How Viral will a YouTube video be?

-- STATS 503 Project Report

Chittaranjan Velambur Rajan Jifan Li Mengyao Wang

Abstract

Many social media users in the United states believe that they used Youtube more during the lockdowns. We intend to investigate what factors can lead to more views, explore underlying trends or patterns of the dataset, and predict the number of views a Youtube video will receive. With the dataset collected from the Youtube API, six different methods were applied in this project, including LDA, Multinomial, KNN, Random Forest, GAM and Neutral Network.

Introduction:

While YouTube has been a popular video sharing and social media website since inception, we are particularly interested in analyzing viewership trends of YouTube videos published during the COVID-19 pandemic. Surveys reveal that roughly 64% of people admitted to increased usage of YouTube during lockdowns due to having more free time[1]. But is increased viewership just a consequence of free time or other reasons? In this context, analysis and information about what kinds of videos or which characteristics of videos can attract larger viewership is a useful insight for any content creator to have.

In this report, we explore underlying trends and patterns of the data, and discuss the usage of statistical methods for predicting the number of views a YouTube video will receive, based on factors related to the video's content, more of which are described in the Data section below.

Data:

Data Collection:

As there were no readily available datasets that were suitable for answering the research questions posed in this report, the YouTube Data API was used to curate a custom dataset[2]. The data collection occurs in three stages:

1. Search: In this stage, the API is used to "search" for relevant YouTube videos within these six categories – Music, Comedy, Entertainment, Education, News/Politics, Science/Tech. This set of API calls only returns the IDs of videos matching the search criteria, and is stored as JSON files for the next stage.
2. List: The API provides a "list" functionality using which metadata about a series of videos can be retrieved. This stage uses the video IDs from the search result and fetches metadata for each of the videos. The metadata includes these fields:
    a. Video ID: The ID when appended to a YouTube URL leads to the video
    b. Published At: Timestamp of when the video was released to YouTube
    c. Category: One of the six categories as specified in the Search Query
    d. Time: One of "before", "during", or "after" reasonable time frames between 2019 and 2022 that reflect the severity of the pandemic.
    e. Title: Name of the video as displayed on YouTube.
    f. Duration: ISO 1806 representation of the length of the video
    g. Definition: One of "hd" or "sd" indicating the picture quality of the video
    h. Caption: Boolean indicating whether the video has Closed Captioning
    i. Made For Kids: Boolean indicating whether the video was meant for kids
    j. Views: The number of views the video has
    k. Likes: The number of likes the video has
    l. Comments: The number of comments the video has

3. Process: The list responses are stored as JSON files, and processed by another script that creates a familiar “N x P” representation in a CSV format. This format has the variables described above for each of the videos listed.

Data Processing:

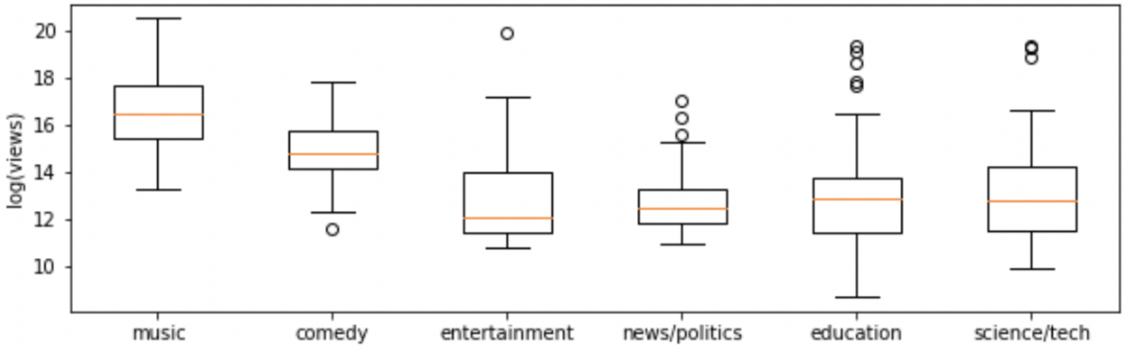
The data is already in a CSV format at this stage, but still requires some preprocessing to enable easier analysis.

