**Final report**

**Executive Summary**

Every minute, there is over 400 hours length of video generated on YouTube. It is attracting more and more content generators and advertisements. Videos with high views affect people’s opinions and willings dramatically. In other words, making a YouTube video more popular contains many valuable business interests such as advertising.

However, whether a video gets more view counts depends on multiple factors. Some videos do have fabulous content, yet not being noticed. In fact, people click into a video even before they see the content inside. Thus, how to describe and package a video becomes very important.

There are many tutorials online trying to teach people how to increase their YouTube videos’ views. However, these guidances are based on experiences and are biased because of the lack of data.

Our project is to find out what attributes besides the video content are affecting the number of views. Further, we want to give guidance on how to construct and package a video in order to get it more viewed.

The dataset we used contains over 40,000 records of videos. We first did feature engineering, added more features according to some online guidance and more features which we think would be important, and leave them to be tested in models.

We trained 4 different models: linear regression, random forest, decision tree, and gradient boost. We utilized 10-fold cross validation in every model. We used these models for two goals, one is to find out the top 5 important attributes, the other is to test whether we can use those attributes to predict a video’s number of views.

We found out the five most important attributes are channel frequency, video category, title length, tag length and publishing hour. However, as the R square score are under 0.3 for all models, and MSE of every model is pretty high, we can’t say we can predict a video’s view simply according to how the video is packaged.

In conclusion, we generate several suggestions for stakeholders. We suggest YouTubers and channels to post as frequently as possible, post earlier in a day, use less than 10 tags, avoid stop words in tags and post more videos about film, music, entertainment, gaming, people and blogs.

Our project also has some defects. Because we utilized data from YouTube’s trending channel, the categories are not covering all possible channels. YouTube generates a large number of videos every day, but we didn’t use a very large dataset and compute big data analysis.

**Problem Statement**

As one of the biggest video-sharing website, YouTube has provided many business and advertising opportunities. Over 400 hours of videos are uploaded on YouTube in every one minute. There are many traditional media platforms as well as individual online celebrities joining the game. However, only a few of the videos became a hit.

Besides having good content, many wrapping factors are influencing the number of views on YouTube. Even thousands of tutorials can be found online, trying to teach people how to make a trending video. However, most of the tutorials are based on personal experience, they are biased because of the lack of data.

Based on this context, the team analyses what factors are contributing to high views on YouTube using more than 40,000 of video statistics and utilize the data science technology to reveal how to package a video to help it become popular.

**Project goal**

(1) Find important attributes that are powerfully influencing a video’s number of views.

(2) Give specific guidance on how to construct the video’s title, tags, descriptions, publishing time period in a day, etc., in order to have a higher number of views.

**Dataset Description**

The data comes from the Kaggle dataset. It is the trending YouTube Video Statistics collected using YouTube API, which is a record of several months of daily top trending videos on YouTube (US region). The size of the dataset is 40,000, containing 16 attributes including both numerical and categorical features. Sample numerical features are views counts, likes counts, dislikes counts, publish time, comment counts. Sample categorical features are video title, category, channel title, tag words, video descriptions, thumbnails.

**Feature Engineering**

1. **Initial Dataset**

Here is an example of raw data:

**Video\_id:** 2kyS6SvSYSE

**Trending\_date:** 17.14.11

**Title:** WE WANT TO TALK ABOUT OUR MARRIAGE

**Channel\_title:** CaseyNeistat

**Category\_id:** category\_id

**Publish\_time:** publish\_time

**Tags:** Shantell Martin

**Views:** 748374

**Likes:** 57527

**Dislikes:** 2966

**Comment\_count:** 15954

**Comments\_disabled:** False

**Ratings\_disabled:** False

**Video\_error\_or\_removed:** False

**Description:** One year after the presidential election, John Oliver discusses what we've learned so far and enlists our catheter cowboy to teach Donald Trump what he hasn't.\n\nConnect with Last Week Tonight online...\n\nSubscribe to the Last Week Tonight YouTube channel for more almost news as it almost happens: www.youtube.com/user/LastWeekTonight\n\nFind Last Week Tonight on Facebook like your mom would: http://Facebook.com/LastWeekTonight\n\nFollow us on Twitter for news about jokes and jokes about news: http://Twitter.com/last week to:night\n\nVisit our official site for all that other stuff at once: http://www.hbo.com/lastweektonight

**2. New dataset**

Since one of our goal is to confirm the online suggestions about how to make your video more popular, we need to extract features that fit the rules we want to confirm. Based on the initial dataset, we followed some online guidance to construct more features. The final features constructed are as follows.

