YouTube Trending Investigation and Prediction

Shihan Wang, Rufeng Lin, Xi Jiang, He Sun

Rady School of Management, San Diego, CA, U.S.A.

Abstract

Breaking down the trending algorithm of YouTube is a common goal for YouTubers and media corporation. Once the content generators identify the secrets behind, they could leverage the power to make frequent appearance on the trending list. This study is designed to assist video makers in boosting the likelihood of making a trending video through certain approaches. To realize the objective and obtain the most precise prediction, there models have been implement and compared (Linear Regression, Lasso, and Random Forest). Based on text mining and regression models' performance, it is found that random forest regression is the optimal models among all the trials.

1. Introduction

YouTube, as a worldwide leading video-sharing platform, offers a wide variety of user-generated content and corporate media videos. As of May 2019, more than 500 hours of video content are uploaded to YouTube every minute (Hale, 2019). To inform the audience of the trending news or hit videos, YouTube lists top trending videos on its website. The trending section is one of the most influential, yet mysterious features (Townsend, 2018). Nearly all of the content generators were eager to be listed on trending page so that the view count of their videos will grow exponentially and bring them appreciable incomes. Meanwhile, those content generators will gain numerous subscribers in just minutes, enabling them to make more money through higher video views in the future. Given the huge benefit of popping up into the trending page, more and more YouTubers are trying to crack YouTube's trending algorithms to increase the likelihood of turning into trending videos. To realize the objective and obtain the most precise prediction, there models have been implement and compared (Linear Regression, Lasso and Random Forest). Based on text mining and regression models' performance, it is found that random forest regression is the optimal models among all the trials.

In this work, the primary task is to investigate the key features in predicting the view counts and trending time (days between publishing and making into trending section) of YouTube videos using a daily record of the top trending YouTube videos dataset acquired from Kaggle.

The data of English-speaking regions (United States, Great Britain, Canada) are extracted from the original version and used as the final dataset, consisting of up to 200 listed trending videos per day from Nov. 2017 to Jun. 2018 in terms of trending data. 120,747 samples are analyzed in this work and the major columns are trending date, title, channel title, category, publish time, tags, views count, the number of likes and dislikes, comments count, description, and regions. A key column called trending time is equal to the difference in days between publishing and trending. To get a preliminary understanding of the dataset, exploratory analysis is essential for visualizing the distribution of dataset in terms of predicted variables and multiple potential predictors.

First, visualization shows the number of trending videos in top 10 categories and the majority of trending videos come from the entertainment category within selected regions, followed by the category of music, people and blogs, and news and politics. Established on this distribution, it is reasonable to infer the category is a good start for the baseline model, but cleaning and processing work needs to be done for this variable.

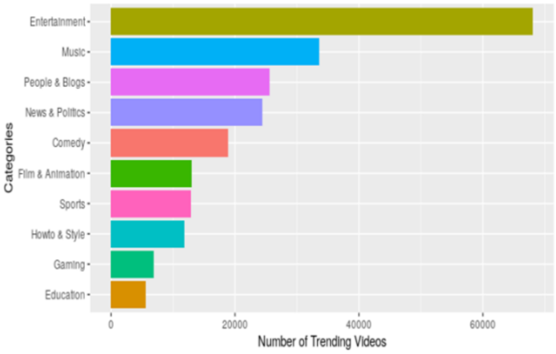


Figure 1: Number of Trending Videos in Top 10 Categories

Second, word cloud is implemented to explore the most popular tags and conclude that “funny” and “comedy” tags are most frequently labeled on the trending videos. That being said, entertaining content is more likely to make into the trending page, indicating that the tag variables should be included in the initial model.

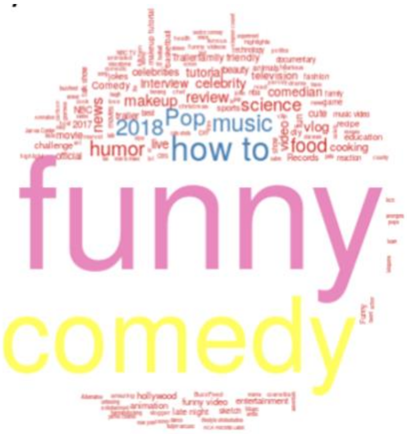


Figure 2: Tags by Popularity

Third, distributions of trending time are visualized across three regions. Observably, trending videos from Canada spend the least time to make a hit. Specifically, most of the videos are listed under trending tab within 6 days after initially publishing. While things are quite distinct from the United States and Great Britain, the trending times are significantly longer than the CA region. In detail, data of US trending time is mostly distributed within 15 days, nevertheless, data of GB is more evenly distributed within the range of 40 days. Given the analysis of trending time distribution, it can be persuasively forecasted that region variable is considerably powerful in building the model.

