Predicting YouTube Trending Videos using Logistic Regression and Decision Tree

Aina Fatin binti Ahmad Fadzil, Yap Wei Li
16028797, 16042145

Abstract

YouTube is a video sharing platform, which allows users to upload, view, rate, share, comment on videos and subscribe to other users. The spectrum of the YouTube library varies from categories such as music, sports, short movies, entertainment, comedy, family, trailers, education and many more. Since YouTube was acquired by Google in 2006, YouTube has become the biggest website for online videos with an amount of over one billion hours of screen time on the site per day. Content creators are able to monetize their Youtube videos through the features such as Advertising revenue, Channel memberships, Merchandise shelf, Super Chat, and Youtube Premium Revenue. To our best knowledge, no prediction model has been performed to predict trending of YouTube videos. Therefore, in this paper, we aim to answer the following problem statement 1) What models best predict the trending of YouTube videos 2) The difference between decision tree and logistic regression based on results and 3) What is the variables that determine a YouTube video to be in the trending list? using logistic regression and decision tree. Logistic regression and decision tree are different in terms of interpretation, variables, readability and predictors and sample size. Logistic regression and decision tree models show the same outcome of the variables that produces the best predictive model. The variable that can determine a YouTube's videos to be in the trending list are views. Based on our results of decision tree, variable such as likes, dislikes and comments are not the determinants of YouTube trending videos.

1. Introduction

YouTube is one of the biggest platforms for video content, hence it is necessary to improve as much as possible in terms of the quality of the proposed content as well as the ability to increase its visibility. One of the main features on YouTube is the possibility to monetize videos in order to earn money on YouTube. As suggested by YouTube Help (n.d.), some of the ways to monetize the videos includes adding advertisements to the videos, channel memberships and merchandise shelf. Subsequently, by applying an in-depth analysis on thousands of videos, we can find several key factors in order to increase views, likes, comments and finally, the income. However, to our best knowledge, no prediction model has been performed to predict trending of YouTube videos. Therefore, in this paper, we aim to answer the following problem statement:

1) What models best predict the trending of YouTube videos?
2) The difference between decision tree and logistic regression based on results.
3) What are the variables that determine a YouTube video to be in the trending list?

We aim to use logistic regression and decision tree to find out models best predictive trending of YouTube videos and perform decision tree to discover the variables that determine a YouTube's videos to be in the trending list.

The scope of work highlighted in this paper is as follows:

1. Relationship between likes, views and comment counts
    a. Analysis of likes
    b. Analysis of views
    c. Analysis of the duration of time between the publish date and trending date
2. Analysis of title and tags using word cloud
    a. Word cloud of video titles according to categories
    b. Word cloud of video tags according to categories

The proposed models for this analysis is logistic regression and decision tree. We used SAS Enterprise Guide ("Graphical User Interface for SAS, SAS Enterprise Guide", 2019) and Enterprise Miner ("Data Mining Software, Model Development and Deployment, SAS Enterprise Miner", 2019) for logistic regression and decision tree respectively. This paper is structured as follows: Section 2 is about literature review, Section 3 discussed about the data, Section 4 is about the Logistic regression, followed by Section 5 that discussed about Decision Tree and lastly, Section 6 includes the comparison of both models, conclusion and implication.

2. Literature Review

<u>What is YouTube?</u>

YouTube is a video sharing platform, which allows users to upload, view, rate, share, comment on videos and subscribe to other users. According to Keller (2018), the spectrum of the YouTube library varies from categories such as music, sports, short movies, entertainment, comedy, family, trailers, education and many more. Since YouTube was acquired by Google in 2006, YouTube has become the biggest website for online videos with an amount of over one billion hours of screen time on the site per day. Additionally, according to Brandwatch (2019), the number of videos adds up to 400 hours of total video time are uploaded on YouTube for every minute. This video sharing platform is available in over 91 countries with a variety of videos available in 80 different Languages, which is equivalent to 95% of the population of the Internet.

Based on the videos that are shared on YouTube, Trending would show the current popular activity on YouTube and in the world to the viewers (YouTube Help, n.d.). Trending would display the same list of videos to all the users of YouTube, which ranges from a predictable video such as a new movie trailer or a recent music video to unpredictable videos such as a viral video. In approximately every 15 minutes, videos may rank higher, lower or stay in the same position in the trending list. As YouTube does not accept payment for a video to be placed on Trending, it is clear that only a certain criterion is chosen for a video to be included in Trending. As users interact and connect through YouTube videos, the three main components that can be analysed based on the dataset are likes, comments and views. As defined by Comnetwork (n.d.), firstly, likes refers to the number of users clicking on the 'thumbs up' icon, indicating that they enjoy the video. Next, comments are feedback from people regarding the video or content that they have watched. Lastly, views are the number of users that have watched the video.

How to Monetize Your Channel?

Content creators are able to monetize their Youtube videos through the following features (YouTube Help, n.d.):

- Advertising revenue: Earn ad revenue from display, overlay, and video ads.
- Channel memberships: Member of your channel make persist monthly payments in exchange for exclusive perks that you offer.
- Merchandise shelf: Your subscribers can browse and purchase official branded merchandise that's demonstrated on your watch pages.
- Super Chat: Your subscribers purchase to get their messages highlighted in chat streams.
- Youtube Premium Revenue: Obtain part of a YouTube Premium subscriber's subscription fee when they watch your videos.

Example of Successful YouTubers:

The example of successful youtuber is Ryan, a 6 year old owner of Youtube account "Ryan's Toy Review'. He earned estimately $11 million (Hogue, 2019). Furthermore, Daniel Middleton, a youtuber who has 18 million subscribers, is estimated to have earned $16.5 million by playing Minecraft in this videos. Moreover, according to O'Kane (2019), two of the highly paid YouTubers are PewDiePie and Logan Paul. Both of them generated as high $14.5 million and $15.5 million respectively over their YouTube content.

However, not all videos are able to get into the trending list; therefore, logistic regression and decision tree are used in our study to predict YouTube trending videos.

3. Data

3.1 Data Collection and Demographics

In order to run the decision tree and logistic regression, the data is collected from two sources which are Kaggle and Youtube. This is due to Kaggle only provide the trending youtube dataset, hence; we decided to manually collect the non-trending youtube videos' data from Youtube. The dataset included 295 non-trending data and 1052 trending data. As YouTube is available worldwide, this paper focuses on the videos that are trending and non-trending in the United States. For the trending data, a random seed is generated to collect the top 20 videos that appear for each category, so that the data is randomized and unbiased. This is step is done to ensure that the ratio of trending videos to non-trending videos is valid to carry out prediction. As for the non-trending data, all the data was collected between the 6th and 7th May of 2019 to ensure that none of the videos were trending during those days. The data consists of 16 attributes about the demographics of Youtube videos, which is demonstrated in Table 1.

Variable Name	Data Type	Roles	Description
video_id	Nominal	Input	The ID of each youtube video, it is extracted from the link of video.
trending_date	Date	Input	The date when the video is trending.
Trending	Binary	Target	It is a derived variable from trending_date. Value 1 indicates trending while value 0 indicates non-trending.
title	Nominal	Input	The title of the youtube video.
channel_title	Nominal	Input	The channel's name of the video.
category_id	Nominal	Input	There are 17 category IDs used in this sample dataset. Each of the category ids and it's description will be provided in table 2.
publish_time	Date	Input	The published date of the video.
tags	Nominal	Input	The tags of the video.