**Input features:**

1) Channel Frequency

According to the dataset, during 213 days, different channels have different numbers of video posting. We calculate the video posting frequency and categorize it into three types: low, moderate and high. Then coded three categories into 0, 0.5, 1.

2) Publish hour

Initial dataset indicates when the video was published. It is accurate to second. The publishing time may have a correlation with views of a video. However, we just need to know the rough interval when the video was posted. Thus, we constructed this feature of publishing hour, with values from 0 to 23 (both inclusive), then normalized this column into [0, 1].

3) Video Category

Initial dataset includes the category of a video. Categories may include sports, music, animal, etc., indicating the overall content category. We decided to retain this feature, as it may determine a video’s view. This attribute values from 0 to 22, all are integers.

4) Tag Words Length

This column contains integer values, indicates the number of tags used. It ranges from 0 to 69, with a mean of 19.74. We then normalized this column into [0, 1].

5) Title Length

This column contains integer values, indicates the length of words of the title. It ranges from 0 to 24, with a mean of 8.53. We then normalized this column into [0, 1].

6) Number of Stop Words in Title

Containing to much words that are meaningless in your title may get your audience bored quickly. We added up the number of stop words in the title as an attribute, then normalized it into [0, 1].

7) Number of Stop Words in Tag

Tags are designed to show the keyword of a video. However, if it contains to many stop words, it won’t work for representing the content and effects in searching. We counted the stop words in tags and normalized it into [0, 1].

8) Brackets in Title

This is a guidance saying that if you contain brackets in your video’s title, it will attract more clicks. To confirm this, we added this attribute of whether contains brackets in title. It is a binary value.

9) Numbers in Title

Similar as brackets in title, we want to confirm how numbers in the title are affecting the views. This is also a binary value.

10) Special Characters in Title

To generalize the constraint, we want to see if containing any special characters other than just words and numbers will affect a video’s view. This a binary value.

11) Contains Facebook Links in Descriptions

Some online guidance indicates that containing your own social media link will help your video be more popular. To confirm this suggestion, we included this feature of whether the YouTuber contains a Facebook link in their video description. This is a binary value.

12) Contains Instagram Links in Descriptions

Similar as the 11th, we want to see if including an Instagram will make a difference. At the same time, we can compare which social media is functioning better.

13) Category of title length

We categorized title length into three types, short, medium and long. We then coded it into 0, 0.5 and 1.

14) Category of tags length

We categorized tags length into three types, short, medium and long. We then coded it into 0, 0.5 and 1.

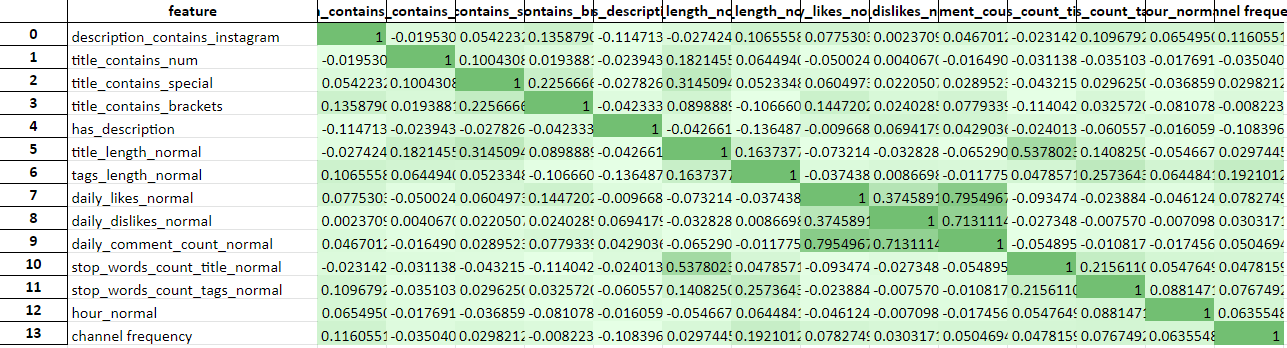
**Output feature:** Average number of daily views

Since the videos were posted online for different length of days, we normalized the video views by dividing the number of days they have been existed online.

