# ML Regression Assignment

## Introduction

Yojo.com is one of the main actors in the online shopping industry. One of the key aspects of its success is the continuous monitoring of its client satisfaction.

In a previous project, the company developed a survey analysis among its clients in order to predict their satisfaction level. From this project, data analysts in the company found out that an important aspect of customers satisfaction is driven by product reviews, specifically by popular reviews (defined as reviews collecting a large number of likes). From these findings, the company decided to study the characteristics that made a review to be popular, in order to design an early intervention in case of potential popular negative feedbacks.

The analysis of the reviews will be made using textual features of the reviews, for example, the number of non-stop words[1], the subject, and the sentiment associated with the title and content.

In this assignment, you will use Machine Learning techniques to predict the number of likes that a review will obtain based on a set of textual features.

## Dataset description

You will receive two datasets containing a list of reviews together with the textual information of their title and content.
There is a total of 38000 records and 38 explanatory features divided into two datasets.

• model.csv: the dataset contains the information of 28000 reviews with the respective target variable. You must use this data to create and evaluate your model.

• predictions.csv: the dataset contains the information of 10000 reviews **without** the target variable. You are requested to provide the predictions for this set of records.

## Target variable:

The target attribute likes is an integer and corresponds to the number of likes collected by the review.

The task is formulated as a regression task. Your grade will be based on both the **MAE** metric, and the modeling process presented in the **report**.

[1] Stop-words are common words in a language (like articles, prepositions, pronouns, conjunctions, etc) that does not add much information to the text.

## Attribute information:

| n | Variable | Description |
|---|---|---|
| 1 | age_days | Days between the article publication and |
| 2 | n_tokens_title | Number of words in the title |
| 3 | n_tokens_review | Number of words in the content |
| 4 | n_unique_tokens | Rate of unique words in the content |
| 5 | n_non_stop_words | Rate of non-stop words in the content |
| 6 | n_non_stop_unique_tokens | Rate of unique non-stop words in the |
| 7 | num_hrefs | Number of links in the content |
| 8 | num_self_hrefs | Number of links to other products |
| 9 | num_imgs | Number of images |
| 10 | num_videos | Number of videos |
| 11 | average_token_length | Average length of the words |
| 12 | num_keywords | Number of keywords in the metadata |
| 13 | product_category | category of the product: business, cleaning,...,other |
| 14 | self_reference_min_shares | Minimum likes of referenced articles |
| 15 | self_reference_max_shares | Maximum likes of referenced articles |
| 16 | self_reference_avg_sharess | Average likes of referenced articles |
| 17 | day | Publication day: mon .. sun |
| 18 | topic_quality | Percentage of the content speaking about quality |
| 19 | topic_shipping | Percentage of the content speaking about shipping |
| 20 | topic_packaging | Percentage of the content speaking about packaging |
| 21 | topic_description | Percentage of the content speaking about the description |
| 22 | topic_others | Percentage of the content speaking about other topics |
| 23 | global_subjectivity | Content text subjectivity (0-Objective 1-Subjective) |
| 24 | global_sentiment_polarity | Text sentiment polarity (-1-Negative 1-Positive) |
| 25 | global_rate_positive_words | Rate of positive words in the content |
| 26 | global_rate_negative_words | Rate of negative words in the content |
| 27 | rate_positive_words | Rate of positive words among non-neutral |
| 28 | rate_negative_words | Rate of negative words among non-neutral |
| 29 | avg_positive_polarity | Avg. polarity of positive words |
| 30 | min_positive_polarity | Min. polarity of positive words |
| 31 | max_positive_polarity | Max. polarity of positive words |
| 32 | avg_negative_polarity | Avg. polarity of negative words |
| 33 | min_negative_polarity | Min. polarity of negative words |
| 34 | max_negative_polarity | ax. polarity of negative words |
| 35 | title_subjectivity | Title subjectivity |
| 36 | title_sentiment_polarity | Title polarity |
| 37 | abs_title_subjectivity | Absolute subjectivity level |
| 38 | abs_title_sentiment_polarity | Absolute polarity level |
| 39 | **likes** | **Number of likes (target)** |

## Important dates and submission instructions

**1. Model Training Data Release: 03 December 2021, 19:00.**

**2. Training set analysis and and model identification: 10 December 20:00.**
You are kindly asked to submit the following supporting information in the WeBeep page of the course:

a) A **brief report** of the step-by-step methodology (i.e., pre-processing, visualization, training, testing, etc.) that you have followed to develop your model, this document must illustrate the motivation behind your selected approach.

• File Format: .pdf

• Filename: 6-digit student code (e.g., 123456.pdf)

b) **The commented python code** that you used in your model. Please submit only your final model, all preliminary attempts can be described in the report. Comments in the code must ensure that the code is easy to follow.
• File Format: .ipynb, .py
• Filename: 6-digit student code, e.g., 123456.ipynb or 123456.py (note that this is **not** your 8-digit POLIMI personal code)

**3. Prediction Data Release: 10 December 21:00.**

**4. Prediction Submission: 12 December 20:00.**
You are kindly requested to strictly follow the described submission guidelines:
• File Format: .csv
• Filename: 6-digit student code (e.g., 123456.csv)
• Column Format: **A single** column named "target"
• Row Format: Your predictions with **the same number of rows** and in the same order as the **prediction** test set.

Example:

| Target |
|---|
| 12 |
| 14 |
| 55 |
| 22 |
| 43 |
| 110 |

## Further instructions

- **Verify** the integrity and coding of your uploaded files in platform.
- The assignment can be developed in groups with a maximum number of three participants.
- Nevertheless, **submission is individual**, therefore each student must upload his/her own submission files (even if they are the same for all participants in the group).
- **Verify** the integrity and coding of your uploaded files in platform.

- **Any submission that does not respect the guidelines (submission after deadline, empty file, wrong student code) will not be graded**.