ML Regression Assignment – Report

The analysis is aimed at developing a model for the company *Yojo.com* in order to predict the number of likes that a review will obtain and design an early intervention in case of potential popular negative feedbacks. The metrics considered to validate it is the MAE, but before going through the model creation, a time-consuming preliminary process of inspection and preparation of the provided dataset is required.

Data analysis and data preparation

The whole dataset contains 28000 observations and 38 different explanatory variables of which 2 categorical ones ('product_category' and 'day'). A check to detect any missing or duplicate values was put in place and it turned out that there were none.

At this point a scatterplot considering all the variables against the target variable 'likes' was built to get a first qualitative insights about the distribution of the variables.

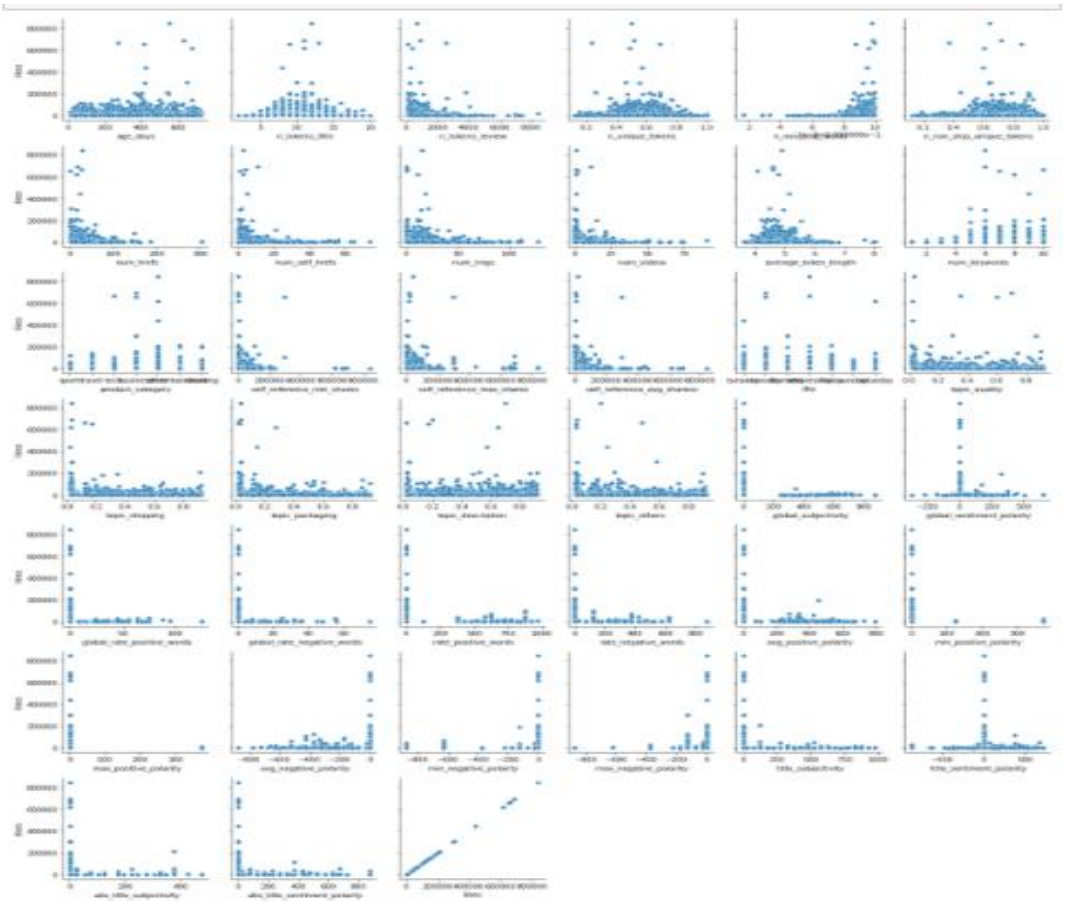

Figure 1: Scatterplot of all the variables against the 'likes' variable

By analysing the target variable 'likes' it was evident the presence of high extreme values and it was decided to apply a box-cox transformation to not directly remove the mentioned extreme values because, even if they are few, they are relevant for the computation of the MAE metrics of the developed models.

Another consideration that emerged by looking at the scatterplot was that some variables (' global_subjectivity', ' global_sentiment_polarity', ' global_rate_positive_words', ' global_rate_negative_words', ' rate_positive_words', ' rate_negative_words', ' avg_positive_polarity', ' min_positive_polarity', ' max_positive_polarity', ' avg_negative_polarity', ' min_negative_polarity', ' max_negative_polarity', ' title_subjectivity', ' title_sentiment_polarity', ' abs_title_subjectivity' and ' abs_title_sentiment_polarity') presented observations that were strongly out of scale, since by definition they should have fallen into specific intervals like [-1,1], [-1,0] or [0,1]. Focusing just on these observations it was assumed that they were mistakes, and they were divided by 1000 to make them comparable to the other observations (ex 800 become 0.8 and it is a plausible value compared to the other ones).