- Check for nulls: Some videos do not have all data points associated with them due to special reasons like being marked as “Private” or “Unlisted”. We choose to discard these records. This finally leaves us with 770 rows in the dataset, with each row indicating a unique video.
- Convert to numerical formats: Some fields like “Published At” and “Duration” are represented in alphanumeric ISO format. These are processed into their numerical equivalents as that is more suitable for learning algorithms.
- Log Transformation: As the number of views, comments, and likes, have a large range, a log-transformation is applied to these variables to stabilize their variance.
- Derived Variables: In order to neutralize the differences between older and newer videos (as older videos are likely to have more views due to being available for longer), a new variable called “Age” is introduced to the dataset which indicates how old a video is. And a log-transformation is applied to “Age”. The “Published At” variable is also broken into its component parts like Year, Month, Day, Hour, Minutes, and Second. Finally, there are a total of 13 predictors in the dataset.

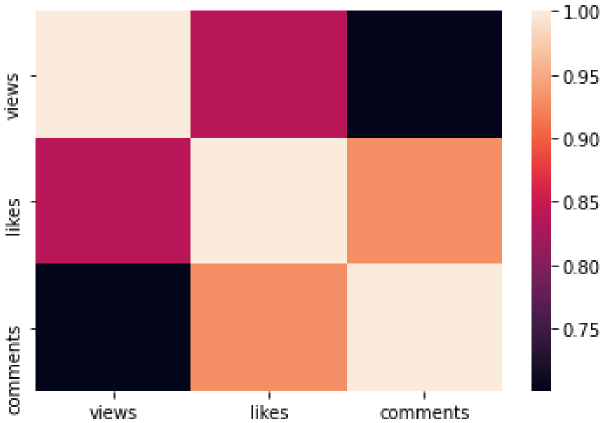
Exploratory Data Analysis:

We present some key findings and results from the Exploratory Data Analysis.

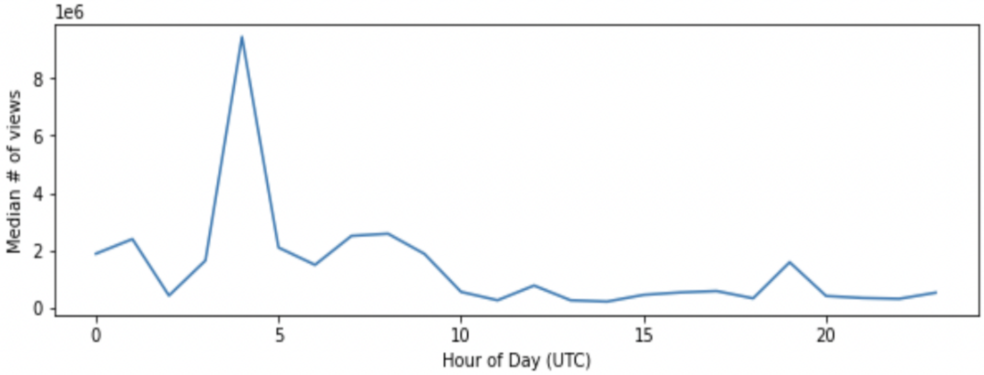
1. **Views across Categories**: The boxplot figure shown below depicts a trend which is unsurprising. Videos from Music and Comedy categories have a larger number of views, and this is likely because videos from these categories are more accessible, thereby garnering views from larger audiences.



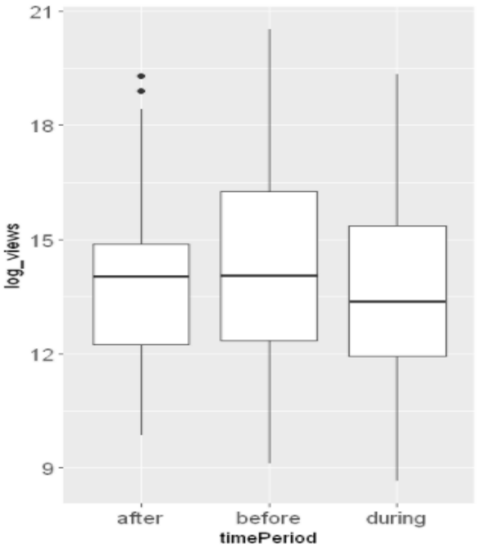
2. **High Correlations**: As the matrix below shows, the correlations between the number of views, comments, and likes was quite high. Due to this, it was deemed reasonable to use just one of these variables both while looking for trends during EDA, as well as during prediction. The number of views was selected as the response variable.