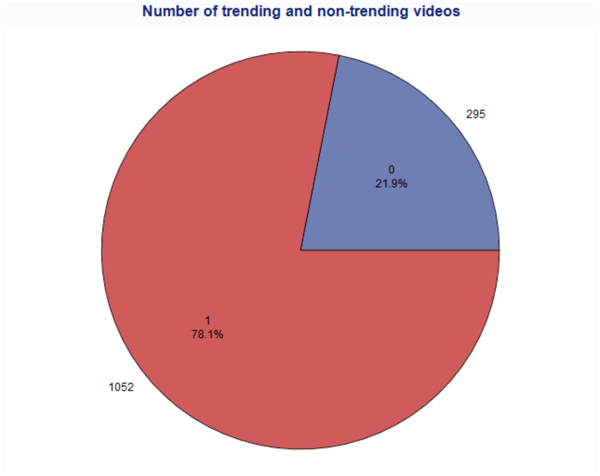
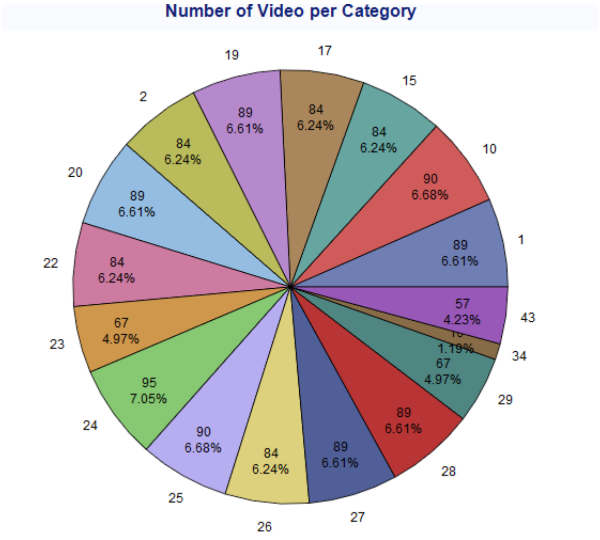
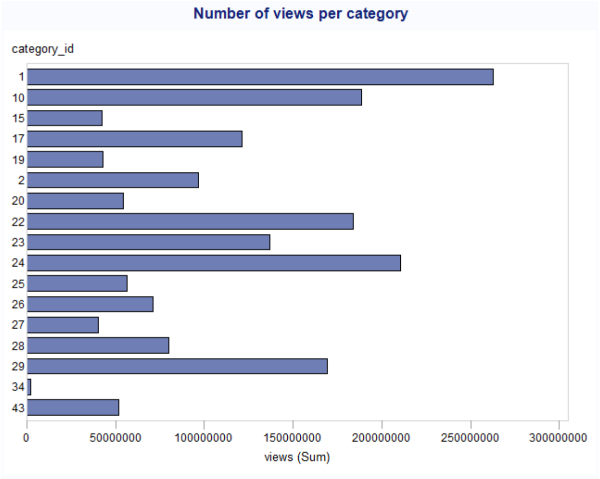
views	Discrete	Input/Target	The number of views a video has.
likes	Discrete	Input	The number of likes a video has.
dislikes	Discrete	Input	The number of dislikes a video has.
comment_count	Discrete	Input/Target	The number of comments a video has.
comment_disabled	Nominal	Input	Setting of the video regarding the comment. FALSE indicates comment is allowed, while TRUE means that comments are now allowed.
ratings_disabled	Nominal	Input	Setting of the video regarding the rating. FALSE indicates rating is allowed, while TRUE means that ratings are now allowed.
video_error_or_removed	Nominal	Input	Is the video removed or not.
Description	Nominal	Input	The description of the video. It is written by the video uploader, the description usually contains what's the video about and social media accounts of the youtuber.
date_diff	Date	Input	calculate the number of days it takes for a video to trend

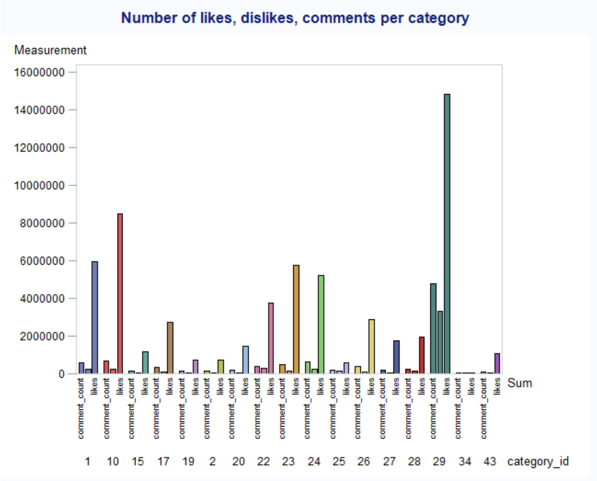
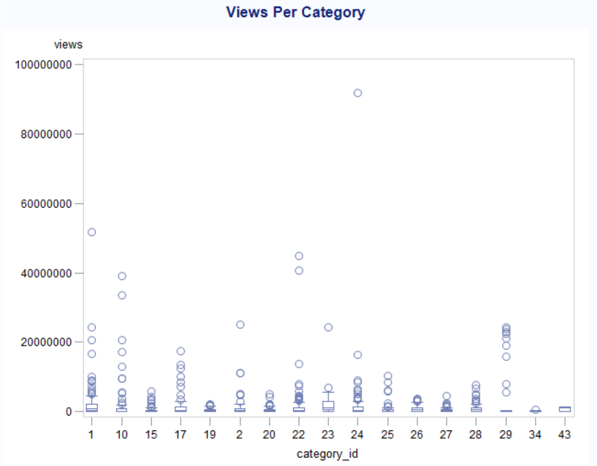
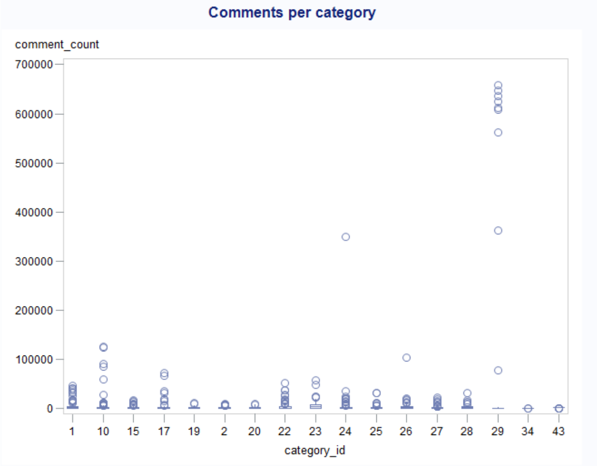
Table 1: The data types, roles and description of each attribute.

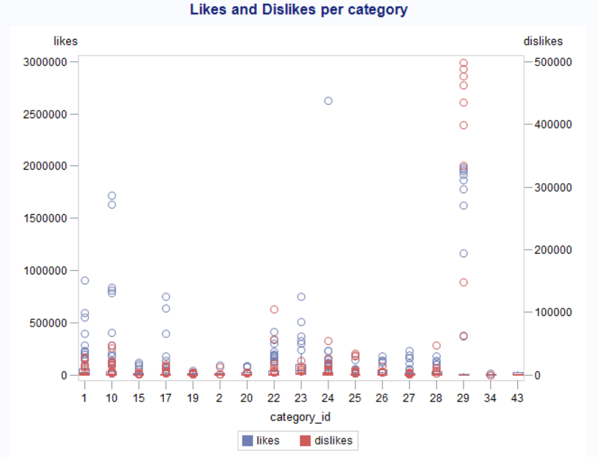
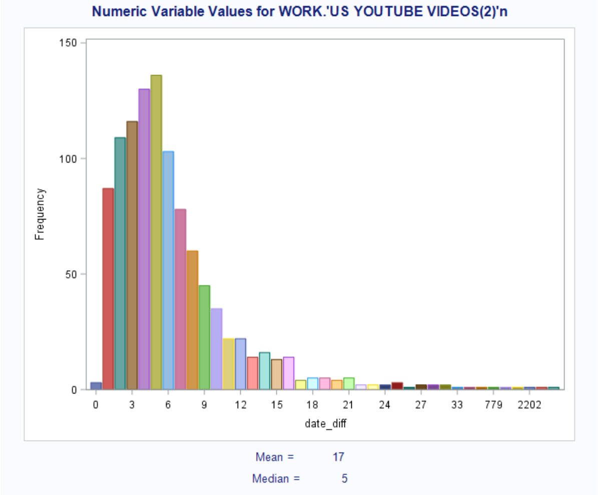
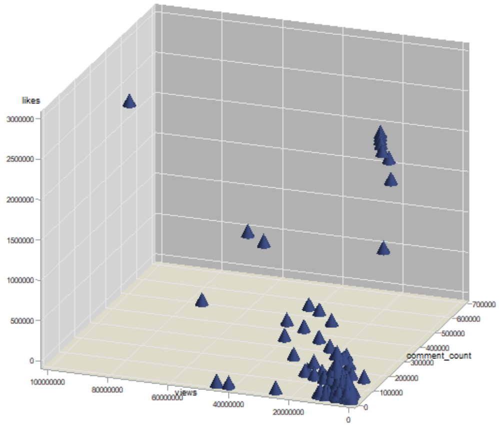
3.2 Data Understanding

