Machine Learning(MO444)
Final Series Forecasting

Felipe Galvão
RA:116790
felipelemes@outlook.com

William Tustumi
RA:120281
whatust@gmail.com

Leo Yuuki Omori Omi
RA 138684
leoyuuki@gmail.com

João Pedro Ramos Lopes
RA 139546
email address

I. Introduction

This project will focus on the problem of Web Traffic Time Series Forecasting hosted on Kaggle [1]. The problem focuses on the problem of forecasting the future values of multiple time series. Sequential or temporal observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, to audio and video processing. The field of time series encapsulates many different problems, ranging from analysis and inference to classification and forecast. On our project we will forecast future web traffic for approximately Wikipedia pages, with a dataset provided by Kaggle [2].

II. Dataset Analysis

A. Dataset

The dataset provided is comprised of 803 days of visits measurements from 145063 combination of wikipedia page and access agent. An page name and header example can be seen on table I. The entries with NaN are correspondent to the days when the wikipedia page did not existed yet, thus those values where substituted by 0 in order to treat this edge case.

	2015 07/01	2015 07/02	2015 07/03	2015 07/04	2015 07/05
2NE1_zh.wikipedia.org_all-...	18	11	5	13	14
2PM_zh.wikipedia.org_all-a...	11	14	15	18	11
3C_zh.wikipedia.org_all-ac...	1	0	1	1	0
4minute_zh.wikipedia.org_a...	35	13	10	94	4
52_Hz_I_Love_You_zh.wik...	NaN	NaN	NaN	NaN	NaN
5566_zh.wikipedia.org_all-...	12	7	4	5	20
91Days_zh.wikipedia.org_al...	NaN	NaN	NaN	NaN	NaN
A'N'D_zh.wikipedia.org_all...	118	26	30	24	29

Table I
Data Example from Kaggle series forecast

The first step was performing an analysis of the data to understand which patterns could leverage our learning methods. Plotting graphs of the traffic from a few pages we observed that different pages show remarkably distinguish trends in traffic. This suggests that a good model needs to explicitly differentiate pages. For specific entries we note a strong year-to-year autocorrelation. We also observed weaker week-to-week and month-to-month trends. One clear example of this case was the Thanksgiving holiday page that showed an yearly spike around a Thursday on late November, which is the day of the holiday.

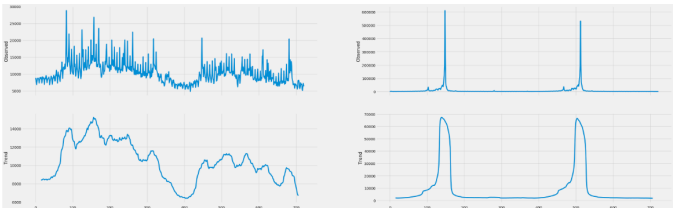

Figure 1. Observed Traffic and Traffic Trend from two different pages

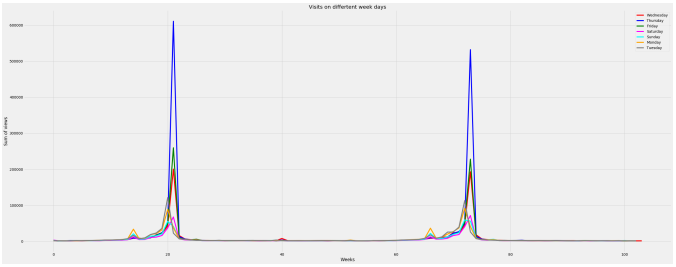

Figure 2. Observed Traffic separated by weekday for the thanksgiving page.

Additionally, we have to deal with spikes in traffic not following a regular pattern. We note a short-term dependency following those spikes, which means the days immediately before the prediction are important to consider.

III. Experiment Results & Discussion

IV. Conclusion & Future Work

References

[1] Kaggle. Web traffic time series forecasting. https://www.kaggle.com/c/web-traffic-time-series-forecasting.
[2] Kaggle. Web traffic time series forecasting data.