Viral Video Data Modelling

Project Deliverable 2

Oklahoma State University

MSIS 5223 Programming for Data Science and Analytics II

Dr. Bryan Hammer

Team Members: Abhishek Ashwin Bhale (20144060)
Shivani Shashikant Achrekar (20212861)
Harshita Menon Achat (20212860)
Divya Peddapeta (20228697)

Table of Contents

Executive Summary

YouTube is one of the largest video sharing platforms where users and the general public can watch, like, share, comment and upload their videos. YouTube platform consists of two types of users: Video creators (people who have channels and upload videos to them) and Video viewers (people who watch videos, interact with videos and subscribe to channels). Many users use it for entertainment purposes, some to watch tutorials or for keeping up with their favorite artists' latest music videos and so much more. (source: https://www.lifewire.com/youtube-101-3481847)

As of February 2017, there were more than 400 hours of content uploaded to YouTube each minute, and one billion hours of content being watched on YouTube every day. As of August 2018, the website is ranked as the second-most popular site in the world, according to Alexa Internet.

Videos uploaded on YouTube get circulated through the viral process of Internet sharing. These videos may be informative, comic or sometimes deeply emotional. Our objective through this project is to build models to predict the virality of the video.

Statement of Scope

The purpose of this project is to analyze the viral YouTube videos based on the number of comments, likes, dislikes and the view counts given by the users. This analysis will be done using the search keywords ‘Viral Videos’ on YouTube. Predictive modeling techniques such as Neural Network, Linear Regression, Regression tree, Logistics Regression will be used to build models on the numeric data which is scraped from YouTube.

Project Objectives

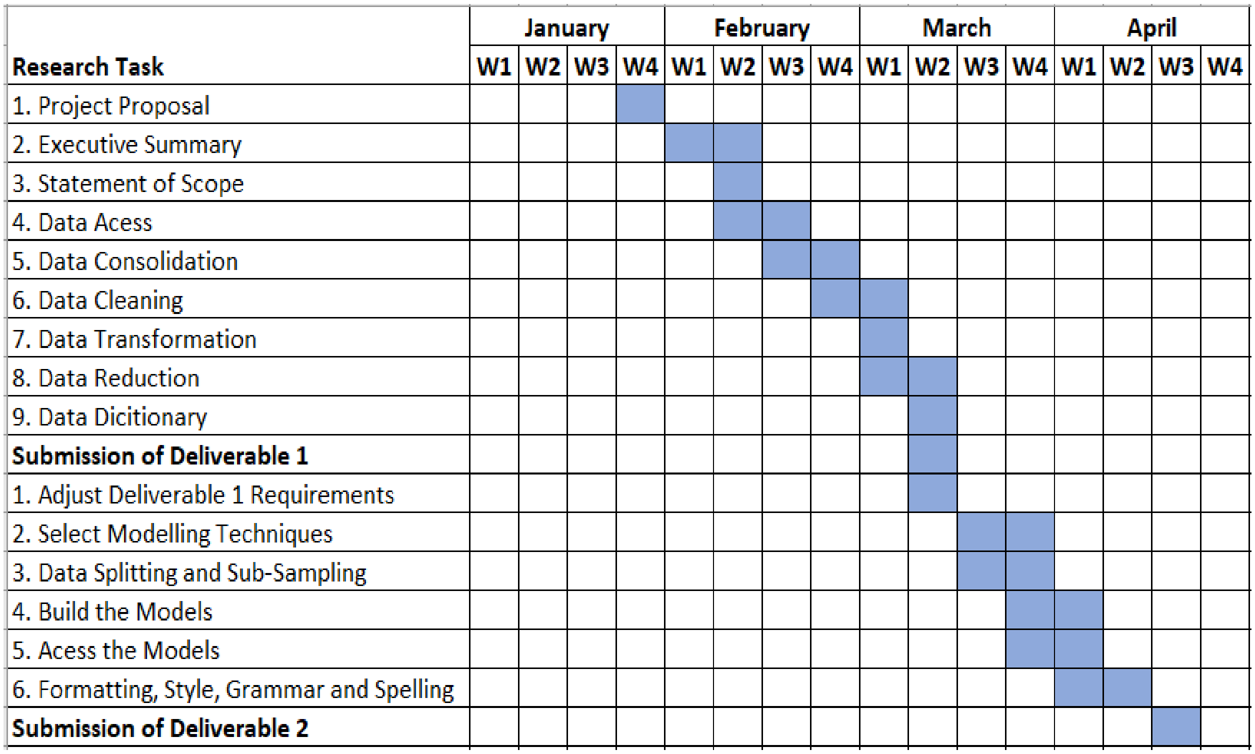
- To perform data splitting and subsampling for training, testing and validating scraped data sets to build models.
- To build a Regression tree using a Target Variable and Predictor Variables.
- To create predictive modeling techniques such as Linear Regression, Logistic Regression and Neural Networks based on the requirement of the data.
- To assess the models and selecting the best fit by comparing their R-square values and accuracy assessments.