Before carrying out predictive analysis, data understanding is required to study the background of data. Table 2 shows a summarization in understanding the data.

title and graph/charts	**description**
number of trending and non-trending videos	Almost one third of the data in this sample is not trending.

 	
number of videos per category  	Number of videos per category are almost even except for the category ID 34 which is comedy.
youtube video views per category  	From this sample dataset, we can observe that category 1 which is Film & Animation has the highest views followed by category 24 (Entertainment) among the other category, while category 34 (Comedy) has the least views.

number of likes, dislikes and comments per category  	The bar chart shows that category 29 which is Nonprofit & Activism has the most likes, dislikes and comments, while category 34 (Comedy) has the least likes, dislikes and comments.
views per category  	All of the categories have outliers except category 43 (shows). Category 24 (entertainment) has the most obvious outlier which is close to 1 hundred million views.
comments per category  	All of the categories have outliers. Category 29 (Nonprofits & Activism) has the most obvious outlier which is close to 700,000 comments.

likes and dislikes per category  	All of the categories have outliers except category 43 (shows). Category 29 (Nonprofits & Activism) has the most obvious outlier for dislikes which is close to 500,000 dislikes while category 24 (entertainment) has the most prominent outlier for likes which is around 2.7 million likes, followed by category 29 (Nonprofits & Activism).
difference of days between trending dates and publishing dates  	The bar chart is skew to the right.The mean is 17 which means that a video would take an average of 17 days to trend after video is published. 50% of the videos take 5 days to trend after video is published. A few videos take 779 days (almost 2 years) to 2202 days (almost 6 years) to trend.
relationship between views, likes, and comment counts  	From this 3D scatter plot, we can observe that the videos with the most likes and comment counts are those with the most views. In fact, likes and comments counts are useful predictors for the number of views of a video.

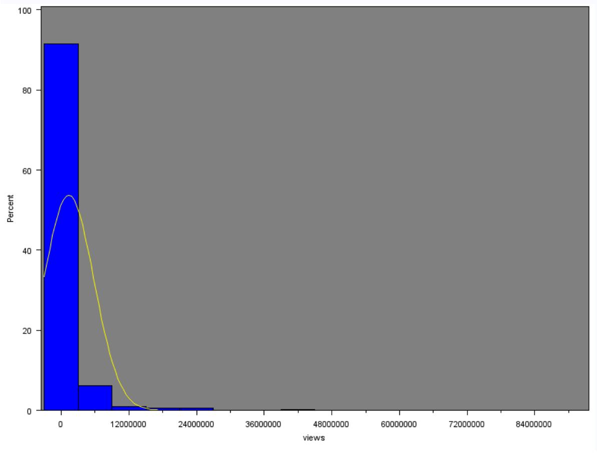
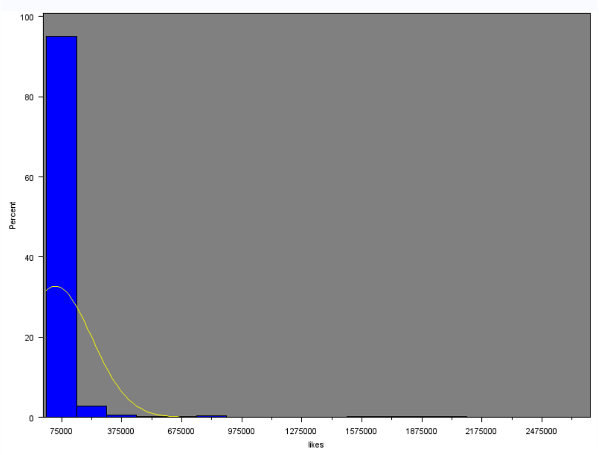
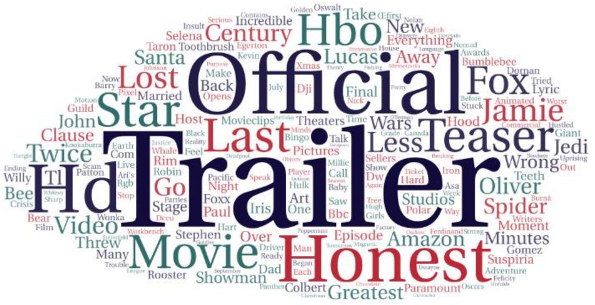
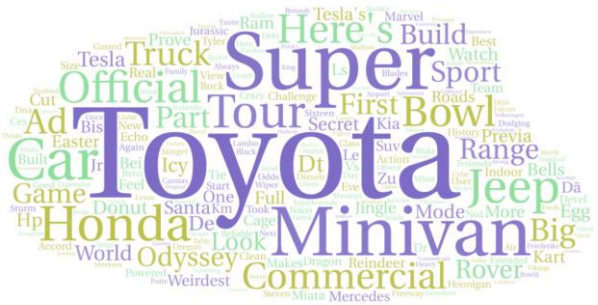
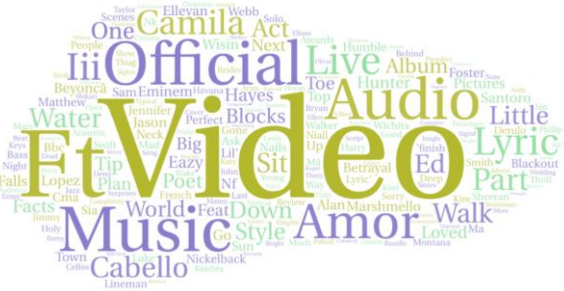
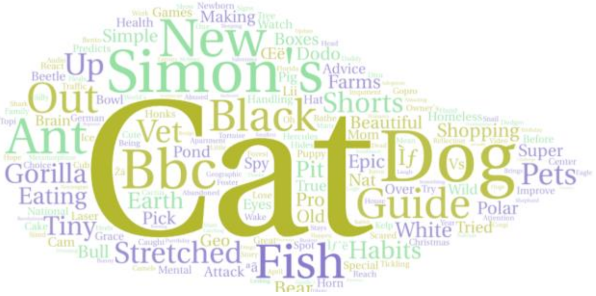
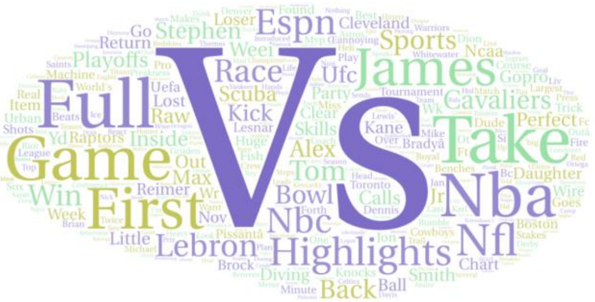
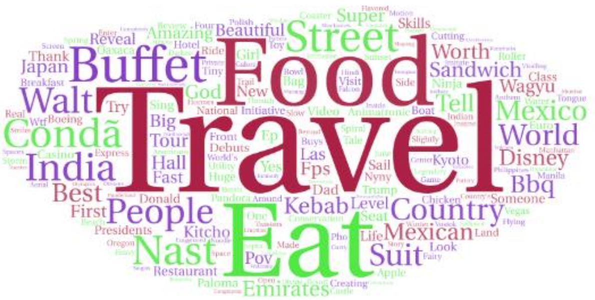
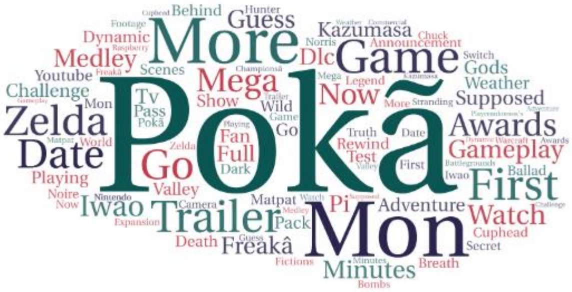
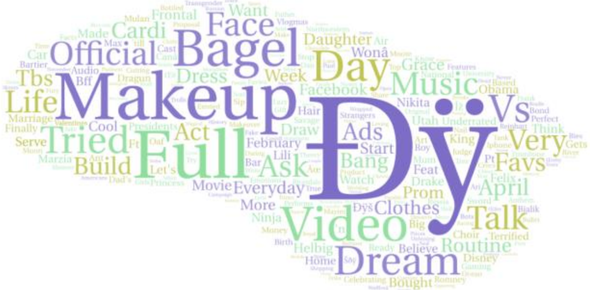
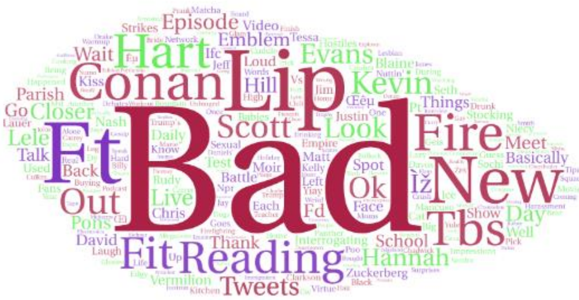
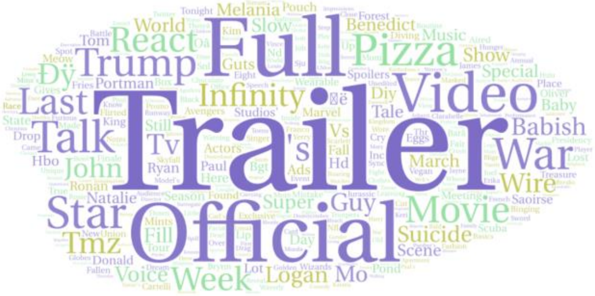
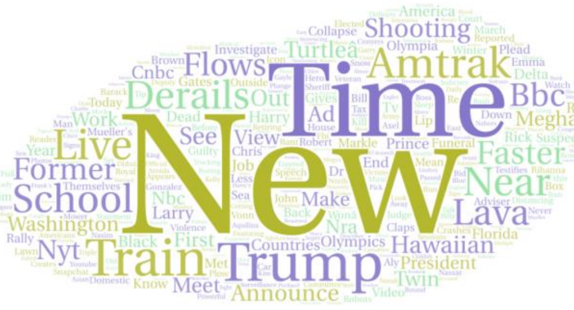
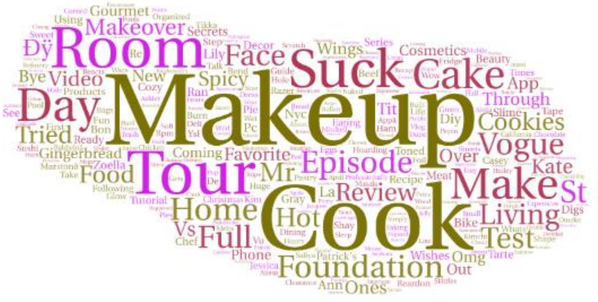
analysis of the views  	The bar chart for the views is skewed to the right.
analysis of the likes  	As for the likes, the bar chart is skewed to the right.