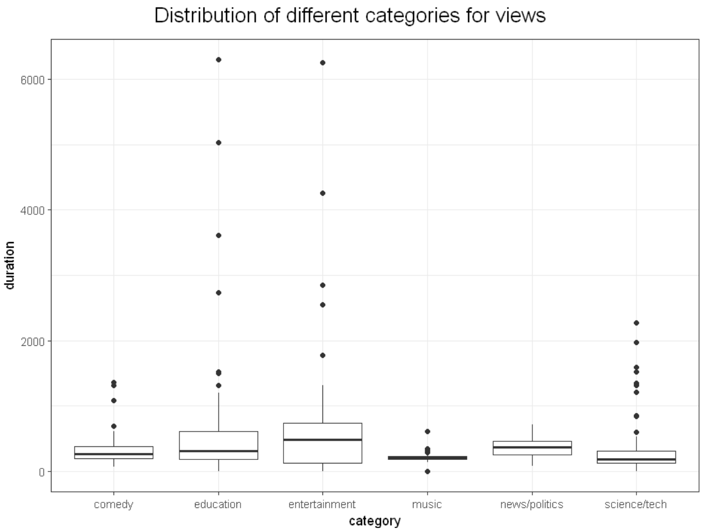
3. **Publish Time of a Video:** A pattern occurs when looking at the publication times of videos, specifically the hour of day they were released at. Both in terms of frequency and median number of views, there was a peak at 04:00 UTC. This corresponds to late evening in the United States, and early afternoon in most parts of Asia. These two geographies are the top two consumers of YouTube content, and it is likely that the spike in viewership comes from the fact that new videos released during "active" hours in these locations are likely to receive higher views.

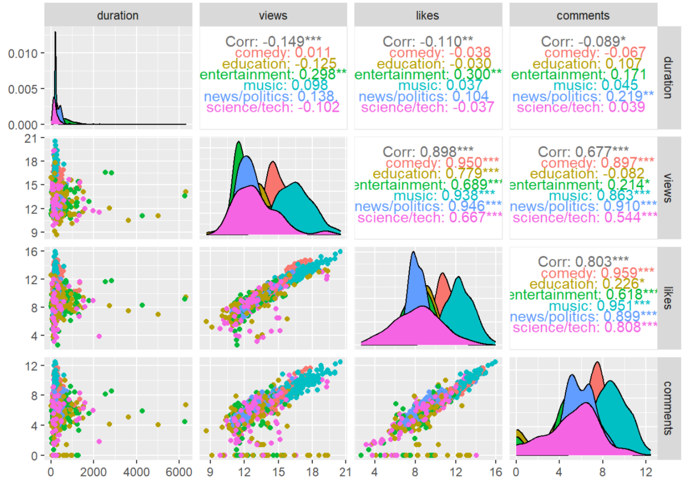


4. **Pandemic Effect:** There was no evidence of a pandemic effect in the data. We expected there to be an increase in the number of views of videos released during the pandemic as compared to those released before or after, but there was a general lack of trend in this comparison. The results are however inconclusive, as this is only a very small sample of all the YouTube videos.

