Machine Learning and Data Mining project: Twitter Food Popularity

Gianluca Guglielmo[1] and Giulia Tranquillini[2]

[1] problem statement, solution design, solution development, data gathering, writing
[2] problem statement, solution design, solution development, data gathering, writing

Course of AA 2020-2021 - DSSC & Theoretical Physics

1 Problem statement

The purpose of this work is to obtain an effective method for predicting the relative popularity of a tweet from an American fast-food company with respect to their historical retweeting performances. We've chosen this niche because of their similarities in their social media strategies. We thought this would be a useful tool for an emerging company to make the most out of its Twitter presence, so we tried to assess performance using handy predictors for their Social Media Manager to exploit, such as best time and date for posting.

2 Assessment and performance indexes

We downloaded and analyzed a database consisting of about 45000 tweets from 13 established American fast-food companies.
These companies were chosen after consulting the "List of major fast food chains" article on Wikipedia. We then selected by hand a batch of them that respected these prerequisites: a clear social media strategy, at least 1M followers and a limited number of replies to users.
Articles [1] and [2] considered the number of retweets to be an optimal response variable for this kind of problem. However, since we had to deal with Twitter accounts that had different *reaches* and *engagements*, we couldn't simply use retweets to estimate relative popularity. We decided to use as numerical response variable a *popularity rate* from 1 (not popular) to 10 (extremely popular) according to which decile of each fast-food's retweets distribution they belonged to. We then set up a regression problem.

Data was collected using the Twitter APIs and each account initially gave us about 3500 tweets. Pre-processing involved cleaning out about 40000 tweets that were customer service replies from companies, hence not intended to be popular. On the other hand, by filtering for appropriate words and emojis, we left in the dataset not just promotional tweets but also about 2000 tweets that were humorous replies to customer *memes* about specific food items, which proved to be successful and frequently received retweets and replies.

3 Proposed solution

We implemented a variety of appropriate regression methods and compared their performances using the RMSE. The baseline for comparison was a predictor that, for every observation, output the mean value of the ten quantiles (5.5).
To filter out the less relevant predictors and speed up the optimizing process, we tested all of them together by training a *Random Forest*, selecting the most promising ones by looking at the *Increase in Node Purity* index. Using only the strong predictors, we then tuned the *RF*, performed a *Support Vector Regression* (SVR) and finally a *Boosting*.

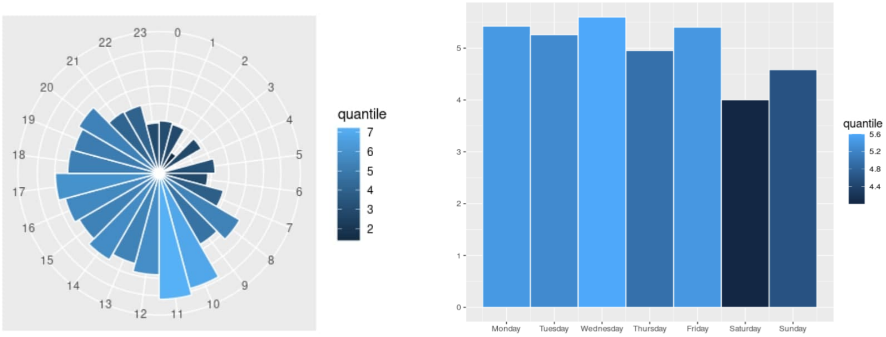

Figure 1: Dataset's Tweet popularity by hour and by weekday

4 Experimental evaluation

4.1 Data

According to both the cited articles and to our personal analysis, we initially extracted 40 categorical and numerical features from tweets' *data* and *metadata*. From the tweets' texts we extracted categorical features indicating the presence of links, emojis, questions, long words with more than 10 characters and a variable indicating whether the first word was capitalized or not. On the numerical side, we extracted the total number of characters, mentions, hashtags, sentences and, using POS tagging, also the number of verbs, adverbs, adjectives, nouns

and pronouns. Moreover, using text mining after removing stop words, punctuation and stemming the terms, we tried to integrate a Bag-of-words approach with the 20 most frequent words in the dataset. We chose these text formatting guidelines since they can be useful to a SMM when developing new tweets.
From *metadata* we extracted the weekday, time, the presence of either photos or videos and information indicating whether the tweet is a reply, a retweet, a quote or neither of them.