Table 2: Data understanding and description

As for word analysis, it provides us a clear visualisation of the themes of the moment, the types of videos that work by categories. Hence,we extract a set of words that returns frequently by removing as common words (the, to, of, be, and etc.) and stemming. The outcome return a bag of words associated with the category and the text in the form of word cloud. Table 3 provides an analysis of title and section Table 4 provides analysis of tags.

[Category_id 1: Film & Animation]	[Category_id 2: Autos & Vehicles]

 The most commonly used word in the trending videos' title in the sample data set are Trailer, Official, Honest, Movie and HD.	 The most commonly used word in the trending videos' title in the sample data set are Toyota, Tour, Minivan, Super and Car and Honda.
[Category_id 10: Music]  The most commonly used word in the trending videos' title in the sample data set are Video, Official, Music, Audio and Ft.	[category_id 15: Pets and Animals]  The most commonly used word in the trending videos' title for category Pets and Animals in the sample data set are Cat, Simon's, Dog, Fish, Black, Guide and Ant.
[category id 17: Sports]  The most commonly used word in the trending videos' title for category Sports in the sample data set are Vs (versus), Full, Nba, Game, First and Highlights.	[category id 19: Travel & Events]  The most commonly used word in the trending videos' title for category Travel in the sample data set are Travel, Eat, Food, Buffet, Street, People and Country.
[category id 20: Gaming]	[category id 22: People & Blogs]

 The most commonly used word in the trending videos' title for category Gaming in the sample data set are Poka, More, Game, Mon, Go and More.	 The most commonly used word in the trending videos' title for category People & Blogs in the sample data set are Makeup, Dy, Video, Dream, Full, Official, Bagel and Face.
[category id 23: Comedy]  The most commonly used word in the trending videos' title for category Pets and Animals in the sample data set are Bad, Lip, Reading, New, Conan, Fire and Ft.	[category id 24: Entertainment]  The most commonly used word in the trending videos' title for category Pets and Animals in the sample data set are Trailer, Full, Official, Video, Pizza, Trump and Movie.
[category id 25: News and Politics]  The most commonly used word in the trending videos' title for category News and Politics in the sample data set are New, Time, Train, Trump, Live, Amtrak, Near, Lava.	[category id 26: Howto & Style]  The most commonly used word in the trending videos' title for category News and Politics in the sample data set are Makeup, Tour, Suck, Cake, Cook, Day, Room, Episode and Make.
[category id 27: Education]	[category id 28: Science and Technology]

