Predicting News Popularity on Kaggle - Report

Anna Corretger, Santhosh Narayanan, Guglielmo Pelino

Exploratory Data Analysis

Our first step consisted in basic explorations of the data: identify possible outliers, handle missing values in the dataset and almost constant variables, explore possible correlations between features.

Outliers

We removed the following observation: dataset <- dataset[-which(train$n_unique_tokens == 701),]

Missing values and handling with/without imputation

We noted that in the dataset there were many missing values: for treating them we used the function "handle.missing" in MedMast package, which either removes or imputes them. Later on in the modelling analysis we realized that imputing them was not a good idea, expecially for classification trees methods such as random forest: thus, we created a binary feature which was a flag for the missing values. Also, we realized that dataset$n_non_stop_words was a constant feature, and thus we removed it from the dataset.

Correlated predictors

We observed that dataset$rate_negative_words was completely determined by the rate of positive words which was present as another feature, and thus removed it from the dataset.

New features

As previously seen in the missing values section, we first of all added a flag feature called "missing.flag" which is 1 if the row had missing or nonsensical values.
We then added date variables (year, month and quarter) in order to exploit time and seasonality as a factor in the underlying dynamics of popularity of the outcoming news.
For analyzing the actual text in the url of the article, we created different bins for different popular topics categories in the url: tech brands, gossip, politics, sports, socialMedia, cars, tvshows (check Max_0.5355.R file in Code folder on github), which were stored in 7 distinct binary additional features.

Our MetaModel: Rolling Windows

Model Selection & Tuning

- Random forest, Neural Networks, Boosting, SVM choice of parameters and motivations
- Comparison between them -plots for optimizing parameters

Final Model

- Ensemble techniques: averaging different predictions
- Anna's text mining to get 4's and 5's