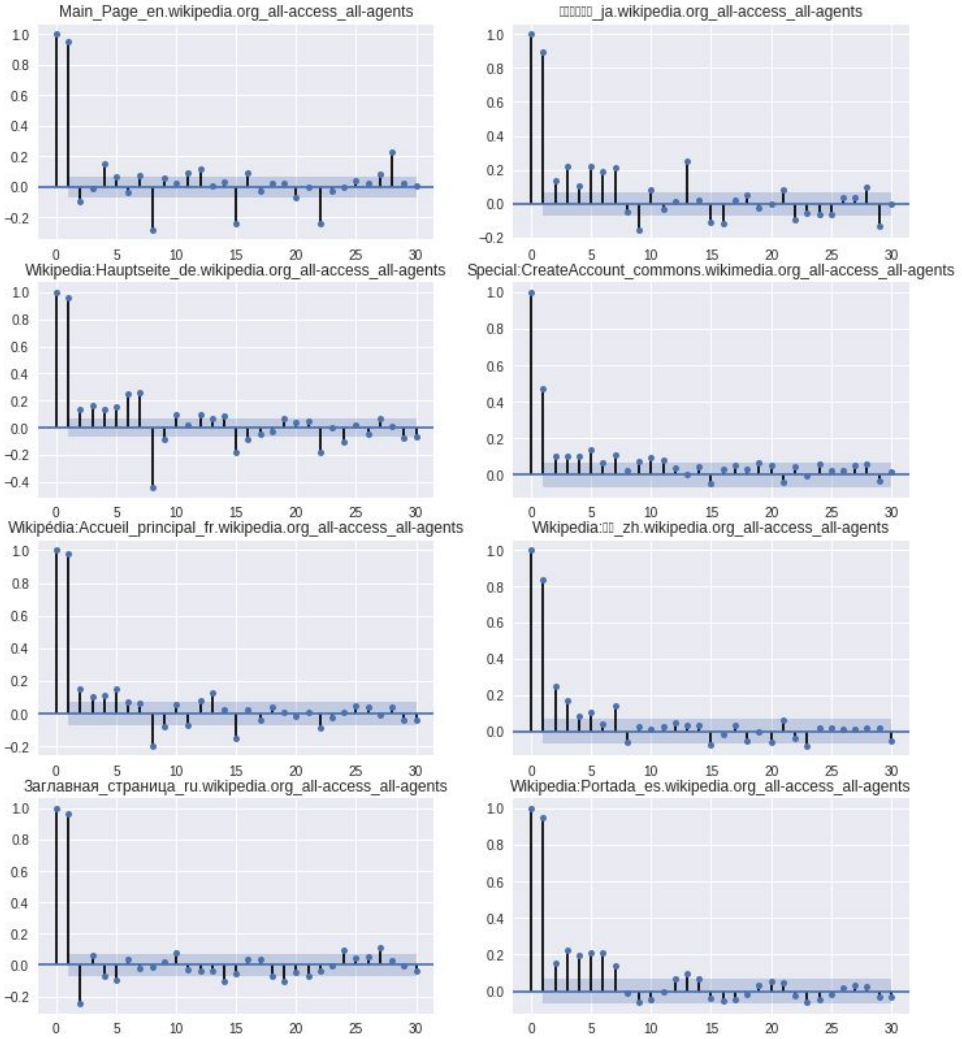
Data Exploration: are visits related to language?