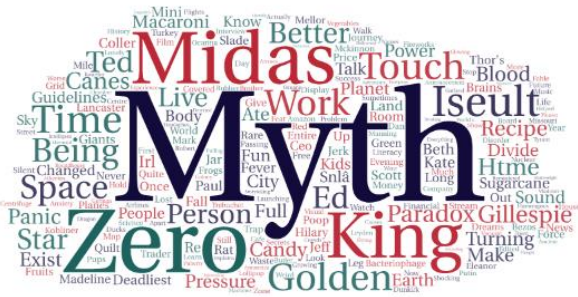
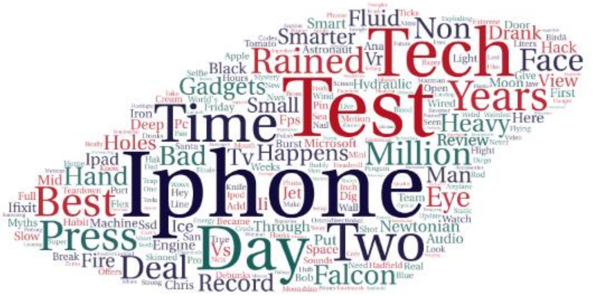
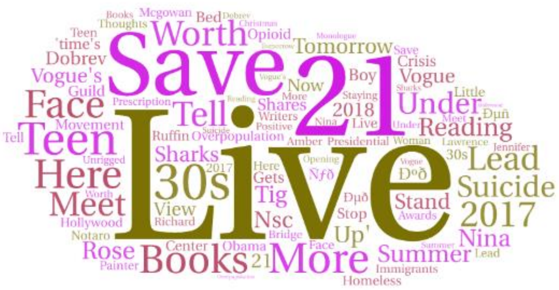
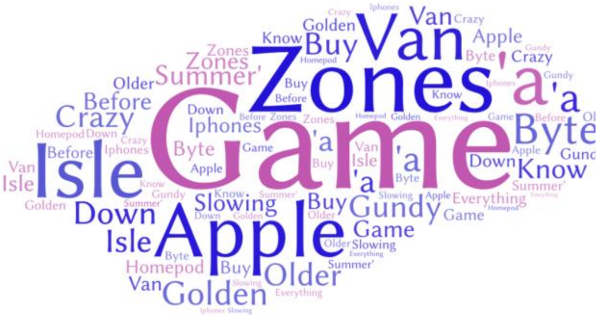
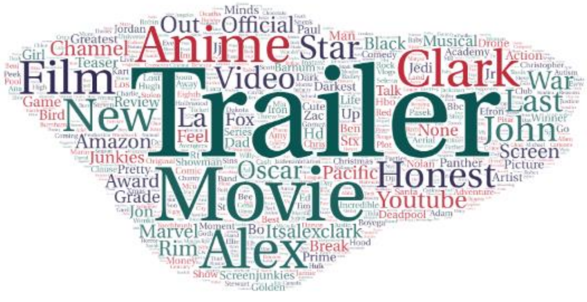
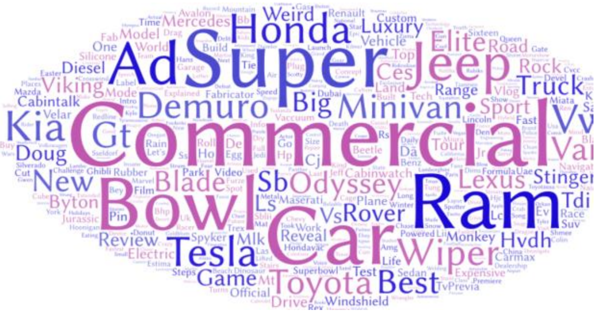
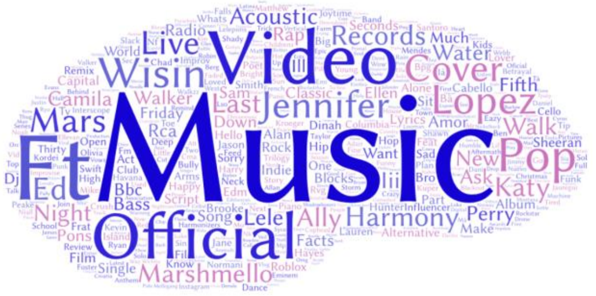
 The most commonly used word in the trending videos' title for category Education in the sample data set are Myth, King, work, Zero, King, Midas and Touch.	 The most commonly used word in the trending videos' title for category Science & Technology in the sample data set are Iphone, Day,Test, Time, Two, Tech and Years.
[category id 29: Nonprofits and Activism]  The most commonly used word in the trending videos' title for category Nonprofits and Activism in the sample data set are Live, Books, Save, More, 21 and Teen.	[category id 43: Shows]  The most commonly used word in the trending videos' title for category Shows in the sample data set are Game, Zones, Apple, Isle and Van.

Table 3: Analysis of Title

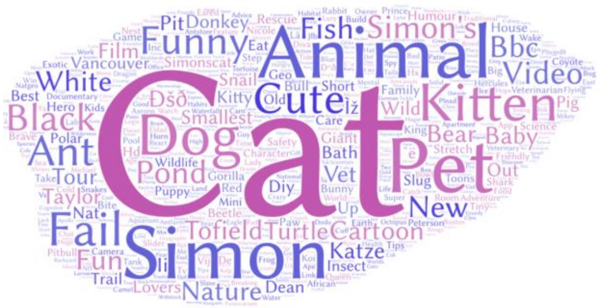
[category_id 1: Film & Animation]  The most commonly used words in the trending video tags for category Film & Animation in the sample data set are Trailer, Movie, Clark, Anime, Alex and Film.	[category id 2: Autos & Vehicles]  The most commonly used words in the trending video tags in the sample data set are Commercial, Bowl, Ram, Car, Super, Ad, Tesla and Honda.

Category id 10: Music



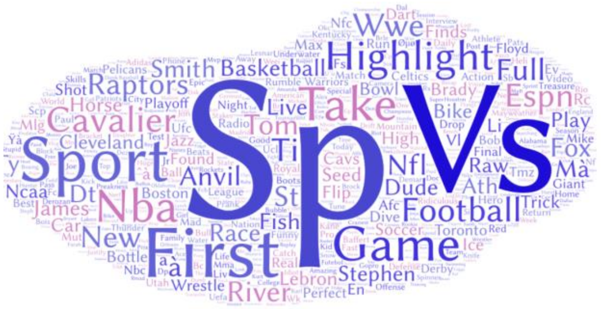
The most commonly used words in the trending video tags for category Music in the sample data set are Music, Video, Official, Ft, Pop, Lopez and Mars

[category id 15: Pets & Animals]



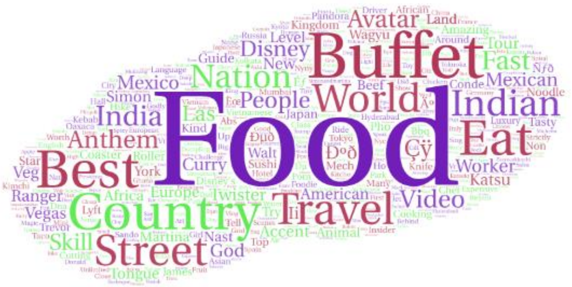
The most commonly used words in the trending video tags for category Pets & Animals in the sample data set are Cat, Dog, Pet, Animal, Kitten, Simon, Fail and Funny.

[category id 17: Sports]



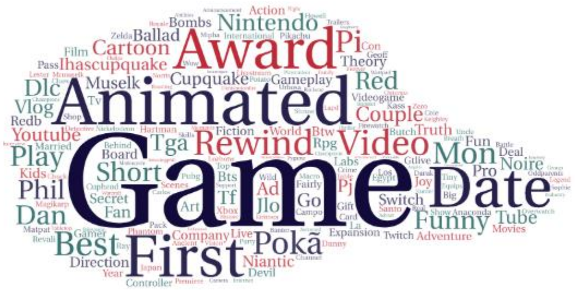
The most commonly used words in the trending video tags for category Sports in the sample data set are Sp, Vs, First, Sport, Game, Football, Highlight and Take.

[category id 19: Travel & Events]



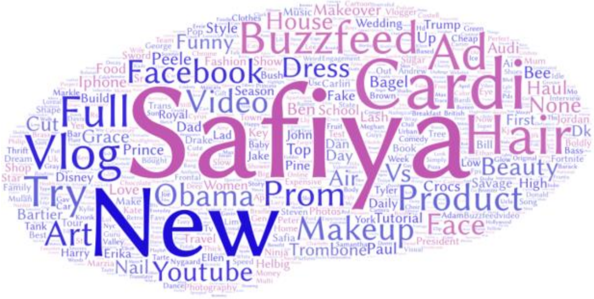
The most commonly used words in the trending video tags for category Travel & Events in the sample data set are Food, Buffet, Best, Country, Travel, Street, World and Eat.

[category id 20: Gaming]



The most commonly used words in the trending video tags for category Gaming in the sample data set are Game, Animated. Award, First, Date, Video and Rewind.

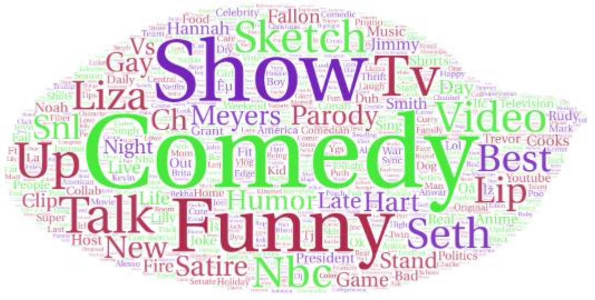
[category id 22: People & Blogs]



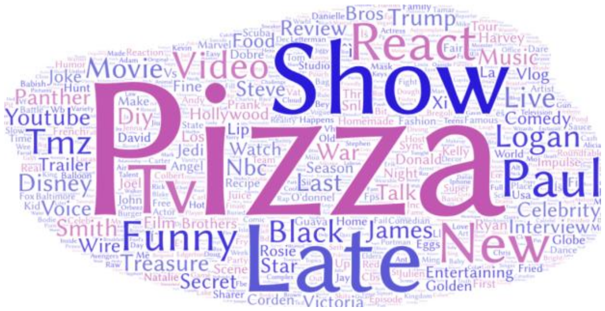
The most commonly used words in the trending video tags for category People & Blogs in the sample data set are Safiya, New, Vlog, Full, Ad, Buzzfeed and Facebook.