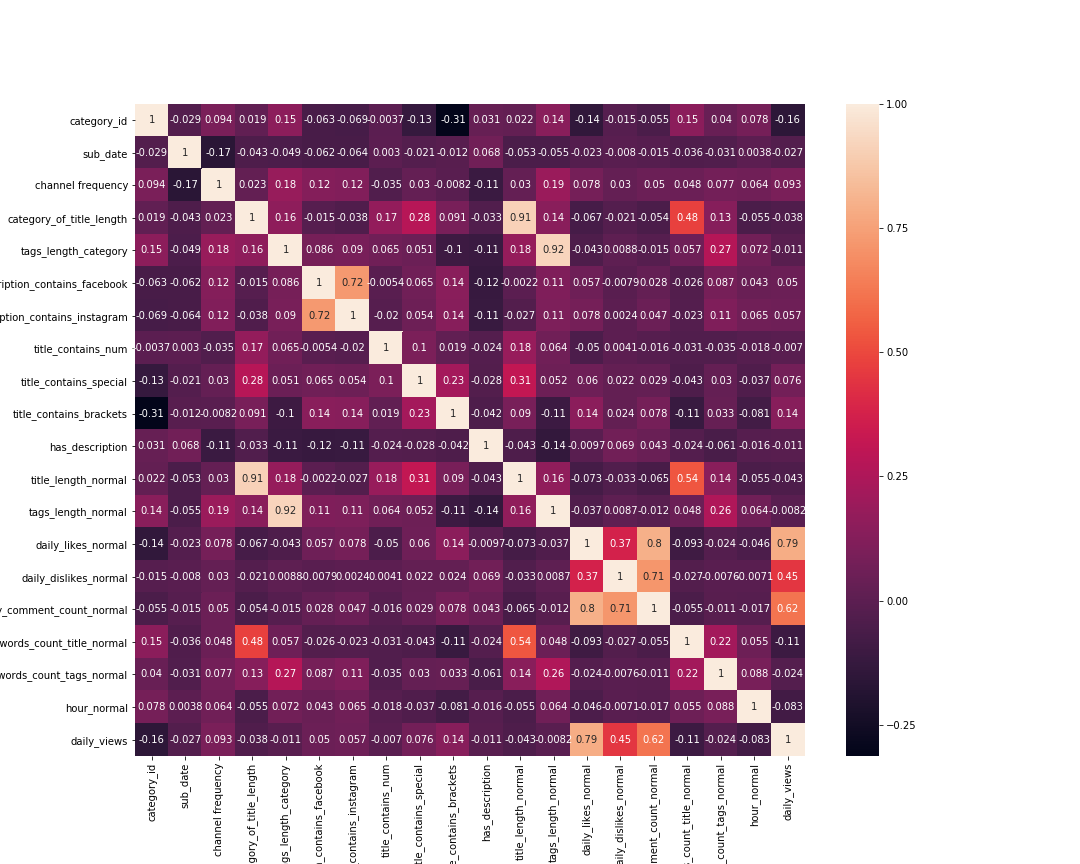
To do this, we first calculated the days by subtracting record day by its publish day, to get the total day it is online. Then we divide the accumulated views by total days, to get average number of daily views, as the output of our model.

**3. Correlation Analysis**

**Pearson Correlation Coefficient**

****

**Heat Map**



According to the correlation analysis, we found out that the “ Title Length” and “Number of Stop Words in Tags” have a strong correlation, so we only use “Title Length” when training the linear regression model.

**Model Development**

1. **Linear regression model**

The model is based on thelog ("daily\_views") and normalized input parameter to [0,1], using 10 folds cross-validation and LASSO.

