Data Science Path. Final Project.

Web traffic forecasting - Temporal series with ARIMA

Created by Sandy Chiereghin

Introduction

For this project, I selected to present a work using the data from a digital product's web page "DisVoize".

DisVoize is a mobile application that allows real-time audio transmission during language interpretations, guided tours and onsite and online events. What I will analyze isn't the performance of the mobile application. I decided to analyse the performance and the web traffic of the commercial product website: disvoize.com.

For that, was analysing the data from the disvoize.com Google Analytics tool connected to the webpage disvoize.com.

In order to handle the data coming from the website, for build the Panda Dataframe I used the Google Analytics Reporting API v4.

The Google Analytics Reporting API v4 provides programmatic methods to access report data in Google Analytics and handle them using several programming languages, Python included.

Basically, what I did using the Google Analytics Reporting API, was build custom dashboards to display Google Analytics data and put them in a Pandas Dataframe.

The API's key features:

- Metric expressions
  The API allows users to request built-in metrics and combinations of metrics expressed in mathematical operations. For example, you can use the expression ga:goal1completions/ga: sessions to request the goal completions per number of sessions.
- Multiple date ranges
  The API allows you to select the date ranges in a single request.
- Cohorts and Lifetime value
  The API has a rich vocabulary to request Cohort and Lifetime value reports.
- Multiple segments
  The API enables you to get multiple segments in a single request.

Here is a Python quick start guide: https://developers.google.com/analytics/devguides/reporting/core/v4/quickstart/service-py

I tried to take advantage of all the power of this API. I prepared different datasets and customised them in order to perform the different steps of my study exercise. Starting from an exploratory analysis, passing to the Machine Learning exercise using temporal series algorithms to forecast web traffic related to a determinate period of time.

Since I have direct ownership of the database, using the Google Analytics Reporting API I can set up different Pandas data frames depending on the proposal of it. For this reason, I will present different types of datasets related to the different proposals. There will be a dataset performing exploratory analysis and visualizations and finally, I will present the dataset in which I will perform web traffic forecasting with the relevant data and information.

What kind of metrics we can access with Google Analytics Reporting API:

Using this https://ga-dev-tools.web.app/dimensions-metrics-explorer platform we can discover the possible data we can collect and use and how to combine them as Dimensions and Metrics.

Common parameter to all data frames:

The collected data is related to the following period of time: a: '2019-01-01', 'endDate': '2022-11-10'

Main data frame (for forecasting) information:

ga:date	ga:users	ga:newUsers	ga:sessions	ga:percentNewSessions	ga:avgSessionDuration	ga:timeOnPage	ga:avgTimeOnPage	ga:exitRate	ga:avgSessionDura
20200402	0.0	8.0	13.0	61.538462	335.230769	4357.0	544.625000	61.904762	335.230
20200403	0.0	3.0	5.0	60.000000	0.000000	0.0	0.000000	100.000000	0.000
20200406	0.0	1.0	1.0	100.000000	0.000000	0.0	0.000000	100.000000	0.000
20200407	0.0	1.0	2.0	50.000000	4.500000	9.0	9.000000	66.666667	4.500
20200408	0.0	8.0	9.0	88.888889	17.888889	161.0	40.250000	69.230769	17.888
...	...	...	...	...	...	...	...	...	
20221116	0.0	1.0	1.0	100.000000	0.000000	0.0	0.000000	50.000000	0.000
20221117	0.0	13.0	23.0	56.521739	247.565217	5694.0	99.894737	28.750000	247.565
20221118	0.0	0.0	3.0	0.000000	12.666667	38.0	5.428571	30.000000	12.666
20221119	0.0	4.0	4.0	100.000000	103.000000	412.0	29.428571	22.222222	103.000
20221120	4.0	4.0	5.0	80.000000	0.000000	0.0	0.000000	50.000000	0.000

Number of Instances: 937

Number of Attributes: 10

Attributes Information:

date: (object) The date of the session is formatted as YYYYMMDD.

users: (float64) The total number of users for the requested time period.

newUsers: (float64) The number of sessions marked as a user's first sessions.

sessions: (float64) The total number of sessions.

percentNewSessions: (float64) The percentage of sessions by users who had never visited the property before.

avgSessionDuration: (float64) The average duration (in seconds) of users' sessions.

timeOnPage: (float64) Time (in seconds) users spent on a particular page, calculated by subtracting the initial view time for a particular page from the initial view time for a subsequent page. This metric does not apply to exit pages of the property.

avgTimeOnPage: (float64) The average time users spent viewing this page or a set of pages.

exitRate: (float64) The percentage of exits from the property that occurred out of the total pageviews.

avgSessionDuration: (float64) The average duration (in seconds) of users' sessions.

dtype: object

Other parameters:

totalEvents: (float64) The total number of events for the profile, across all categories.

bounceRate: (float64) The percentage of single-page session (i.e., session in which the person left the property from the first page).

Country: (object) Users' country, derived from their IP addresses or Geographical IDs.

Region: (object) Users' region, derived from their IP addresses or Geographical IDs. In U.S., a region is a state, New York, for example.

City: (object) Users' city, derived from their IP addresses or Geographical IDs.

Objectives:

General:

- Using the historical metrics of Disvoize.com I want to create web traffic forecasting as a strategic study.

Specific:

- Learn how to explode data from Google Analytics using Python.
- Learn about temporal series machine learning algorithms and how to apply them to digital product strategies.
- Make an exploratory analysis related to the selected data and project.
- Build a general study case related to the traffic performance of disvoize.com.