Trend analysis



Seasonality analysis

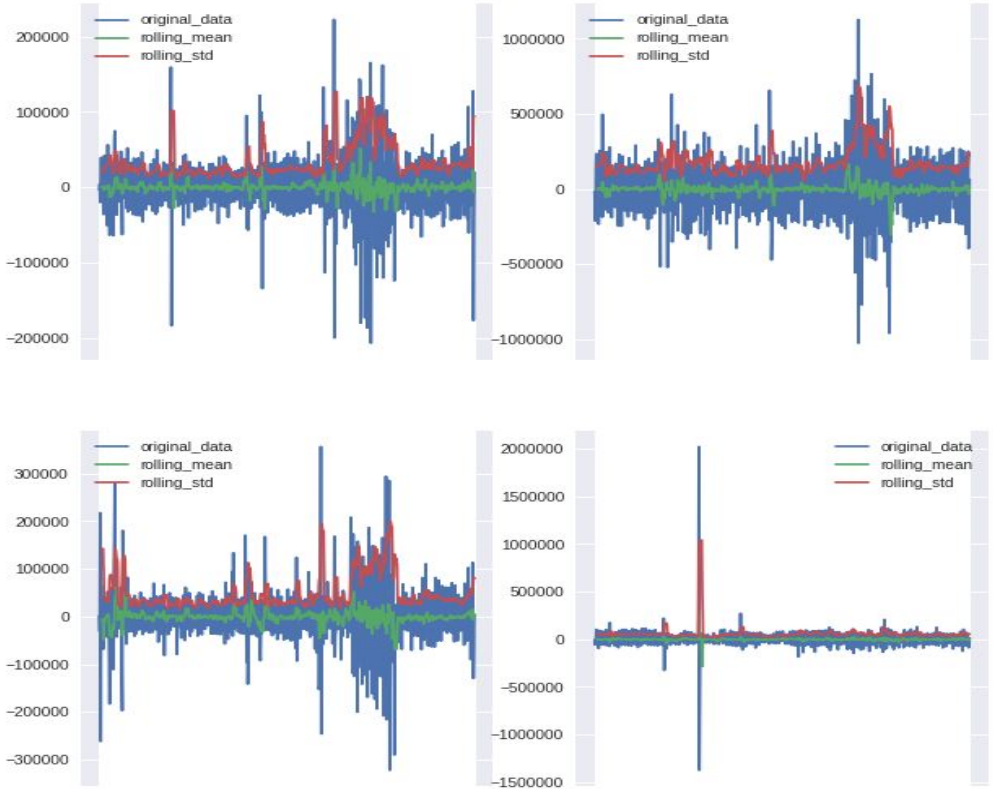
- PACF: partial autocorrelation function
- In general, visits are mostly related to visits of the previous day
- Can be used to determine the order of autoregressive model



Stationarity analysis

- Stationary: not trend or seasonal effects
- Augmented Dickey-Fuller test was used to test stationary. If p-value < 0.05, reject the null hypothesis and the time series is stationary
- Data transformation: first order difference