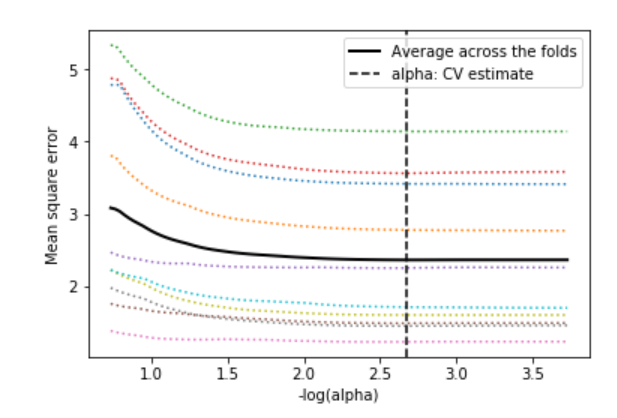
**Input parameters with weights:**

1. Channel Frequency: 2.35
2. Publish hour: -0.302
3. Tag Words Length: 0.589
4. Title Length: -0.463
5. Number of Stop Words in Tags: 0
6. Brackets in Title: 0.230
7. Numbers in Title: 0.087
8. Special Characters in Title: 0.149
9. Contains Facebook Links in Descriptions: -0.051
10. Contains Instagram Links in Descriptions: 0.189
11. Category 1 'Film and Animation': 0.246
12. Category 2 'Cars and Vehicles': 0
13. Category 10 'Music': 0.700
14. Category 15 'Pets and Animals': -0.325
15. Category 17 'Sport': 0
16. Category 19 'Travel and Events': 0
17. Category 20 'Gaming': 0.601
18. Category 22 'People and Blogs': 0.052
19. Category 23 'Comedy': 0.164
20. Category 24 'Entertainment': 0.002
21. Category 25 'News and Politics': -0.934
22. Category 26 'How to and Style': -0.385
23. Category 27 Education': -0.395
24. Category 28 'Science and Technology': -0.120
25. Category 29 'NonProfits and Activism': 0
26. Category 43 'DK': 0
27. Intercept:9.33

**R Square:** 0.235

**MSE:** 2.466

**Linear regression plot**

****

According to the result of the plot, there is no specific correlation between the real value and the predict value, which means the prediction of the linear regression does not performance that kind of well.

**(2) Random Forest Regressor**

We also tried nonlinear models to reveal the correlations between attributes and video views.

The random forest regressor is an ensemble learning method for regression by constructing multiple decision trees during training. The outputs are mean predictions (regressions) of individual tree. Therefore, the random forest regressor can give out the importance of features as indication for future application.

The input data are normalized and the input features are :

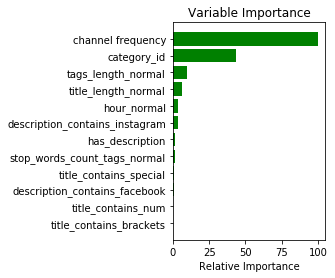
1. Video Category
2. Channel Frequency
3. Publish hour
4. Tag Words Length
5. Title Length
6. Number of Stop Words in Tags
7. Brackets in Title
8. Numbers in Title
9. Special Characters in Title
10. Contains Facebook Links in Descriptions
11. Contains Instagram Links in Descriptions

The output is the log value of the daily views.

The Video Category is added as an additional input feature in the random forest regressor as to the input features in the linear regression model. A 10-fold cross-validation is used to train the data and overall training for all data points is performed as the last step. Some parameters for this model are shown below:

1. Maximum tree depth: In order to avoid overfitting, the maximum tree depth is set to 5.
2. Split criterion: To measure the quality of split, the mean squared error is used as criterion.
3. Number of estimators: 700.

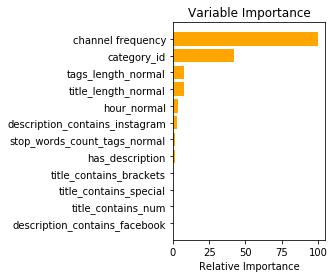
The result gives out the relative importance of each feature with the percentage of 100% as the most important feature:



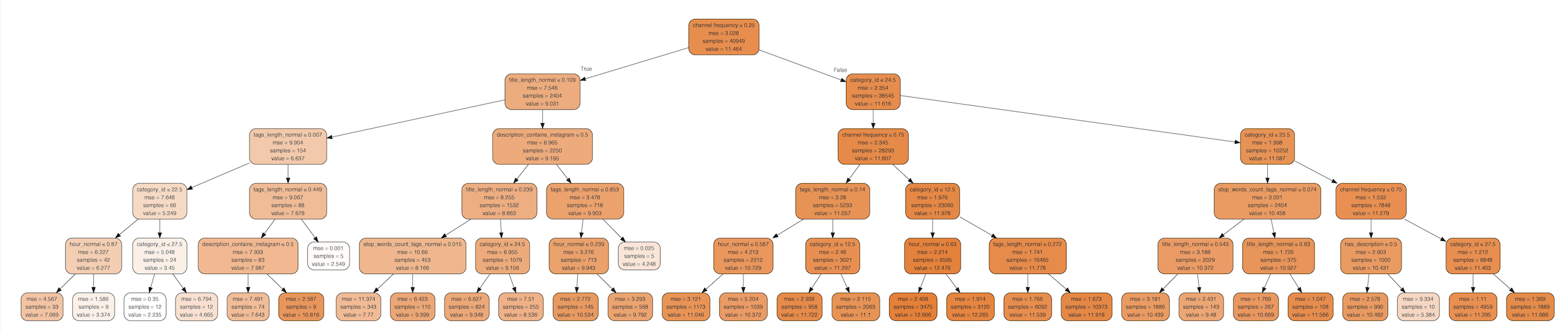
As shown on the figure, the most relevant feature to the daily views is the channel frequency. The video category also has a great effect on the average daily views, while title contains numbers and title contains brackets have the least importance.