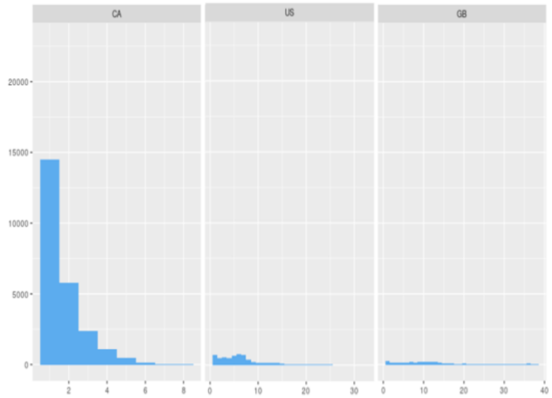


Figure 3: Distribution of Trending Time

Fourth, the database is divided into the title variables and extract the most commonly used title words in each popular category and summarize the times that each word shown up. Interestingly, certain words appear much more frequently than the expectation, leading to take title word count into consideration in constructing the predicting model.

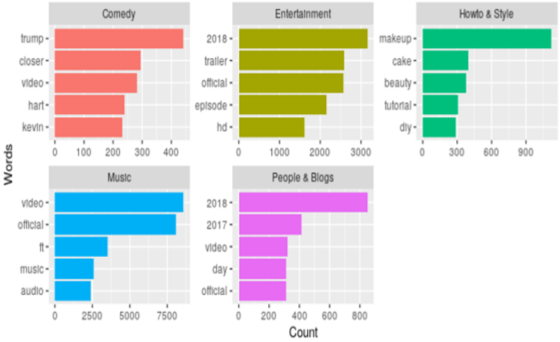


Figure 4: Top 5 Title Words by Top5 Categories

According to the data exploration and visualization, the variables discussed above are included in the model. Additionally, an advanced research should be conducted to appropriately transform and recode those variables to build a better predictive model.

2. Variable Selection

2.1 Assess the validity of prediction

After determining the dataset and which exciting phenomenon to study, the first thing to do is making sure what kind of dependent variables are needed to evaluate the initial hypotheses and what type of result might contradict common sense.

To make this research more valuable, the popularity of a video directly reflects its commercial value (no matter in a positive or a negative way). What is more, YouTube, the No.1 popular video website throughout the world, is obviously the primary investment objective of many related business companies. Therefore, the principal task of this research is to find out which video is the most valuable, according to an investor's perspective.

2.2 Process some duplicate data, outliers and classification data

(a) Preprocess the valuable data

First, for validating the validity of the model, it is common to divide this dataset into a training set (80%) and a validation set (20%)

Although the entire database has 10 data of countries from different continents, to reasonably analyze the features of tags and titles, only the data from English-speaking countries (U.S, Canada, and Great British) are selected.

(b) Remove duplicates:

After carefully examining the database, there are many duplications exist in the dataset since some videos might keep trending or appearing in data lists of different countries.

Given the actual economic use, the US data has the highest priority, and then Canada for the North America area since the commercial value of the US, Canada, and UK markets decreases in sequence. For example, when a video is trending in the lists of 3 countries simultaneously, only the data of the US will be reserved. Analogously, when a video is not trending in the US but in Canada and in the UK, only the rows of Canada will be kept.

For those videos staying in the trending leader board for several days, the shortest time interval between the publishing date and trending date is picked because it is more reasonable to focus on the fastest trending time.

(c) Remove outliers:

To make the model more accurate, those data which has a trending time more than 100 days got eliminated because they are not economically significant but deviate the prediction model.

(d) Process skewed data:

Meanwhile, some variables are also skewed, even though extreme outliers are removed. For example, some videos may publish in early 2000 but become trending recently, so their trending time will be much longer than the majority data. To reduce skewness, it is reasonable to apply a log transform to those variables with the skewness' absolute value higher than the threshold. According to the result of the regression, the optimal threshold is 0.75.

2.3 Select and filter predictors

When selecting and filtering variables, the first task it to observe the significance of each variable and how important it is to the problem

For a video, the most significant parameter is undoubtedly the number of views. It is the simplest way to assess whether a video is popular, so it is more evident for investors who are looking for an appropriate video to run their advertisement or post the names of their brands aside the video.

(a) Select variables based on common sense

For the initial trial, the variables below are incorporated into the baseline model:

1. The number of likes, dislikes and comments
2. The time interval between the publish date and trending date (the date when the video enters the Top 200 Leaderboards)
3. The length of the title

Intuitively, the number of likes and comments are the most significant and impactful predictors. Also, the number of dislikes might result in a more substantial amount of views.

