Web Traffic Time Series Forecasting Project: Data Preprocessing Report

Domain Description

Our time series dataset was found on Kaggle, a website which does machine learning model creation competitions. In this case, our dataset came from a Google competition from 6 years ago (https://www.kaggle.com/competitions/web-traffic-time-series-forecasting). The goal of this competition was to predict the web traffic of specific pages on Wikipedia.

Our dataset itself contains two main attributes: page name and the amount of page views on a specific date. The page name gives the Wikipedia article name and extra metadata like the language of the page, and the access method of the page. The page views are given on a day-to-day basis as an integer. In the case of this dataset, each row represents a specific Wikipedia page, and each column represents a specific date.

The goal of our project will thus be to model the web traffic of the web pages in the future. The current attributes have a lot of possibilities for extra feature engineering, which gives a lot of room for improvement later on. Some examples of this are: extracting the language metadata from the page, aggregating a weekly count, monthly count, week-of-the-day count etc.

Data Preprocessing

As our dataset only consists of one type of value, our cleaning process is fairly straightforward. Our initial analysis of the data indicates that there is a high amount of missing values. In the case of the rows, this is 20.67%, in the case of the columns this is an even higher percentage of 99.88%.

A missing value either indicates that they were not able to get a measurement of the traffic on that specific day, or that the value is 0. This gives us three possible ways to deal with the missing values: fill them in with 0, fill them in with some kind of average or mean, or remove them. To not taint the integrity of the dataset, especially since each row represents a complete time series for an article, we choose to drop the rows with missing values.

In absolute numbers, this leads to the following:

- **Original Data Shape:** 145,063 rows, 804 columns.
- ***Cleaned Data Shape**:* 115,084 rows, 804 columns.