After the scaling of these variables an exhaustive process of outliers' elimination was performed.

This task was accomplished by following this line of reasoning:

- For those variables for which a theoretical distribution was very similar to the actual one by looking at the overall closeness of the points to the line in the q-q plot, the observations more distant from the line were dropped. For example, in the image below the distribution of the variable "global rate of negative words" is compared with an exponential one and it is evident that the observations higher than 0.125 are outliers.

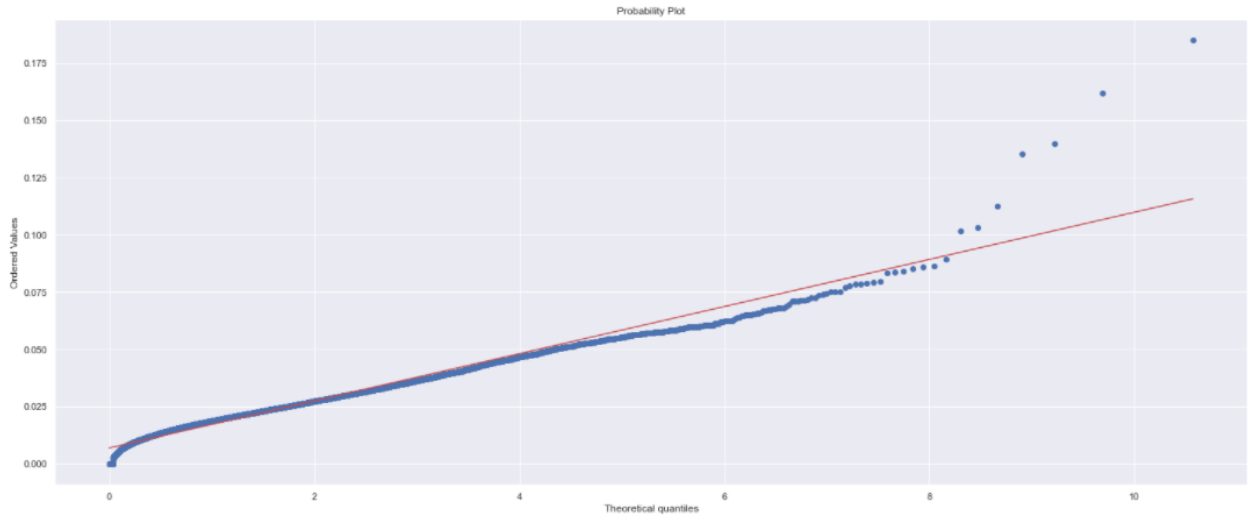

Figure 2: q-q plot considering the 'global rate of negative words' variable against an exponential distribution

- For the other ones, since it was not found a theoretical distribution overall similar to the actual one, a more naïve approach was adopted, that consisted in dropping just very few observations that

appeared very different (distant) from the vast majority by looking at their histograms and at the scatterplot.

- By looking at the dispersion of the points in the scatterplot below (Figure 3), where the target variable is the transformed likes, it was decided to remove those observations associated to a value of the transformed likes lower than 2.5.

At the end of this process 1064 observations were removed.

At this point the dataset was split into two sub-datasets: one containing just the categorical variables and the other one containing just the numerical ones.

For what concerns the numerical dataset, a variable reduction process was performed. The two variables ' abs_title_subjectivity' and ' abs_title_sentiment_polarity' have been dropped since they did not add more information with respect to the variables ‘ title_subjectivity' and ' title_sentiment_polarity'. Moreover, the correlation matrix has allowed to identify the pairs of variables that were highly correlated between each other (a threshold of 0.8 has been considered): for each pair, the variable less correlated with respect to the target (‘ rate_negative_words', ' self_reference_max_shares', ' self_reference_min_shares', ' n_unique_tokens') has been removed. Then, the numerical variables have been standardised to avoid biases in the algorithms caused by the different magnitude of the values.

