Web Traffic Time Series Forecasting

Forecast future traffic to Wikipedia pages

Web Traffic Time Series Forecasting

Overview

Start

Jul 13, 2017

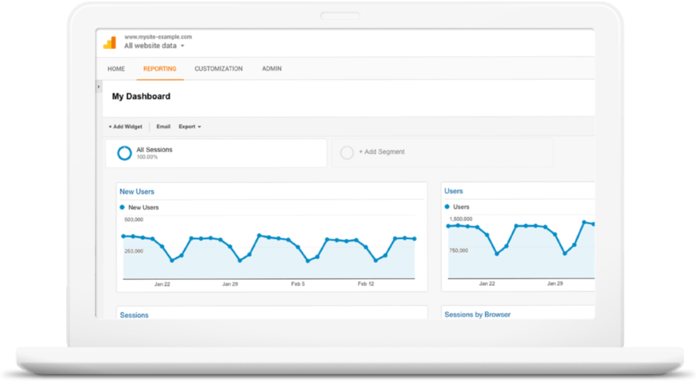
Close

Nov 15, 2017

Merger & Entry

Description

This competition focuses on the problem of forecasting the future values of multiple time series, as it has always been one of the most challenging problems in the field. More specifically, we aim the competition at testing state-of-the-art methods designed by the participants, on the problem of forecasting future web traffic for approximately 145,000 Wikipedia articles.



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. What can you do to help predict future views?

This competition will run as two stages and involves prediction of actual future events. There will be a training stage during which the leaderboard is based on historical data, followed by a stage where participants are scored on real future events.

You have complete freedom in how to produce your forecasts: e.g. use of univariate vs multi-variate models, use of metadata (article identifier), hierarchical time series modeling (for different types of traffic), data augmentation (e.g. using Google Trends data to extend the dataset), anomaly and outlier detection and cleaning, different strategies for missing value imputation, and many more types of approaches.

We thank Google Inc. and Voleon for sponsorship of this competition, and Oren Anava and Vitaly Kuznetsov for organizing it.

Kaggle is excited to partner with research groups to push forward the frontier of machine learning. Research competitions make use of Kaggle's platform and experience, but are largely organized by the research group's data science team. Any questions or concerns regarding the competition data, quality, or topic will be addressed by them.

Evaluation

Submissions are evaluated on SMAPE between forecasts and actual values. We define SMAPE = 0 when the actual and predicted values are both 0.

Submission File

For each article and day combination (see key.csv), you must predict the web traffic. The file should contain a header and have the following format:

```
Id,Visits
bf4edcf969af,0
929ed2bf52b9,0
ff29d0f51d5c0
etc.
```

Due to the large file size and number of rows, submissions may take a few minutes to score. Thank you for your patience.

Prizes

1. $12,000
2. $8,000
3. $5,000

Top submissions will also have the opportunity to present their work at the **NIPS Time Series Workshop** in Long Beach, California, co-located with the top machine learning conference **NIPS 2017**. Attending the workshop is not required to participate in the competition, however only teams that are attending the workshop will be considered to present their work.

Attendees presenting in person are responsible for all costs associated with travel, expenses, and fees to attend NIPS 2017.

Timeline

This competition has a training phase and a future forecasting phase. During the training phase, participants build models and predict on historical values. During the future phase, participants will forecast future traffic values.

- **September 1st, 2017** - Deadline to accept competition rules.
- **September 1st, 2017** - Team Merger deadline. This is the last day participants may join or merge teams.
- **September 1st, 2017** - Final dataset is released.
- **September 12th 7:59 PM UTC** - Final submission deadline.

Competition winners will be revealed after November 13, 2017.

All deadlines are at 11:59 PM UTC on the corresponding day unless otherwise noted. The competition organizers reserve the right to update the contest timeline if they deem it necessary.

Citation

Maggie, Oren Anava, Vitaly Kuznetsov, and Will Cukierski. Web Traffic Time Series Forecasting. https://kaggle.com/competitions/web-traffic-time-series-forecasting, 2017. Kaggle.

Dataset Description

The training dataset consists of approximately 145k time series. Each of these time series represent a number of daily views of a different Wikipedia article, starting from July, 1st, 2015 up until December 31st, 2016. The leaderboard during the training stage is based on traffic from January, 1st, 2017 up until March 1st, 2017.

The second stage will use training data up until September 1st, 2017. The final ranking of the competition will be based on predictions of daily views between September 13th, 2017 and November 13th, 2017 for each article in the dataset. You will submit your forecasts for these dates by September 12th.

For each time series, you are provided the name of the article as well as the type of traffic that this time series represent (all, mobile, desktop, spider). You may use this metadata and any other publicly available data to make predictions. Unfortunately, the data source for this dataset does not distinguish between traffic values of zero and missing values. A missing value may mean the traffic was zero or that the data is not available for that day.

To reduce the submission file size, each page and date combination has been given a shorter Id. The mapping between page names and the submission Id is given in the key files.

File descriptions

Files used for the first stage will end in '_1'. Files used for the second stage will end in '_2'. Both will have identical formats. The complete training data for the second stage will be made available prior to the second stage.

- **train_*.csv** - contains traffic data. This a csv file where each row corresponds to a particular article and each column correspond to a particular date. Some entries are missing data. The page names contain the Wikipedia project (e.g. en.wikipedia.org), type of access (e.g. desktop) and type of agent (e.g. spider). In other words, each article name has the following format: 'name_project_access_agent' (e.g. 'AKB48_zh.wikipedia.org_all-access_spider').
- **key_*.csv** - gives the mapping between the page names and the shortened Id column used for prediction
- **sample_submission_*.csv** - a submission file showing the correct format

Files

6 files

Size

611.85 MB

Type

zip

License

Subject to Competition Rules

key_1.csv.zip(100.61 MB)

Unable to show preview

Previews for binary data are not supported

Data Explorer

611.85 MB

key_1.csv.zip

key_2.csv.zip

sample_submission_1.csv.zip

sample_submission_2.csv.zip

train_1.csv.zip

train_2.csv.zip