Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
June Call, A.Y. 2022/2023

Last update: June 1, 2023

1 Project dates

Start date: May 31, 2022 at 23:59 (CET)
Due date: June 17, 2022 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Predicting online news popularity is vital, with far-reaching implications for various fields of study, including media, advertising, and communication. As the consumption and dissemination of news continue to shift to digital platforms, understanding the factors that contribute to the sharing of news articles on the internet is critical. Accurate predictions can provide valuable insights into the underlying mechanisms of online information diffusion and help inform strategies for news organizations and advertisers seeking to maximize their reach. In this project, you are required to predict the number of shares of each described news article. Moreover, this task presents an opportunity to contribute to the broader discourse on the role of social media and online communication in shaping public opinion and the contemporary media landscape. Thus, the prediction of online news popularity is a significant area of research with considerable theoretical and practical implications.

2.1 Dataset

> **Warning:** For this project, you are not allowed to use external datasets other than the one provided. Adoption of external resources, including pre-trained models, will result in failure of the exam.

The dataset contains 39797 instances describing some online news articles. The dataset does not share the original content but some associated statistics. The original content can be publicly accessed and retrieved using the provided URLs. Several metadata characterizes each record. The following is a short description of each of them:

- url: URL of the article (non-predictive)
- timedelta: Days between the article publication and the dataset acquisition.
- n_tokens_title: Number of words in the title.
- n_tokens_content: Number of words in the content.
- n_unique_tokens: Rate of unique words in the content.

- n_non_stop_words: Rate of non-stop words in the content.
- n_non_stop_unique_tokens: Rate of unique non-stop words in the content.
- num_hrefs: Number of links.
- num_self_hrefs: Number of links to other articles published by Mashable.
- num_imgs: Number of images.
- num_videos: Number of videos.
- average_token_length: Average length of the words in the content.
- num_keywords: Number of keywords in the metadata.
- data_channel: the type of data channel. It can be ’Lifestyle’, ’Entertainment’, ’Business’, ’Social Media’, ’Tech’ or ’World’.
- kw_min_min: Worst keyword (min. shares).
- kw_max_min: Worst keyword (max. shares).
- kw_avg_min: Worst keyword (avg. shares).
- kw_min_max: Best keyword (min. shares).
- kw_max_max: Best keyword (max. shares).
- kw_avg_max: Best keyword (avg. shares).
- kw_min_avg: Avg. keyword (min. shares).
- kw_max_avg: Avg. keyword (max. shares).
- kw_avg_avg: Avg. keyword (avg. shares).
- self_reference_min_shares: Min. shares of referenced articles in Mashable
- self_reference_max_shares: Max. shares of referenced articles in Mashable
- self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
- weekday: the day of the week the article was published.
- LDA_00: Closeness to LDA topic 0.
- LDA_01: Closeness to LDA topic 1.
- LDA_02: Closeness to LDA topic 2.
- LDA_03: Closeness to LDA topic 3.
- LDA_04: Closeness to LDA topic 4.
- global_subjectivity: Text subjectivity.
- global_sentiment_polarity: Text sentiment polarity.
- global_rate_positive_words: Rate of positive words in the content.
- global_rate_negative_words: Rate of negative words in the content.
- rate_positive_words: Rate of positive words among non-neutral tokens.
- rate_negative_words: Rate of negative words among non-neutral tokens.
- avg_positive_polarity: Avg. polarity of positive words.

- min_positive_polarity: Min. polarity of positive words.
- max_positive_polarity: Max. polarity of positive words.
- avg_negative_polarity: Avg. polarity of negative words.
- min_negative_polarity: Min. polarity of negative words.
- max_negative_polarity: Max. polarity of negative words.
- title_subjectivity: Title subjectivity.
- title_sentiment_polarity: Title polarity.
- abs_title_subjectivity: Absolute subjectivity level.
- abs_title_sentiment_polarity: Absolute polarity level.
- shares: Number of shares (label).

The dataset is located at:
`https://drive.google.com/file/d/1QTyU4vNW3WrMIkIQwUtl7cEmm0nVi-IL/view?usp=sharing`

Within the archive, you will find the following elements:

- **development.csv** (development set): a comma-separated values file containing the records from the development set. This portion does have the `shares` column, which you should use to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the records corresponding to the evaluation set. This portion does not have the `shares` column.
- **sample_submission.csv**: a sample submission file.

2.2 Task

You are required to build a regression pipeline to predict the number of shares of each described news in the Evaluation Set.

2.3 Evaluation metric

Your submissions will be evaluated through RMSE (Root-Mean-Square Error).

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
31715,4349.984540028923
31716,5774.650007043754
31717,5281.4213237379245
31718,2264.97388674423
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set. It corresponds to the column `id` in the evaluation CSV file.
- the `Predicted` label for the corresponding record.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `lorenzo.vaiani@polito.it`. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_june_2023**.

> ℹ **Info:** A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing `.zip` extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.