Variables

We have scraped the name of the video, video uploader, number of views, likes, dislikes, video id, length of the video in seconds, duration, category, number of comments and the published date. The variables that can be used are Likes, Dislikes, Viewcount, length of the video in seconds, Category and number of comments. We have selected Viewcount as our target variable. The predictor variables will be Likes, Dislikes, No of comments and Length of the video in seconds.

Project Schedule

We have used a GANTT chart for our project schedule. Every week we met on Wednesdays and planned to meet on every other following Wednesday. Worked for 2 to 3 hours every time we met. All the tasks are done together as a group. The project took around 5 weeks to complete. In every meeting, a new plan was created, and we worked on these plans for the following week. We had planned the entire schedule until the final presentation according to the GANTT chart below. Due to the current Pandemic outbreak, we had to make a few changes in our schedule. Instead of meeting in person, we switched to zoom meetings to maintain social distancing. Though there was a change in the pattern and schedule of the project, we were resilient to the changes and have done all the work so that the project could be completed on time.



	January				February				March				April			
Research Task	**W1**	**W2**	**W3**	**W4**	**W1**	**W2**	**W3**	**W4**	**W1**	**W2**	**W3**	**W4**	**W1**	**W2**	**W3**	**W4**
1. Project Proposal				■												
2. Executive Summary					■	■										
3. Statement of Scope						■										
4. Data Acess						■	■									
5. Data Consolidation							■	■								
6. Data Cleaning								■	■							
7. Data Transformation									■							
8. Data Reduction									■	■						
9. Data Dicitionary										■						
Submission of Deliverable 1										■						
1. Adjust Deliverable 1 Requirements										■						
2. Select Modelling Techniques											■	■				
3. Data Splitting and Sub-Sampling											■	■				
4. Build the Models												■	■			
5. Acess the Models												■	■			
6. Formatting, Style, Grammar and Spelling													■	■		
Submission of Deliverable 2															■	