[category id 23: Comedy]

[category id 24: Entertainment]

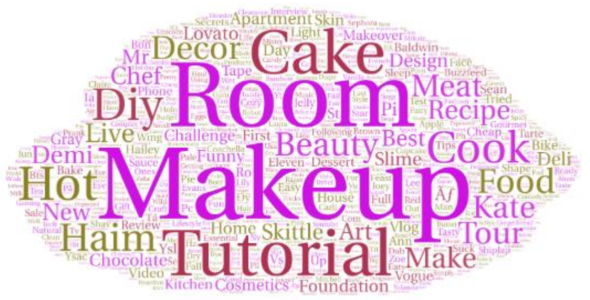


The most commonly used words in the trending video tags for category Comedy in the sample data set are Comedy, Funny, Humor, Show, Tv, Sketch and Up.



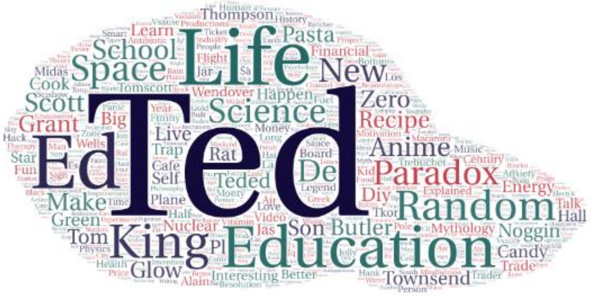
The most commonly used words in the trending video tags for category Entertainment in the sample data set are Pizza, Show, Late, React, Paul, Funny, New and Video.

[category id 26: Howto & Style]



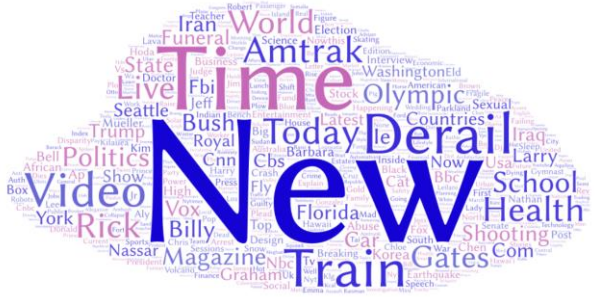
The most commonly used words in the trending video tags for category Howto & Style in the sample data set are Makeup, Room, Tutorial, Cake, Hot, Diy and Cook.

[category id 27: Education]



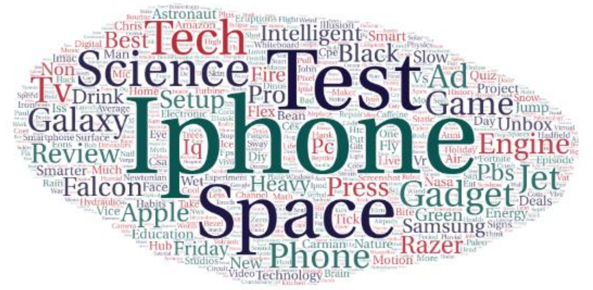
The most commonly used words in the trending video tags for category Education in the sample data set are Ted, Ed, Life, Education, King, Random, Paradox and Space.

[category id 25: News & Politics]



The most commonly used words in the trending video tags for category News & Politics in the sample data set are New, Time, Train, Derail, Amtrak, World, Health and School.

[category id 28: Science & Technology]



The most commonly used words in the trending video tags for category Science & Technology in the sample data set are Iphone, Test, Space, Science, Tech, Gadget and Galaxy

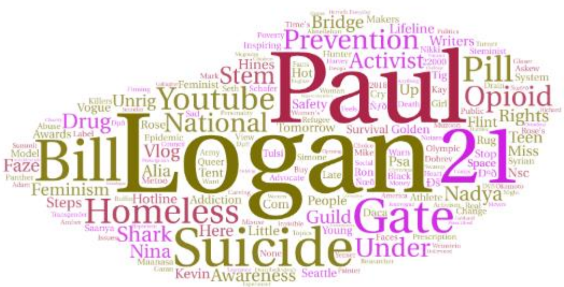
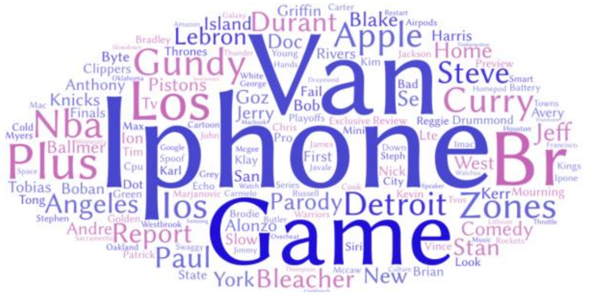
[category id 29: Nonprofits & Activism]  The most commonly used words in the trending video tags for category Nonprofits & Activism in the sample data set are Logan, Paul, 21, Suicide, Gate, Bill, Homeless and Opioid.	[category id 43: Shows]  The most commonly used words in the trending video tags for category Shows in the sample data set are Iphone, Van, Game, Br, Plus, Nba, Los, Gundy and Detroit.

Table 4: Analysis of tags

3.3 Data Quality

Most of the variables (comments_disabled, ratings_disabled, dislikes, likes, views and publish_time) have 1 missing values, while variable comment_count has 8 missing values. The higher number of missing values for comment_count is due to the variable comments_disabled, where the value is 'TRUE'. Thus, comment_count is considered as a missing value as there are no comments.

3.4 Data Preparation

In order to differentiate non-trending and trending data, we decided to add an attribute called trending. Therefore, a non-trending video will have an output of 0, and a trending video will have an output of 1. Besides that, we added an attribute, date_diff, to calculate the number of days it takes for a video to trend. Prior to this step, due to the different format of the dates, we standardized the date into the same format, which is the YYYY-MM-DD format. We also look into the category_id attribute to make sure the category ID exists. As category ID such as trailers and movies are identified differently, they are categorized under category entertainment. Hence, videos under category ID 18 is changed to category ID 24 (Entertainment). Following Table 5 shows the category id and it's name.

Category ID	Category Name
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets and Animals

17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News and Politics
26	Howto & Style
27	Education
28	Science and Technology
29	Nonprofits & Activism
43	Shows

Table 5: Category ID and corresponding Category Name

4 Logistic Regression

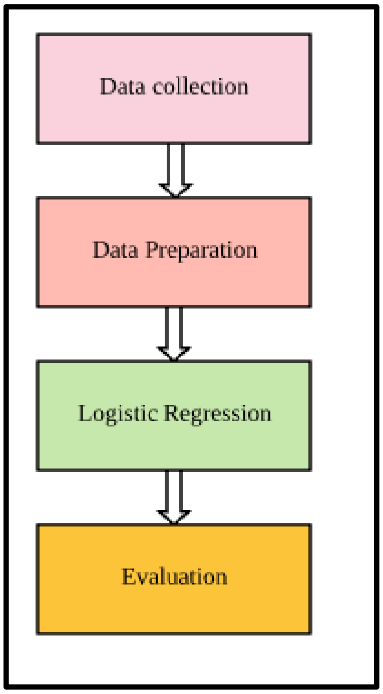


Diagram 1: Steps for Logistic Regression

4.1 Logistic Regression Overview

Logistic regression is a common model used to make dichotomous predictions and has proven to be a successful algorithm (Nie, Rowe, Zhang, Tian & Shi, 2011). Some applications that have utilized logistic regression includes the field of agriculture, credit scoring and accident analysis and prevention. The form of logistic regression model is known as the logit transformation, which produces a linear function of the parameters β_0 and β_1 in a logit equation as follows:

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1$$

According to Nie, Rowe, Zhang, Tian and Shi (2011), the estimation of parameters of the logistic regression model is determined by the least squares function, known as maximum likelihood. Thus, a number of models with varying variables combinations are built in order to find the predictive ability of combinations of different variables. In this paper, we have adopted such method to find the model with successful predictors.

