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

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 observesd that different pages show remarkably different trends in traffic. This suggests that a good model needs to explicitly distinguish pages. For specific pages 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 Wednesday on late November, around the date of the holiday.

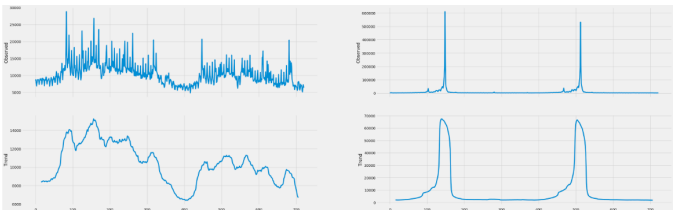

Figure 1. Observed Traffic and Traffic Trend from two different pages

Additionally, we have to deal with spikes intraffic 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.