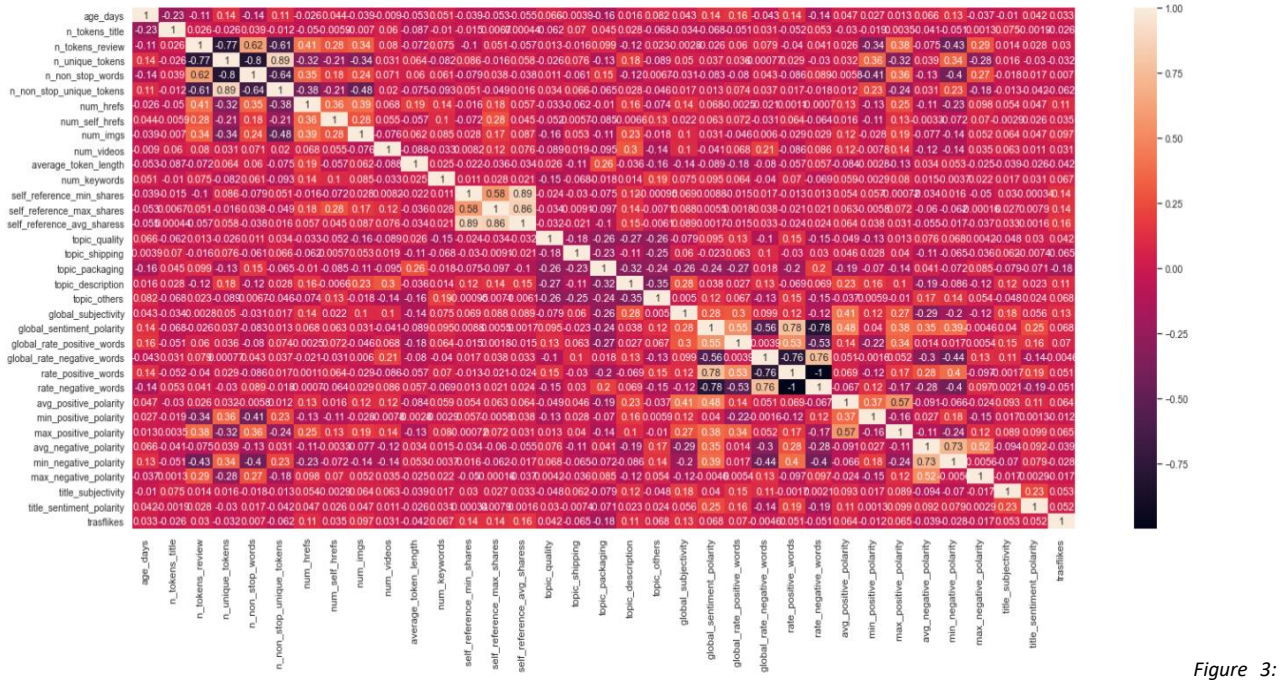

Figure 3: Correlation matrix of the numerical variables

As far as the categorical variables are concerned, dummy variables were created to be included in the developed models.
These two datasets were merged again, leaving out the target variable that has been used to evaluate the performances of the models.

Model training and testing

Data was now ready to train the different algorithms: it was split into training set (70%) and test set (30%). The final goal was to find the best supervised learning algorithm in terms of MAE.

To do so a Grid-search was adopted to find the optimal hyperparameters of the different regression models. A 5-fold cross validation was included into the Grid-search to increase the robustness of the algorithms, in order not to make the results dependent just on the split chosen for the dataset.

In this way the best combination of hyperparameters for each model refers to the mean of the (negative) MAE score metrics.

In some cases, it occurred that the hyperparameters found in the grid search resulted in a significant gap between the MAE train and the MAE test and for this reason some hyperparameters were changed to reduce this distance.

Models' evaluation and final remarks

Many supervised learning algorithms were run, and their respective test MAE are shown in the table below.

Model	**MAE test**
Linear model	2109.63
Ridge	2109.65
Lasso	2109.74
KNN regressor	2110.55
Decision tree regressor	2132.29
MLP regressor	2123.19
Random Forest regressor (450 estimators)	2085.84
Bagging regressor (100 MLP estimators)	2117.46
Gradient boosting regressor (300 estimators)	2083.66
Adaboost (100 estimators)	2116.10

Stacking regressor (estimators: the best Adaboost, the best Random Forest, the best Gradient boosting, the best MLP; final estimator: Ridge regressor)	2075.20
Voting regressor (estimators: the best Adaboost, the best Random Forest, the best Gradient boosting, the best MLP)	2085.63

Looking at these results it is evident that the simpler models perform better than some ensemble more complex algorithms.

However, even if the difference in terms of the test MAE is relatively small, since this metrics is an average considering many thousands of observations, the three best models, the Random Forest, the Gradient Boosting and the Stacking regressors, were considered for the decision of the final model.

By plotting the residuals vs the fitted values and their histograms considering the train set, it was chosen to adopt the Gradient Boosting model because its residuals better respect all the assumptions behind the regression models. In fact, despite of very few observations, they follow a random pattern, and they seem to behave normally with a mean close to 0 and a constant variance.