In this paper, we have also demonstrated the use of logistic regression based on the knowledge discovery rather than applying a new approach. SAS Enterprise Guide has been utilized in order to carry out the analysis for the logistic regression model.

4.2 Modelling and Results

We build models with different variable combinations to find the power of different variables. We used a full model fitted to select the variables during the process of model building. Hence, all the variables are included in the building the model. Since there are 17 variables, we have categorised them accordingly to fit the model as shown in Table 6 below. Model 1 consists of variables that is driven by a viewer's watching experience to observe if a viewer's experience would affect the video trend. Model 2 is built to find out does the variables that are related to the settings of the video affect the video trend. Model 3 is built based on variables of Model 1 and 2 to see if both of the conditions would affect if the video is trending or not trending. Model 4 is built using the variable category_id to find out does the category of the video uploaded affect the trending of video.

	Model 1	**Model 2**	**Model 3**	**Model 4**
Variables	views likes dislikes comment_count	comments_disabled ratings_disabled video_error_or_removed	views likes dislikes comment_count comments_disabled ratings_disabled video_error_or_removed	category_id
Description	Attributes related to the video	Features of the video	Combination of Model 1 and Model 2	Categories of the video

Table 6: The variables included in different models

The results of the model is displayed in Table 7.

	Model 1	Model 2	Model 3	Category_id	Model 4
Dependent variable: Trending Intercept	-0.4002	5.5751	4.6552		-1.7745
X1 video_id				1	0.6609
X2 trending_date				2	0.4031
X3 trending				10	0.7053
X4 title				15	0.4031
X5 channel_title				17	0.4031
X6 category_id				19	0.6609
X7 publish_time				20	0.6609
X8 date_diff				22	0.4031
X9 tags				23	-11.8968
X10 views	-2.44E-7		-1.5E-7	24	0.9021
X11 likes	0.000027		0.000034	25	0.7053
X12 dislikes	-0.00004		0.000026	26	0.4031
X13 comment_count	-0.00098		-0.00117	27	0.6609
X14				28	0.6609
X15 comments_disabled		-0.0351	1.5059	29	0.0341
X16 ratings_disabled		-0.0139	-0.5590	43	16.1269
X17 video_error		-6.8067	-5.9442		
X18 description					

AIC	1393.62	1415.05	1393.62		1418.09
SC	1398.82	1420.26	1398.82		1423.30
-2 Log L	1391.62	1413.05	1391.62		1416.09
Deviance	1155.38	10.65	1131.94		0.0003
Pearson	9.0501E13	10.07	4.08933E13		0.0002
HL	434.13 (<0.0001)	0	650.35 (<0.0001)		0.0001 (1.000)

*HL Hosmer and Lemeshow Goodness-of-Fit

*Significance at the 5% level

Table 7: Logistic regression based prediction model

The measures related to the fit of the model are listed at the bottom of the model result table. All the measures show that Model 3 fits the data best. Although Model 1 has similar results () to Model 3, the variables in Model 1 are also in Model 3, which makes Model 3 the overall best model.

4.3 Model Performance and Interpretation

The four correlation indices, which are Somers' D, Gamma, Tau-a and c are computed as a higher value for these indices would indicate a better predictive ability. Based on the generated results, Somer's D, Gamma and Tau-a values indicate that the indices have a better predictive ability. The c, which is concordance, estimates the probability of an observation with an outcome having a higher predicting probability than an observation without the outcome. The value of c, which is close to 1, indicates that there is a strong predictive ability to indirectly affect the trending. The following Table 8 shows the model performance in terms of the predictive power analysis.

	Model 1	**Model 2**	**Model 3**	**Model 4**
Somer's D	0.733	0.014	0.757	0.250
Gamma	0.733	0.200	0.757	0.328
Tau-a	0.247	0.005	0.255	0.085
c	0.866	0.507	0.878	0.625

Table 8: Prediction performance of logistic regression based models

Therefore, this shows that Model 3 is the overall best model as it has the best predictive power. The value of Model 3 is the closest to 1, especially based on c, which is the concordance (0.878). Thus, Model 3 has a strong predictive ability to predict the videos to trend.

5. Decision Tree

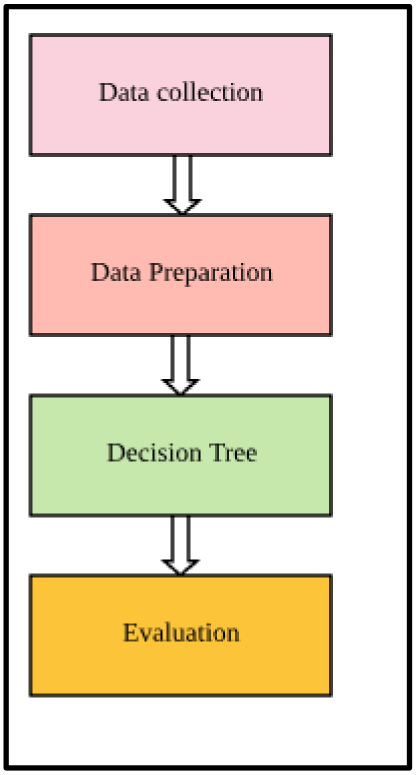


Diagram 2: Step for Decision Tree

5.1 Decision Tree Algorithm

Decision tree is a widely used model and is applied to solve the real-world problems (MRI, credit card churn). Decision tree is a symbolic learning technique that structures information collected from a training dataset in a hierarchical's structure made up of nodes and complications (Nie, Rowe, Zhang, Tian & Shi, 2011). It is easy to understand the results of decision tree due to the structure of the tree (credit card churn). At each branch of the tree, considering the some particular criteria, explanatory variable that has the highest association with the response variable is selected by the decision tree algorithm. The most important variable is known as the root node (metabolic syndrome) . Besides that, the decision tree can be applied to datasets with numerical and categorical data (Nie, Rowe, Zhang, Tian & Shi, 2011). There are 2 types of decision trees which are classification trees and regression trees. The response variable gets its value from a distinct domain, and each class of a leaf node is associated with a probability (Nie, Rowe, Zhang, Tian & Shi, 2011). Decision trees have been widely used in many industries such as bank industry to predict credit card churn (credit card churn), medical industry (metabolic syndrome, MRI) (Nie, Rowe, Zhang, Tian & Shi, 2011). To the best of our knowledge, there is no application to the youtube video trending prediction. This paper does not intend to enhance the existing decision tree algorithms. We use SAS Enterprise Guide to perform Decision Tree.

5.2 Modelling and Results

Similar to the logistic regression model, we build models with different variable combinations to find the power of different variables. The model is split to train with 70% of the data and validate with 30% of the data. Since there are 17 variables, we have categorised them accordingly to fit the model as shown in Table 9 below. Model 1 consists of variables that are driven by a viewer's watching experience to observe if a viewer's experience would affect the video trend. Model 2 is built to find out does the variables that are related to the settings of the video affect the video trend. Model 3 is built based on variables of Model 1 and 2 to see if both of the conditions would affect if the video is trending or not trending.

	Model 1	**Model 2**	**Model 3**
Variables	views likes dislikes comment_count	comments_disabled ratings_disabled video_error_or_removed	views likes dislikes comment_count comments_disabled ratings_disabled video_error_or_removed
Description	Attributes related to the video	Features of the video	Combination of Model 1 and Model 2

Table 9: The variables included in different models

The results of the model is displayed in Table 10.

Model Performance	**Model 1**	**Model 2**	**Model 3**
Average Squared Error (ASE)	0.09	0.1519	0.0831

Table 10: Decision Tree based model's prediction performance.

Based on ASE results of 3 models, the figures are very close to 0 which indicates that we have found a good fitting model to predict Youtube Trending Videos. However, Model 3 (containing attributes related to the video and features of the video) is the best decision tree model because it has the lowest ASE.