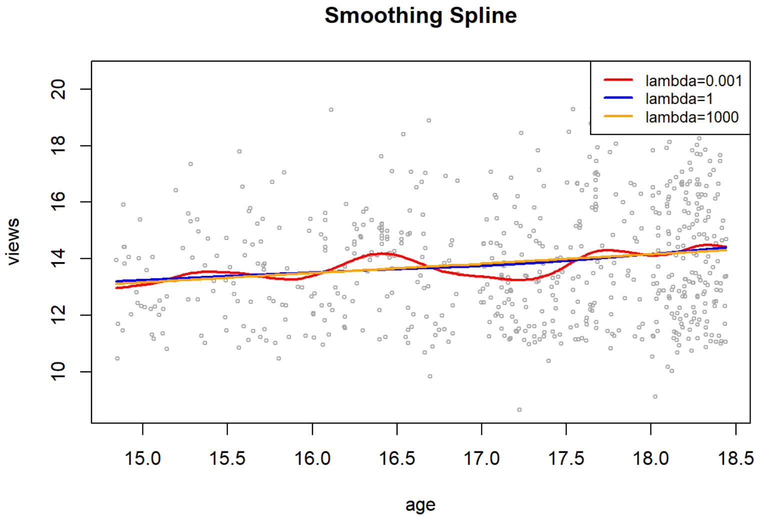


5. **Duration across Categories:** The music category seems to have the smallest durations of all categories. Entertainment and Education categories have a wider range of durations since Entertainment could include TV shows or candid interviews, and education includes tutorials or lectures which are long in length. And in the second plot, there is mostly constant variance across all values of duration, so we expect duration to not have a big impact on the number of views, likes, and comments.





6. **Age and Views:** Using smoothing spline with different values of lambda, we can see older videos, which are available for longer, tend to have more views.



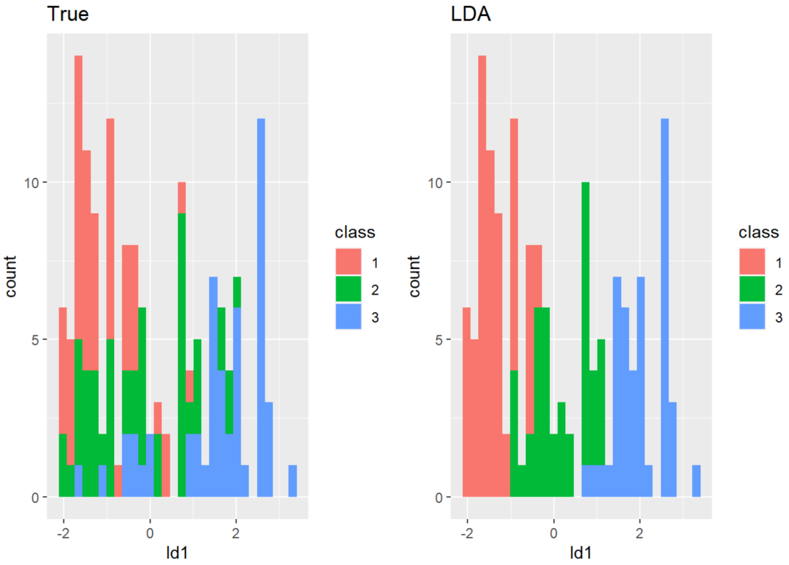
Results:

We split our dataset into train and test set (8:2), and convert views into three different levels by quantile("high" for >66.7%, "medium" for 33.3% to 66.7%, and "low" for <33.3%) to apply the classification methods. We tried many different methods and picked some of them to show below.

Firstly, we intend to try LDA, which is a simple parametric method and has advantages in interpretation.

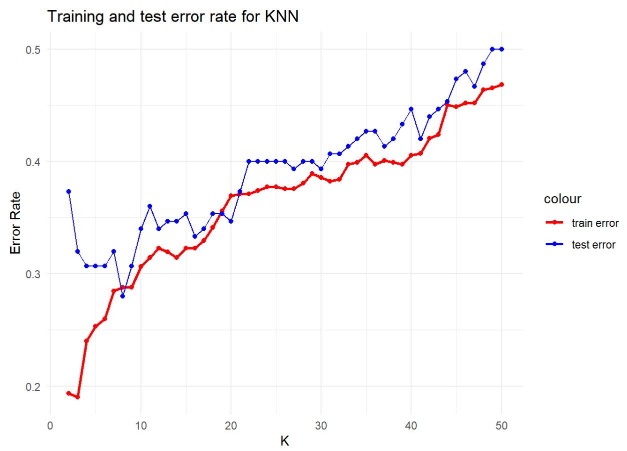
LDA:

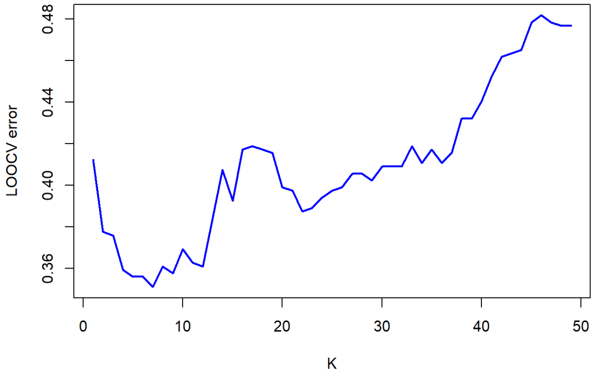
LDA has advantages in interpretation, but may have significant limitations when dealing with complex data. We try LDA first, to see whether our data can be handled by a simple parameter method. The overall test accuracy is 67%, and in the plot we can see class 2, which represents "medium", is mixed with other classes. Then we compute the test accuracy for each class, and find that the accuracy of "low" is 78%, the accuracy of "high" is 76%, but the accuracy of "medium" is only 46%. Based on this result, we think the boundaries might be too complex for LDA, and we decide to try KNN, which can adapt to any boundary shape.



KNN:

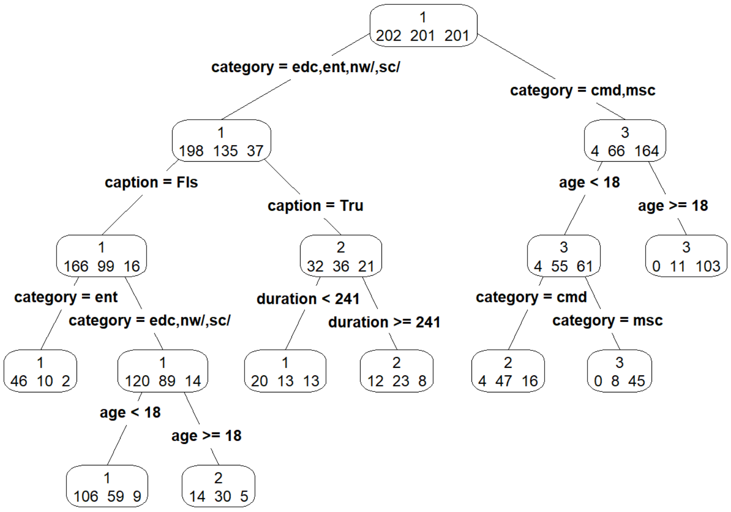
KNN is a fully non-parametric method which will have limitations in interpretation but can adapt to any boundary shape. In the left plot, we can see the test error increases as k increases when k is larger than 20. Then we use LOOCV to investigate further, LOOCV is computationally intensive but has less bias and is more stable than N-fold CV. And in the right plot we can see when K = 7, the model can get the lowest LOOCV error. We know that the smaller k, the more complex the boundaries, and K = 7 is not a large number, which validates that the boundaries are not simple and not easy to fit for LDA. The overall test accuracy of KNN is 69%, and the accuracy of "medium" is 58%, which is much better than LDA.

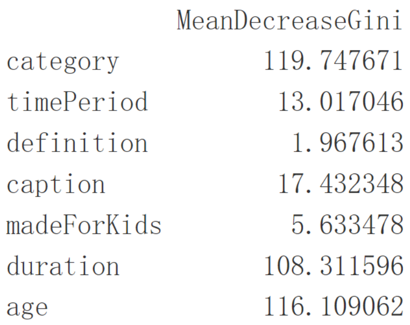




Random Forest:

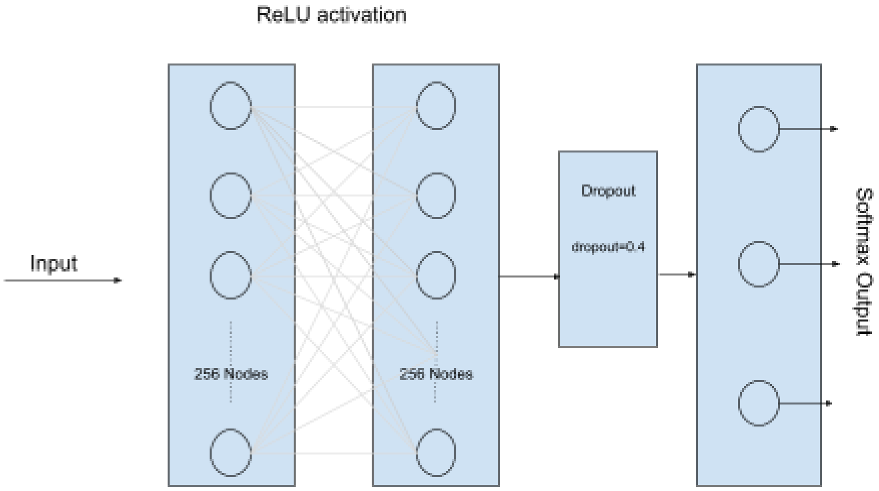
Then we decide to use Random Forest to figure out which variables are more important and see the accuracy of Random Forest. The left plot is a single tree, where 1, 2, and 3 represent “low”, “medium”, and “high”, and we can see how the variables are used for classification. The right plot shows that category, age, and duration are more important than other variables. For category and age, we already know in EDA that they have effects on views. But for time duration, based on the results of EDA, we used to think that it has a small impact on views. The inconsistency with our guess demonstrates the importance of further analysis rather than just guessing. And we find that timePeriod, which represents the pandemic effect, is not as important as age, duration, and category. The accuracy of Random Forest is 67%, which is close to LDA, the accuracy of each class is also close to LDA. This accuracy is reached by setting mtry=3, nodesize=1, ntree=3000. We tried mtry from 3 to 5, ntree from 1000 to 3000, nodesize from 1 to 3. When mtry is 4 and 5, the test error is higher than mtry=3, which may be because of the overfitting.



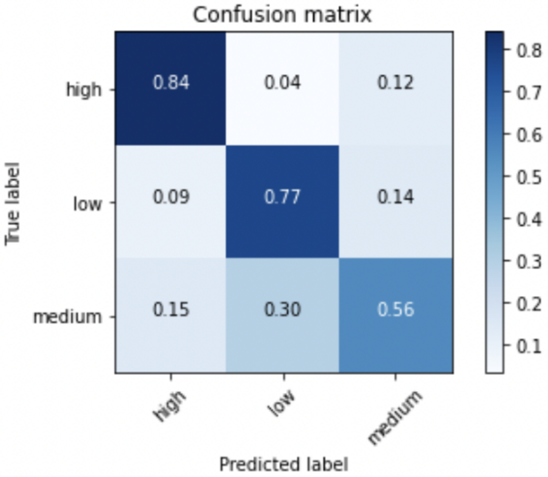

```
             MeanDecreaseGini
category          119.747671
timePeriod         13.017046
definition          1.967613
caption            17.432348
madeForKids         5.633478
duration          108.311596
age               116.109062
```

Neural Network:

The Neural Network (or Multilayer Perceptron) was chosen because it is a versatile method that can work decently well with minimal assumptions. A simple architecture was selected which produced ~72% accuracy on the three-class classification. The architecture was as described in the schematic below:

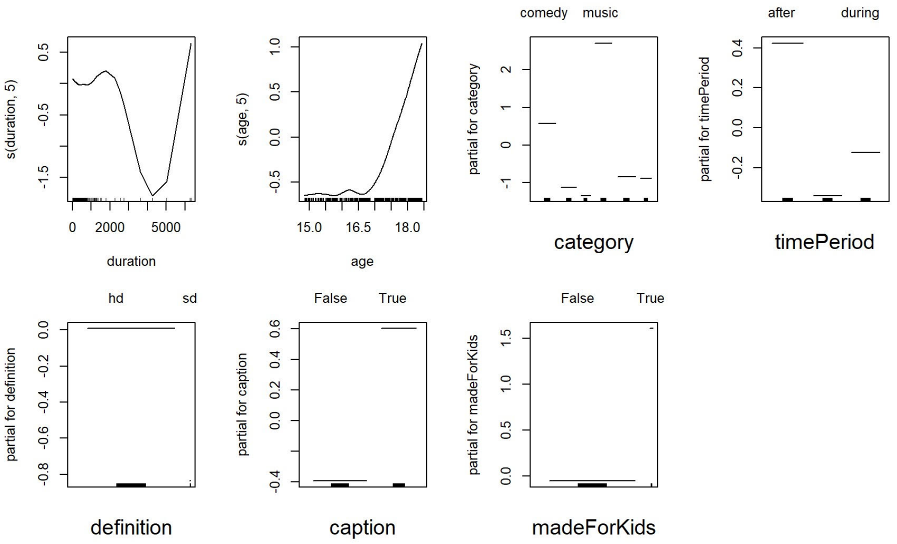


The categorical variables were dummy encoded in order to be interpretable for a neural network. All other numericals variables were standardized to the [0, 1] range so as to avoid issues with any single variable dominating the weights of the neurons. The confusion matrix of the testing error is shown below:



GAM:

Views is originally a numeric variable, we can take advantage of this to apply regression methods on our data and see the impact of each variable on views. Generalized Additive Model, also known as GAM, is a non-linear regression method, which uses smooth non-linear functions of the predictors and adds them together to get the response variable, and we decide to try GAM. We tried several kinds of splines, the plot below uses smoothing splines for duration and age. We can see larger duration will bring more views first, then bring less, and finally bring more views again when it is very large. And we can see the effect of other variables on views below. The videos that have closed captioning, “HD” picture quality, and are made for kids tend to have more views. And we assign the classes to the prediction of GAM to get the classification accuracy, and we find that the test accuracy of the GAMs are only about 64%, which is lower than the other models.



The final results for all attempted models are provided below:

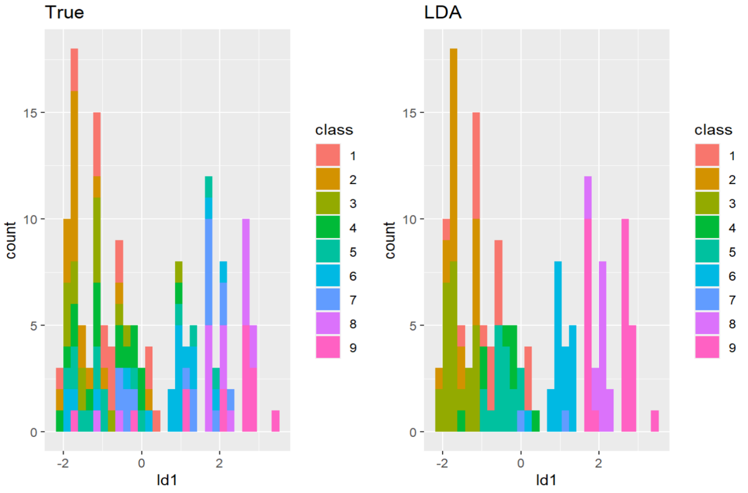
Table 1: Training and Test Error of Views

Model		Training Error				Test Error			
		Overall	Low	Medium	High	Overall	Low	Medium	High
LDA		0.341	0.208	0.552	0.264	0.333	0.22	0.54	0.24
Multinomial		0.343	0.183	0.582	0.264	0.333	0.18	0.56	0.26
KNN		0.286	0.257	0.413	0.189	0.307	0.32	0.42	0.18
Random Forest		0.369	0.322	0.547	0.239	0.327	0.24	0.54	0.20
GAM	Natural Splines	0.353	0.361	0.493	0.204	0.36	0.4	0.5	0.18
	Smoothing Splines	0.356	0.356	0.508	0.204	0.36	0.36	0.54	0.18
	Multiple Splines	0.359	0.396	0.478	0.204	0.36	0.36	0.54	0.18
Neural Network		0.24	0.20	0.37	0.12	0.28	0.23	0.44	0.16

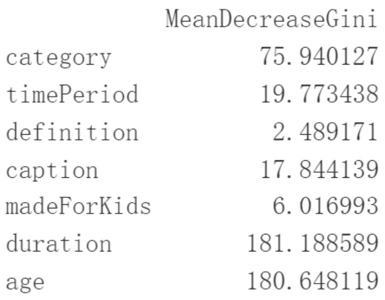
We can see videos with “medium” number of views are the most difficult to deal with. We think it is partly because “medium” has two boundaries, “high” and “low” only has one boundary. Videos which have low views in “medium” might be classified as “low” by our methods, and videos which have high views might be classified as “high”. But for “low”, the highest part of it might be misclassified, but the lowest part of it won’t. For “high”, it is similar.