(b) Create potential predictors

Create six additional columns since the popularity of each tag is another potential predictor which might be significant. The detail idea about how these variables are established is that a big tags dictionary including the repeat times of each tag is built first and for each video's tags, calculate the relationships between them and the whole tag dictionary. For instance, mean tag count is got from each tag's total repeat time divided by the tag number of each data.

1. Average counts of all the tags
2. The sum of all tags' counts
3. Counts of most popular tags and least popular tags
4. If this tag belongs to the top 10 or 20 most popular tags.

In the subsequent trials, the attribute of title and description, like the length of them, are also essential for the precision of the final model:

1. Length of title and description
2. The underlying subjective emotion in the title and description
3. Existence or frequency of specific keywords in the title and description
4. The complexity of title and description

(c) Avoiding multilinear issue

To select predictors for the final model, investigating the relationship between all the numeric variables by drawing the heatmap are vital for avoiding the double counting issues.

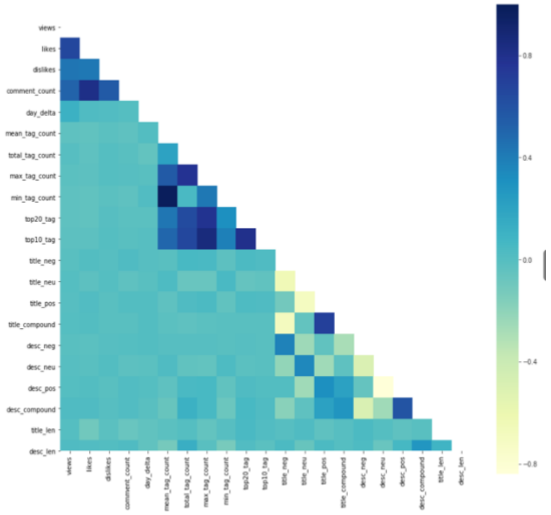


Figure 5: The Heatmap of Variable's Correlation

As the heatmap shows, the Min_Tag_Count and Mean_Tag_Count show the extremely high correlated relationship, so it is necessary to eliminate one variable. Since Mean_Tag_Count has a more significant correlation with the number of views, it is incorporated into the model. Base on this variable-selecting logic, dummy variables like Title_Pos and Desc_Pos (whether this title or description contains a positive sentiment) are also took out from the basic model.

(d) Create dummy variables for categorical features

The cultural and regional preference is another impactful predictor, so the category and country of these videos. For example, videos in Spanish in the UK are not as popular as that in the US. The significant American political news may not be prevalent in other countries but widespread domestically.

1. Category of video (musical, sports, etc.)
2. Certain countries

2.4 Select models and tools for prediction

The prediction objective, number of views and trending time are both numeric variables, so the ideal model comes into mind is the regression model. Here are several optional models:

1. Linear regression
2. Ridge regression
3. Lasso-regression

and tools that may be adopted:

1. Bags-of-words
2. TF-IDF
3. Sentiment analysis
4. Similarity (Jaccard, Cosine, Pearson)

Both Bags-of-words and TF-IDF are really efficient tools to extract the frequency of each keyword. The existence of the most popular keywords in the title or description is pretty likely to result in trending and a large number of views of this video. For example, "Trump" is one of the most popular words according to intuition, so when the title of the video contains this keyword, the video is more likely to get trending and viewed by more people.

3. Model

The target is to predict both the view count and the trending time of the videos. The trending time represents how fast certain video can be trended. By analyzing the data, other related features can be found out through both text mining and dummy encoding methods.

As both the view count and treading time are numerical variables, the simple linear regression could be a good start to train and predict the data. The linear-regression model is simple and efficient, and has the widest accepted fields. However, the results may be not very satisfying and reliable, both underfitting and overfitting can happen. The second model is the Lasso regression, which can process variable selection and regularization at the same. The benefit of this model is that it can handle both the continuous and discrete variable, as well as select the most relative variables to avoid overfitting. On the other hand, Lasso may also ignore nonsignificant variables which may have important practical meaning. The mathematical formula of this model:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \le t.$$

The optimal way to improve this model is changing the lambda parameter to penalty the variables in different extent and compare the MSE and R-square of each regression results. The third model is random forest regressor based on bagging arithmetic, which has higher regression efficiency and more reliable estimators. But the time cost of this method could be relatively high when the dataset and variables are big. The grid search of max_depth and n_estimators at the same time will finally come out a relative best parameter.

Although the whole database has 10 countries' data in different continents, to fairly analyze the features of tags and titles, only the data from English-speaking countries (U.S, Canada, and Great British) is selected. Based on the practical meaning and the object of the prediction, it is more rational to ignore the number of likes, number of dislikes, and comment counts because they are not proper explanatory variables to predict the view count and the trending time. To decide which video has more potential to be trended and popular, concentrating on the tags, titles, description, categories, and countries makes more sense.