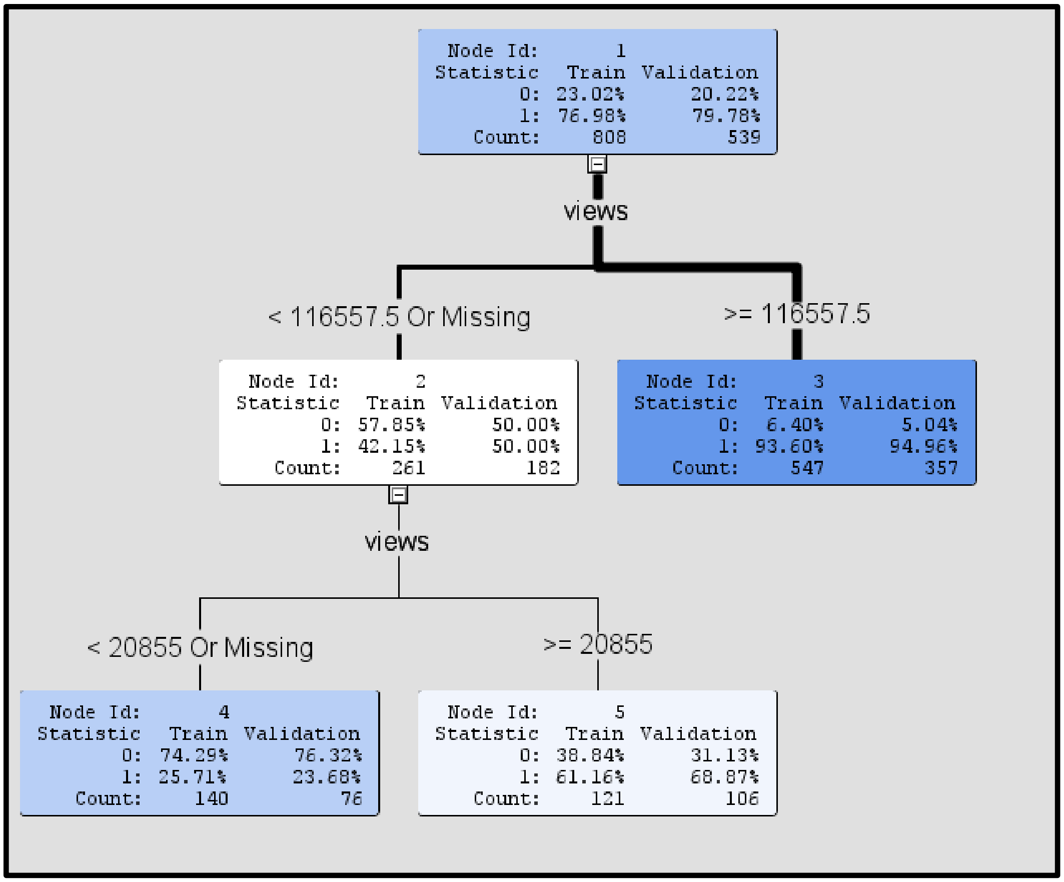


Diagram 3: Decision Tree of Model 3.

Based on diagram 3, the best node for trending=1(trending) is node 3: If views are more than 116558, approximately 94% a video will be trending. On the other side, the best node for trending=0 (non-trending) is node 4, if a video's views are less than 20855 or missing, approximately 75% it will not be trending.

6. Conclusion

6.1 Model Comparison

Although the results of Logistic Regression and Decision Tree are similar, we also found that both of the models have different algorithms in order to carry out their subsequent tasks. The following Table 11 describes the differences between both of these models:

Difference	Logistic Regression	Decision Tree
Interpretation	We find that Logistic Regression is harder to interpret as it involves a lot of measures and indicators as to how successful is a predictive power of a model.	We find that Decision Tree is easier to interpret as the measures and indicators of a successful prediction is supported by a tree diagram. Thus, by just looking at the

	Therefore, without a good statistical knowledge, one would not be able to interpret and understand the outputs of the model.	decision tree diagram, one can interpret the variables involved to determine the best nodes for the best prediction.
Variables	In this case, we find that we are unable to know the predictive power of the variables individually. Thus, we are required to run a hypothesis testing to test the regression coefficients based on the measures of Likelihood Ratio, Score and Wald to find if the variable is effective in the predictive model.	In this scenario, we find that we are able to find the predictive ability of each individual variable. Unlike logistic regression, no specific tests or measures are required in order to determine if the variable is effective or not effective in the predictive model. Instead, the model would use only the effective variables.
Readability	In logistic regression, the model uses an equation-like structure between the independent variables in respect to its dependent variable. Hence, it is harder for one to interpret the values based on the equation in regards to building the predictive model.	In the decision tree model, the tree classifiers produce rules that are like simple English sentences. Therefore, even someone with an analytical background would be able to interpret the variables used in building the predictive model.
Predictors and sample size	For the logistic regression model, this model examines all the simultaneous effects for all the predictors. Regardless of the sample size, the data will be sufficient to analyze all the relevant predictors.	For decision trees, every time the tree splits the data with a predictor, the remaining sample size reduces. It will then not be able to obtain enough data in order to identify further predictors although the predictors may be relevant. Thus, decision trees may be less accurate for a small sample size as not all the predictors are analyzed.

Table 11: Differences between Logistic Regression and Decision Tree

6.2 Conclusion and Implication

Therefore, logistic regression and decision tree models show the same outcome of the variables that produces the best predictive model, indicating that both models are as effective in predicting if a YouTube video will trend or remain as a non-trending video. Furthermore, the variable that can determine a YouTube's videos to be in the trending list are views. Based on our results of decision tree, variable such as likes, dislikes and comments are not the determinants of YouTube trending videos.

There are several implications which can be improved for further research. Firstly, as this analysis only consists of trending and non-trending YouTube videos from the United States, a further analysis can be done including all the countries with access to YouTube. This would be able to provide a better prediction of the videos to trend as it is based on the whole YouTube platform. Furthermore, the method of extraction of non-trending video should be improved. Reason being, the day that the data was collected may not show that the video is trending, however it may trend later on. Lastly, a collection of data across a long period of time may be more sufficient. According to Cooper (2019), over the years, YouTube has constantly changed their algorithm in pushing videos to guide people's behaviour. Thus, the viewers behaviour is constantly changing, which will affect how videos trend over the years as well as the importance for YouTube to engage with their viewers in order to satisfy them.

7. References

Brandwatch. (2019). 48 Fascinating and Incredible YouTube Statistics. Retrieved from https://www.brandwatch.com/blog/youtube-stats/

Comnetwork. (n.d.). A Guide to Measuring Impact on YouTube. Retrieved from https://storytelling.comnetwork.org/explore/19/a-guide-to-measuring-impact-on-youtube

Cooper, P. (2019). How Does the YouTube Algorithm Work? A Guide to Getting More Views. Retrieved from https://blog.hootsuite.com/how-the-youtube-algorithm-works/

Data Mining Software, Model Development and Deployment, SAS Enterprise Miner. (2019). Retrieved from https://www.sas.com/en_us/software/enterprise-miner.html

Graphical User Interface for SAS, SAS Enterprise Guide. (2019). Retrieved from https://www.sas.com/en_us/software/enterprise-guide.html

Joseph Hogue, C. (2019). Can You Still Make Money on YouTube?. Retrieved from https://myworkfromhomemoney.com/make-money-youtube/

Keller, J. (2018). YouTube: Everything you need to know! Retrieved from https://www.imore.com/youtube-everything-you-need-know

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems With Applications*, *38*(12), 15273-15285. doi: 10.1016/j.eswa.2011.06.028

O'Kane, C. (2019). Top 10 highest-paid YouTube stars of 2018, according to Forbes. Retrieved from https://www.cbsnews.com/news/top-10-highest-paid-youtube-stars-of-2018-forbes/

YouTube Help. (n.d.). How to earn money on YouTube - YouTube Help. Retrieved from https://support.google.com/youtube/answer/72857?hl=en

YouTube Help. (n.d.). Trending on YouTube - YouTube Help. Retrieved from https://support.google.com/youtube/answer/7239739?hl=en