Web traffic time series forecasting

M. Nafees Abubacker, Osheen Mohan, Wang Wei Han, Veerendarath Naladala, Shahmir Kazi

***Forecast traffic to each individual webpage is pertinent for programmatic buying of advertisement space. FB Prophet yields robust results. Segmented targeted required for real world application.***

**Problematic**

The marketing industry has seen a multitude of upheavals due to the advent of machine learning techniques. Well known executives have made targeted campaigns synonymous with “Adwords”. However, the emergence of **programmatic buying** has dared to instill a new norm. **Mindshare (GroupM)** with their tool **Xaxis** are a prime example, used widely across EMEA and ASEAN region. However, it is a tool affordable by large firms and offered only on a retainer basis. It promises optimal targeting and maximum exposure, claiming 80% higher conversion.

This team has developed a solution for maximum exposure through forecasting web traffic to each **individual** webpage on a **daily basis**. This would allow the website to **maximize revenue** through **dynamic pricing** and allow the client to attain **maximum “eyeballs”** within a defined period. The ideation and proof of concept was validated through discussion with **Head of Communications at Arcelik A.S** and through the experience of our team, which has been **involved in buying advertisement space** for the **ICC Cricket World Cup** and Pakistan Super League.

**PROPOSED SOLUTION AND RESULTS**

The dataset includes 145,063 Wikipedia webpages with daily web traffic for 26 months. Such a scale ensures that the dataset is heterogenous. Preprocessing handled missing values and feature engineering allowed tags including medium (mobile, desktop or spider) and language. Hence the webpage shown as *“綜藝玩很大\_zh.wikipedia.org\_all-access\_all-agents”* was split into workable parts, each part represented in a column and highlighting a feature.

**Stationary tests** included Augmented Dickey-Fuller **(ADF) and KPSS**. The **ACF** and **PACF** plots are then used for visual assessment. **ARIMA** and **SARIMAX** are applied for forecasting accordingly. The team derives the optimal **(p,d,q)** values through error minimization tuning.

**ARIMA** was used to fit and estimate the DGP of each of the webpages. Accordingly, **SARIMAX** and finally **FB** **Prophet** were used. After due Hyper Parameter tuning through Cross Validation, FB Prophet is used to deliver results on the final test data. The results get even more precise when the dataset is sliced to make it homogenous, on features such as language. The accuracy for above mentioned page is **85.3%.**

**way forward for practical implementation**

The team concluded that this mode is **robust on a homogenous data set** and is fit to exercise as a tool to guide clients’ media purchasing, and also for vendors to maximize their profits. Further feature addition will yield greater insights into **consumer segmentation**.

These may be based on demographic, geographic, income or psychographic factors. It would allow vendors to have bargaining power when negotiating the placement and even cost-per-click rate for pages - especially during high seasonality, with Euro 2020 being prime example.