Data after differencing

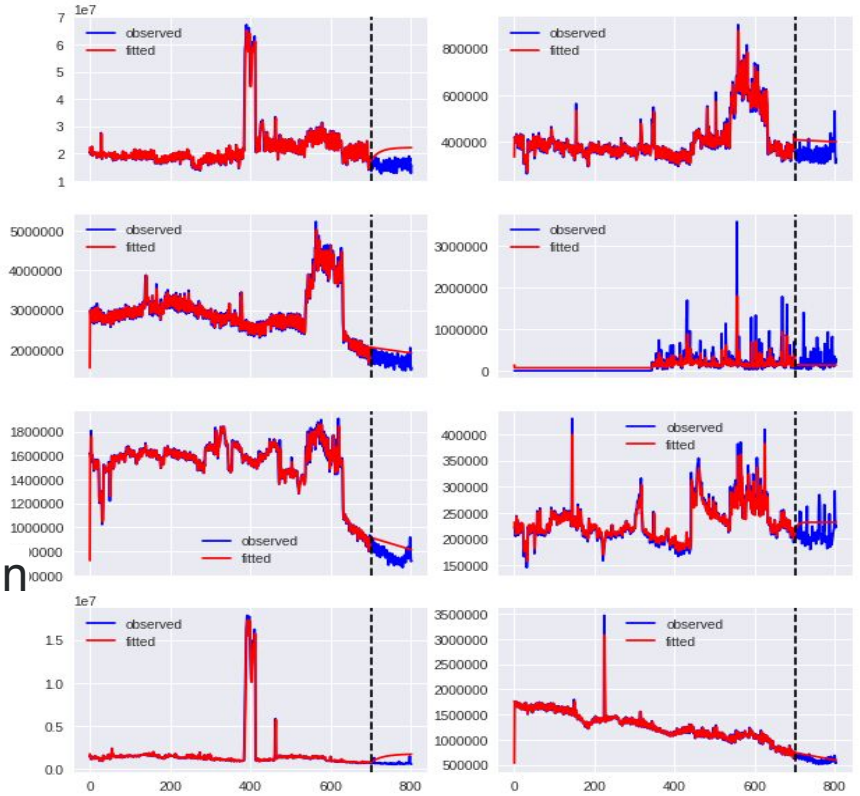


Forecast: ARIMA

- ARIMA = Auto-Regressive Integrated Moving Average
- Assume stationary
- Parameters:
  - Number of AR (Auto-Regressive) terms (p) -> chosen from PACF
  - Number of I (Integrated or Difference) terms (d) -> if differencing is needed
  - Number of MA (Moving Average) terms (q) -> chosen from ACF

ARIMA results

- Metrics: **SMAPE** (Symmetric mean absolute percentage error)
- Mean SMAPE value of the 8 forecasts is 29.0
- In general, ARIMA performs well for in-sample fitting
- Tends to converge to the sample mean for out-of-sample forecast when predicting long forecasting periods

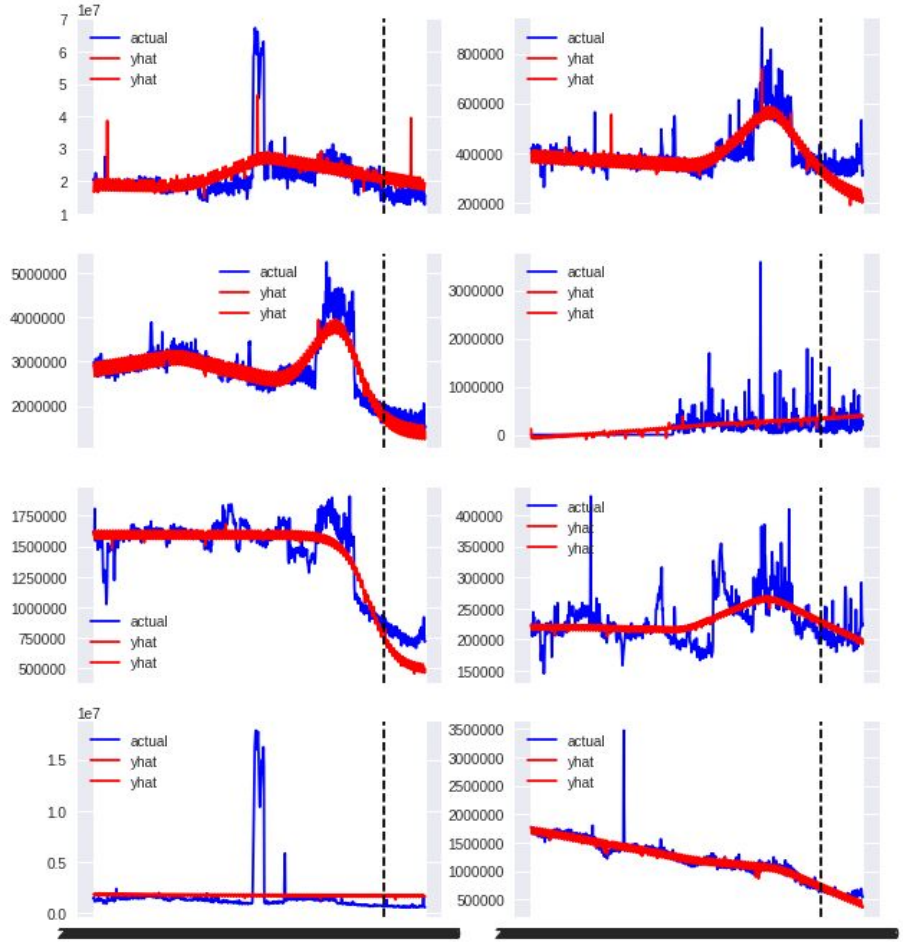


Forecast: Prophet

- The Prophet model decomposes time series to three main model components: trend, seasonality, and holidays
- Parameters:
  - Growth - linear, logistic
  - No of change points
  - Location of change points
  - Seasonality: daily, weekly, monthly, yearly, etc
  - Holidays