**(3) Decision Tree**

As to check the match of results between different model, a decision tree model is also used to find the importance of features. The input features and output values are the same as those for random forest. In addition, the maximum tree depth, split criterion and the number of estimators are also the same as those for random forest. Similarly, a 10-fold cross validation and a final overall-data training are performed to produce a comprehensive model. The result of relative feature importance is shown below:



After comparing the result of feature relative importance, the top 5 most important features are the same as to those for the random forest. However, the relative importance percentage between each features are different from the random forest method. In order to look detailed values for the tree nodes and see how the decision is made, the final decision tree is shown below:



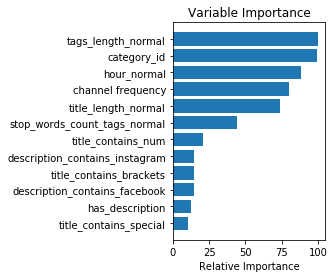
The highest value in the tree leaves follows the decision according to the channel frequency-> video category-> channel frequency -> video category -> publistion hour, which demonstrates the high importance of channel frequency and video category that contribute to the average daily views.

**(4) Gradient Boosting Regressor**

Finally, as an addition method to find the feature importance, the gradient boosting regressor is used to produce a prediction model of an ensemble of weak prediction models. It builds the model in a stage-wise fashion by allowing arbitrary optimization of differentiable loss. The input features and output values are the same as those for the random forest and decision tree methods. Some parameters for the gradient boosting regressor is shown below:

1. Number of estimator: 700 (same as for the decision tree and random forest regressor)
2. Maximum number of tree depth: 5 (same as for the decision tree and random forest regressor)
3. Learning rate: 0.01
4. Loss function: least squares regression
5. Split criterion: MSE

After doing 10-fold cross validation and a final training session for overall data, the result of feature relative importance is shown below:



According to the result, the top five features that has the most relative importance are the same as those for the random forest and decision tree regressor but in different order. However, since gradient boosting regressor catch weak prediction model, the relative importance for each feature is much higher than the results shown in decision tree and random forest.

**(5) Other interesting findings**

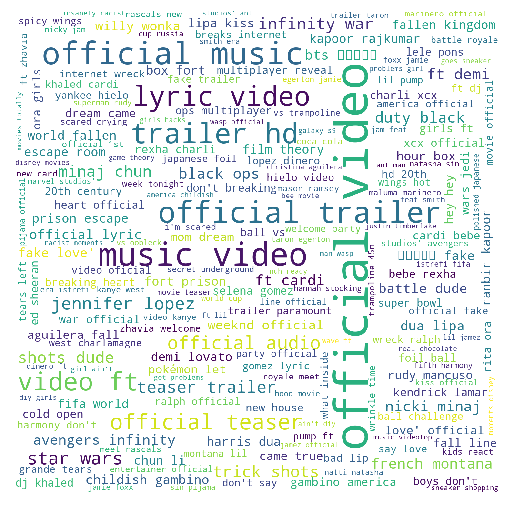
To further analyze the relationship between the pattern of title, tags used and views, we also did a descriptive analysis showing what are the hot words that are related to high views of videos. We achieved this goal by first categorized videos to 3 categories according to the view number. Then we extracted words used in titles and tags of highly viewed videos, then generated word clouds. The results are as follows.

i) Tags used in highly viewed videos



From the above word cloud we can see that the most popular tags in high-viewed videos are “music video”. Other functioning tags include “super bowl”, “dude”, “perfect”, “funny video”, “official video” etc.. We can see that audiences in US region mostly care about music and sports topics. They also click more on funny videos. Videos with verification of “official” institutes are also more welcomed.

ii) Title words used in highly viewed videos



From the above picture we can know that there are much less dominating words used in title in highly viewed videos. This coincides with common sense because tags are often more concise, and people often choose more representative words. However, there is still a pattern that titles with “official video” often get more viewed.

