Title: Wikipedia Web traffic time series forecasting

Desire Aheza

Sreenivasa R. Gattamaneni

Pooja Walia

Harrisburg University of Science and Technology

Abstract

Time series forecasting is the use of a model to predict future values based on previously observed values. In this project we will be analyze, build a model that will be used to predicting the future behavior of time series’ that describe the web traffic for Wikipedia articles. Wikipedia is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation (Foundation, 2018). We will use a dataset that contains about 145k time series, the same dataset was used in a forecasting competition that was organized by google (Google, 2017). We plan to enhance our research by considering other datasets, using R and its popular forecasting libraries and technics.

# Introduction

This is an early exploratory data analysis for the Web traffic time series forecasting a web traffic with tidy R studio to analyze them in graphic use.

The challenges we faced is to predict the future behavior series of time which describes the web traffic for articles. The dataset we are using which contains the 100K plus time series and comes in two different separate files called as train.csv which holds the traffic data, In which each column is a date and each row is an article and key.csv contains a mapping between IDs and names which is used in submission of a file.

# Data Transformation and helper function

To make the data easier for training to handle the split in two different parts in which one

would be article information and another one is time series data from the data columns. We explain the article information into data form Wikimedia, mediawiki and Wikipedia due to the different formatting of the page names. After transformation we will rejoin all the article information into a column dataset.

In order to plot the series of data we generally use a helper function that allows us to

extract the series of time for a specific row number. Basically if has a normalized version which facilitates the comparison between multiple time series curve, to curve for large differences n view count.

# Summary parameter extraction

In the following steps we have a global look at the population parameters to our training time series data. Also we start with Wikipedia data set. The reason behind this approach was to probe the parameter space of the time series information along certain important metrics to observe and identify the break out forecasting strategies.

# Individual observations with extreme parameters

Based on the parameters we can focus on attention on those articles for which time series

parameters are at the extreme space.

We are doing a large linear slope standard deviations, large amplitudes, average views.

# Short term variability

Before turning to forecasting methods, we have a closer look at the characteristic short-term variability which becomes the evident of several plots already. We can see the high view count series on a very regular period which strikingly similar for both of them. We have kind of similar structure that the variance caused by lower number.

The plots generally describes the variability on a weekly scale and also we do have the average variability in the other plot over the day.

# Forecast methods for selected examples

We are identified sample of series with extreme parameters which we can use them to forecast different methods. For example of 100k articles we most likely to depends on the automatic mechanism to make our predictions. Therefore the forecasting method has to perform the robust range of different time series shapes and variabilities.

# Bibliography

Foundation, W. (2018). *Wikipedia*. Retrieved from Wikipedia: https://www.wikipedia.org/

Google. (2017). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/c/web-traffic-time-series-forecasting