Prophet results

- The mean SMAPE of the predicted page visits over the test period is 34.6
- Prophet does not model the in-sample series as well as ARIMA since it is based on trend and seasonality
- The out-of-sample predictions are greatly affected by the trend near the end of the training period

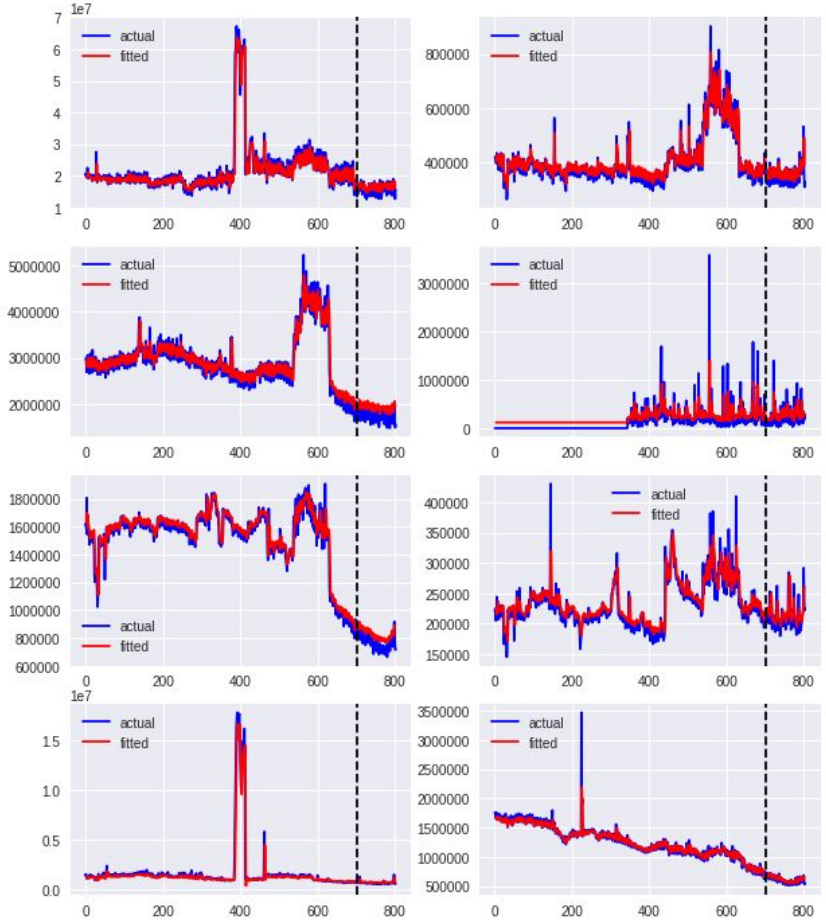


Forecast: LSTM

- **Long Short-Term Memory Network (LSTM)** is a a type of RNN (Recurrent neural network)
- LSTM networks have memory blocks that are connected through layers
- There are three types of gates within a unit:
  - Forget Gate: conditionally decides what information to throw away from the block.
  - Input Gate: conditionally decides which values from the input to update the memory state.
  - Output Gate: conditionally decides what to output based on input and the memory of the block.

LSTM results

- A LSTM network was built with a hidden layer with 4 LSTM blocks or neurons
- A window method is used so multiple recent time steps can be used to make the prediction for the next time step
- The mean SMAPE value of the 8 pages over the test period is 16.1
- LSTM predicts satisfactory values for both in-sample and out-of-sample forecast



Summary

	Advantages	Disadvantages	Results
ARIMA	easy to understand	parameter tuning (p, d, q), requirement of stationarity	Great fitting for in-sample prediction, out-of-sample forecast converge to sample mean, **SMAPE** of test period: 29
Prophet	Work for both stationary and non-stationary data, easy to implement	Parameter tuning (seasonality, change points, holiday etc)	Perform ok for in-sample forecast, out-of-sample forecast depend on trend at the end of training period, **SMAPE** of test period: 34.6
LSTM	Great performance in forecast	Difficult to understand, parameter tuning (no of cells, look-back steps)	Great performance for both in-sample and out-of-sample forecast, **SMAPE** of test period: 16.1