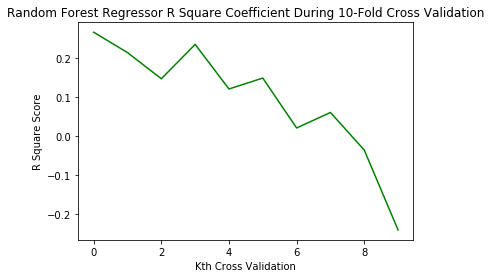
**Performance Evaluation**

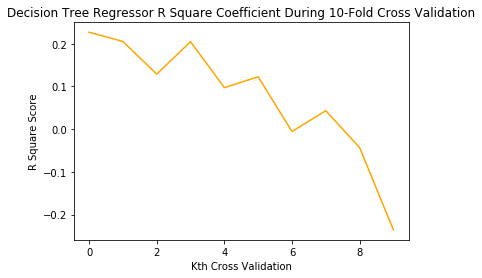
1. **Linear Regression**

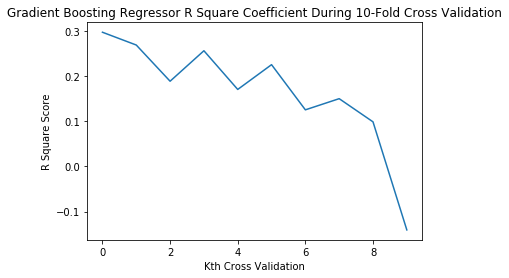
According to the results of the linear regression, although a lot of different methods are tried to improve the performance of the model, the R Square is still only 0.24. However, the result is actually predictable, in common sense, we cannot predict the view of the video just by the features get from the title, tags, and descriptions, etc. without knowing the content of the video.

1. **Random Forest & Decision Tree & Gradient Boosting Regressor**

To check the random forest regressor performance, the R-Squared coefficient for each fold of cross validation is recorded as performance metrics.



****

****

Based on the three R-squared values, all three regression has relatively low R-squared values, which indicates they are not good enough for prediction. However, to the feature importance for weak prediction models, gradient boosting regressor is the most suitable model to use, which has the highest R-squared values.

**Conclusion**

1. **Top features**

According to linear regression model, we found the following attributes have positive effects on views: Channel Frequency, Tag Words Length, Brackets in Title, Numbers in Title, Special Characters in Title, Contains Instagram Links in Descriptions, Category 1 'Film and Animation', Category 10 'Music', Category 20 'Gaming’, Category 22 'People and Blogs', Category 23 'Comedy' and Category 24 'Entertainment’.

Some attributes have negative effect on views. They are: Title Length, Category 15 'Pets and Animals', Category 25 'News and Politics', Category 26 'How to and Style', Category 27 Education' and Category 28 'Science and Technology'.

Other attributes have zero or near zero parameters, meaning they are not relevant to views. They are: Number of Stop Words in Tags, Contains Facebook Links in Descriptions, Category 2 'Cars and Vehicles', Category 17 'Sport', Category 19 'Travel and Events', Category 29 'NonProfits and Activism' and Category 43 'DK'.

According to the random forest, decision trees and gradient boosting models results, besides video contents, there are some other features of video that can affect video views, which could be used for consideration to popularize their Youtube videos.

The top 5 most important features are:

1. channel frequency
2. ideo category
3. title length
4. tag length
5. publish hour

2. **Suggestion for Stakeholders**

The stakeholders for the project is people who want to improve their Youtube video average daily views. After checking the values of each nodes in the decision tree, the top five values from high tree nodes are selected and corresponding original values after inverse-normalization transformation are calculated. Some suggestions are concluded based on linear regression and decision tree:

1. Publish as many as possible videos to arise your channel frequency.
2. Post in the morning, more specifically earlier than 2PM.
3. Use less than 2 stop words in tags.
4. Use less than 10 tags.
5. Use “official video” as tags if possible.
6. Categories of 'Film and Animation', 'Music', 'Gaming’, 'People and Blogs', 'Comedy' and 'Entertainment’ usually get more views, try to post videos in these topics!

**Risk:**

1. The conclusion is drawn by only 40000 samples which not enough.
2. Some of the features such as “category ID” is not complete (only include 16 categories).
3. The data of the video is the “trending video” which is a special part of all of the videos in youtube.
4. Even though we only need the importance of the features instead of the prediction, the prediction results of the model is not so good which may cause less cogency.

**Future Work:**

1. Find more features such as the sequence of words in the description, the attribute of cover pictures to make the model more accuracy and make the further analysis.
2. Get more data of the youtube video from internet.