4.2 Procedure

Random Forest

As described earlier, we first removed the least important predictors. Then, we implemented a search grid for finding the best values of the hyperparameters `mtry` (number of variables to be used for each tree) and `nodesize` (minimum size of terminal nodes), using a 10-fold CV for each pair of values. As expected for a regression, the best value of `mtry`= 8 is almost a third of the parameters left. After the training we observed the variables importance graph according to the Node Impurity Rate :

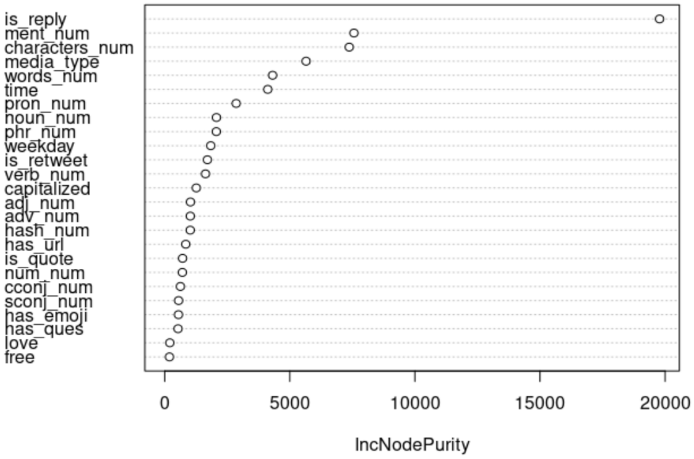

Figure 2: Ranking of variables

Support Vector Regression

We performed *Support Vector Regression* (SVR) to exploit its flexibility and distance-based approach. We tried out different types of kernel to check which fitted best. This preliminary operation was executed without going deeply into

optimization: the *polynomial kernel* was chosen.
A 10-fold cross validation on a 3-dimensional grid was performed to tune the margin of tolerance ε as well as the other hyperparameters:

ε	C	Degree
0.6	1	4

Table 1: SVR Parameters

Boosting

Boosting was chosen to try and exploit the weak learners influence on the predictions. Its tuning was carried over 5 parameters: execution time quickly became an issue so we could not check all the optimal combination and ended up trading some precision to get some speed.

4.3 Results and discussion

The following table sums up the RMSEs for each learning method:

Learning algorithm	RMSE
Mean	2.87
RF	2.04
SVR	1.96
Boosting	2.12

Table 2: Final results

All the methods performed better than the mean-predictor. The SVR is clearly the best one, with an RMSE that is roughly 68% of the baseline's. According to the *Node Purity* plot, bag-of-words predictors were some of the least-significant, hence almost all of them were removed after the first RF trial. The Reply parameter played a big role in the results, since it was the single most important one. This was probably due to the generally low number of retweets in respect to the rest of dataset.
There is great room for improvement. To begin with, we could have used a tf-idf approach instead of a simple BOW, and then we could have built a practical fast-food-related dictionary in which to find the best terms to use. Sentiment analysis could have been tried too, but it probably would not have been useful, since every tweet from a major corporation tries to convey the same positive emotion. Finally, using image detection paired with color analysis [3], we could have found further useful patterns in posts with media attachments.

References

[1] Y. Zhang, Z. Zu, “Predicting the Popularity of Messages in Twitter using a Feature-weighted Model”, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

[2] K. Fiok, W. Karwowski, ”Predicting the Volume of Response to Tweets Posted by a Single Twitter Account”, University of Central Florida, Orlando, 2020.

[3] S. De, A. Maity, V. Goel, S. Shitole and A. Bhattacharya, ”Predicting the popularity of instagram posts for a lifestyle magazine using deep learning”, 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), Mumbai, 2017.