4. Literature Review

The original dataset is obtained from Kaggle, comprised of a list of popular YouTube videos in various categories and different languages. In the prediction tasks, only English-speaking videos are taken into account as analyzing their trending time. There are 19 different features for each trending video defined in the original data, but the predictive model considers 35 observations in total after incorporating more influential factors. With the application of random forest algorithm, the five most significant features are attained, which are description length, description compound, title length, total tag count and max tag count.

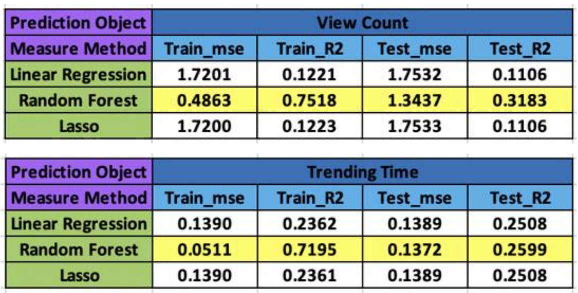
There is a similar dataset called "Credit Card Fraud Detection" on Kaggle, and the common algorithm applied to perform regressions is random forest as well. With the help of this method, the large dimensionality dataset can be handled properly to avoid overfitting problem when running the regression. In addition, in comparison with other traditional regression methods such as linear regression, multiple linear regression and polynomial linear regression, the performance of the predictive model generated by the random forest method is the best in terms of MSE.

In addition to random forest, there are two more state-of-the-art regression methods, projection pursuit regression (PPR) and support vector machine (SVM). According to the test performed in the long-term travel time prediction paper, random forest is the best regression approach among the three methods. Even though the results obtained from PPR are more accurate, there will be more pre-processing work associated when applying it. As PPR and SVM are not tried in the predictive model, it is not assured that whether the other two methods would outperform random forest or not. However, with respect to the referred paper, random forest is in general the most promising method which is consistent with the results.

5. Result and Conclusion

By splitting the database into 80% of training set and the rest to be the testing set, the regression process and the prediction of both the view count and trending time is followed. To examine the performance of each regression model, four estimators are calculated: train MSE, train r-square, test MSE and r-square. The results are shown in the following table:

Table 1: The Regression and Prediction Results



Prediction Object	View Count			
Measure Method	Train_mse	Train_R2	Test_mse	Test_R2
Linear Regression	1.7201	0.1221	1.7532	0.1106
Random Forest	0.4863	0.7518	1.3437	0.3183
Lasso	1.7200	0.1223	1.7533	0.1106

Prediction Object	Trending Time			
Measure Method	Train_mse	Train_R2	Test_mse	Test_R2
Linear Regression	0.1390	0.2362	0.1389	0.2508
Random Forest	0.0511	0.7195	0.1372	0.2599
Lasso	0.1390	0.2361	0.1389	0.2508

In the view count prediction, to get the optimal parameters of each model. Grid search methods are implemented in the random forest model to find the optimal Max_Depth and N_Estimators, the final value of them is 31 and 200. The depth refers to the maximum depth of the tree, it sets how specific the regression is. And the estimators mean the number of trees in the forest, it can decide how strong the final model is. The lasso regression needs to find the optimal alpha through the for loop to give enough penalty to certain variables. Because the variables are not very complicated, the alpha is set to be 0.0001. However, the lasso model and simple linear regression (baseline)'s performance is quite similar. It seems that the penalty is not necessary because the explanatory variables' correlation is basic similar. The random forest model has a much better results compared to the other two models, especially the r square value, which increases to a large extent. However, the random forest model doesn't perform really well in the trending time's prediction. The overfitting situation happens obviously in these aspects because the data structure is not really rational. Two possible explanations of this phenomenon: the dataset is not adequate to engage in proper training; more relevant features need to be figured out.

Regarding the random forest results, top 5 most important features are description length, total tag count, title length, max tag count, and description compound. Because these 5 features all have positive Gini values, it can be assumed that videos with these high features have more potential to be popular. Overall, this model provides some business insights to both the youtubers and the advertising companies.

Reference

[1] Hale, J. (2019, May 7). More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute. Retrieved from https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/

[2] Townsend, B. (2018, February 6). The Secrets Behind The YouTube Trending Page - And How To Get A Video Trending. Retrieved from https://www.wetheunicorns.com/features/how-youtube-trending-page-works/

[3] Mendes-Moreira, J., Jorge, A. M., Sousa, J. F. D., & Soares, C. (2012). Comparing state-of-the-art regression methods for long term travel time prediction. *Intelligent Data Analysis*, 16(3), 427–449.

[4] Nidaguler. (2019, October 20). Credit Card Fraud Detection. Retrieved from https://www.kaggle.com/nidaguler/credit-card-fraud-detection/notebook.