Classification of Likes and Comments:

In addition to views, we are also interested in likes and number of comments, but it will be unfair to compare the likes or comments of two videos if one has a large number of views and the other doesn't. So for each level of views, we create 3 levels of likes and comments by quantile, get a total of 9 levels, and study if the numbers of views are both high or both low, what factors can make more likes or comments. The following plot uses 9 levels, including views and likes, we can see it is much harder to do classification on 9 levels, the overall accuracy is only 33% for views and likes, and 31% for views and comments. Using LOOCV, we find that when K = 18, the model can get the lowest LOOCV error for views and likes, and when K = 7, it can get the lowest LOOCV error for views and comments.



When we study the likes and comments of videos with the same level of views, duration and age become more important, and category become less important. It seems some categories can attract people to view it, but can't bring much more likes and comments compared with other categories. The plot below is for likes, and it is similar for comments.



```
                MeanDecreaseGini
category               75.940127
timePeriod             19.773438
definition              2.489171
caption                17.844139
madeForKids             6.016993
duration              181.188589
age                   180.648119
```

Table 2: 9 Levels of Views and Likes or Views and Comments

Model	Training Error		Test Error	
	Likes	Comments	Likes	Comments
LDA	0.647	0.68	0.667	0.69
Multinomial	0.631	0.659	0.66	0.74
KNN	0.889	0.889	0.913	0.867

Random Forest	0.704	0.687	0.593	0.76

For 9 levels, we can see the error of KNN is higher than the other methods, it is not surprising since the points with the same level of views tend to be close to each other, and it will make it hard for KNN to classify them further.

Conclusions and Discussion:

KNN with k = 7 and Neutral Network perform best in predicting views. Other models including LDA, Multinomial, GAM and Random Forest get the relatively lower accuracy in our dataset. From the results of Random Forest, we can see that category, duration and age are the three most important variables, and madeForKids and definition are the least important factors. In addition to views, likes and comments can also indicate the popularity of videos. Random Forest performs best for 9 levels of views and likes, and category becomes less important. It seems some categories can attract people to view, but cannot bring much more likes and comments compared with other categories.

We have the most error coming from predicting “medium” level videos in all models. This may be because the “medium” has two boundaries that are hard to classify, while “high” and “low” only have one boundary. Maybe we can change the definition of the levels, for example, define "medium" as 25% to 75%, to get higher overall accuracy and more helpful results for our analysis. However, due to time constraints, we haven't tried it.

And in the future, we could try more regression analysis on views, likes and comments rather than just converting them into levels. And we could also try to predict views for the same video but vary the “Age” predictor incrementally and generate the trend line of cumulative views over time.

Contributions:

The team collaborated synchronously over Zoom calls, and also asynchronously via Slack. The topic for the project was arrived at after collective brainstorming, and further pieces of analysis were done independently and compiled together later.

Chittaranjan Velambur Rajan :

- Data Collection and Cleaning
- Exploratory Data Analysis
- Neural Network
- Presentation

Jifan Li :

- Exploratory Data Analysis
- LDA, Multinomial, KNN, Random Forest, GAM for 3 levels of views
- LDA, Multinomial, KNN, Random Forest for 9 levels of views and likes(comments)

Mengyao Wang:

- Exploratory Data Analysis
- LDA, Multinomial, KNN, Classification Tree for 3 levels of views

References:

[1] Estimated U.S. YouTube usage increase due to coronavirus home isolation 2020. https://www.statista.com/statistics/1106313/youtube-usage-increase-due-to-coronavirus-home-usa/
[2] Youtube API. https://developers.google.com/youtube/v3/docs/videos
[3] STATS 503 Lecture Notes. The models, methods and some codes we used can be found in the lecture notes. And we use some of the descriptions in the lecture notes, such as the advantages of a method.

Code:

Dataset and all source code of this project can be found in the following github link:
https://github.com/chittaranjan19/youtube-video-analytics