Data Preparation

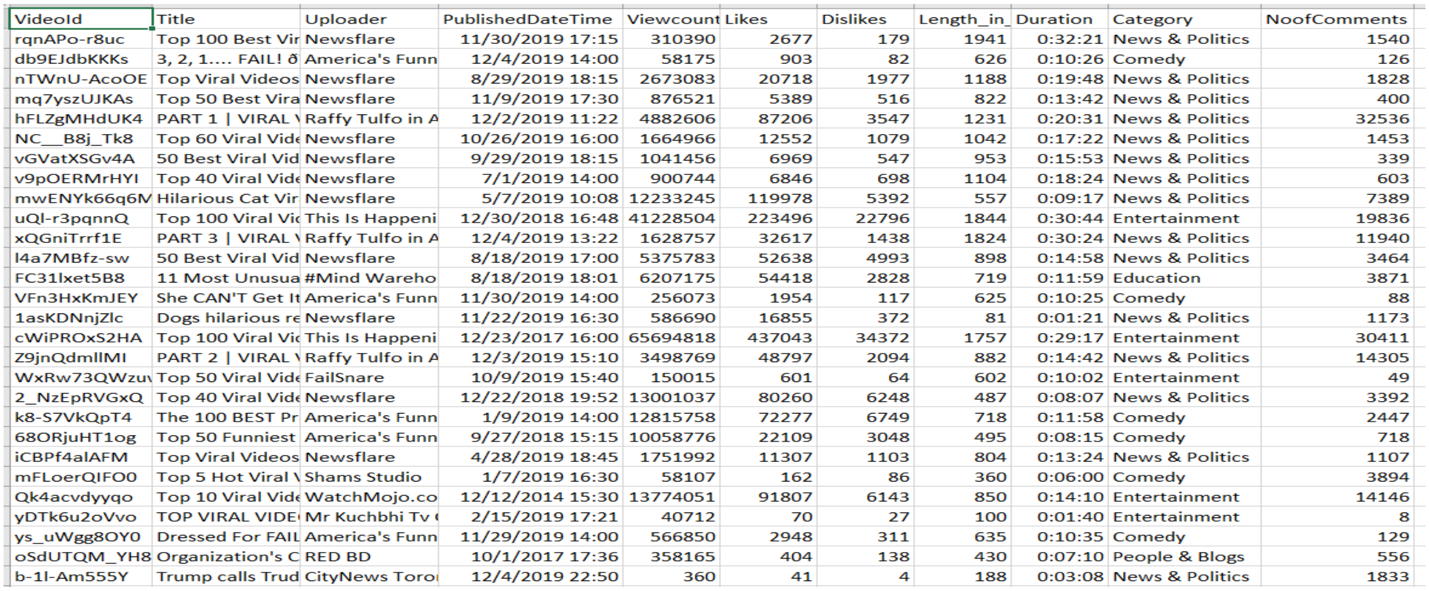
Data Access

Here, we searched for the ‘Viral Videos’ on YouTube as our search query and analyze the top 100 videos that show up after the search. For accessing the data, we used a code that will scrape the links to the top videos that were obtained with the help of Selenium and CSS selectors. We used a library in Python named pafy which pulls the likes, dislikes, view count, name of the video, video Id, uploader, duration, published date, length in seconds and category. All the data comes from YouTube itself.

Data Consolidation

We have done the consolidation process using Python code which we have attached in the Appendix at the end of this document.

CSV file of YouTube Data.



VideoId	Title	Uploader	PublishedDateTime	Viewcount	Likes	Dislikes	Length_in_	Duration	Category	NoofComments
rqnAPo-r8uc	Top 100 Best Vir	Newsflare	11/30/2019 17:15	310390	2677	179	1941	0:32:21	News & Politics	1540
db9EJdbKKKs	3, 2, 1.... FAIL! ð	America's Funn	12/4/2019 14:00	58175	903	82	626	0:10:26	Comedy	126
nTWnU-AcoOE	Top Viral Videos	Newsflare	8/29/2019 18:15	2673083	20718	1977	1188	0:19:48	News & Politics	1828
mq7yszUJKAs	Top 50 Best Vira	Newsflare	11/9/2019 17:30	876521	5389	516	822	0:13:42	News & Politics	400
hFLZgMHdUK4	PART 1 | VIRAL \	Raffy Tulfo in A	12/2/2019 11:22	4882606	87206	3547	1231	0:20:31	News & Politics	32536
NC__B8j_Tk8	Top 60 Viral Vide	Newsflare	10/26/2019 16:00	1664966	12552	1079	1042	0:17:22	News & Politics	1453
vGVatXSGv4A	50 Best Viral Vid	Newsflare	9/29/2019 18:15	1041456	6969	547	953	0:15:53	News & Politics	339
v9pOERMrHYI	Top 40 Viral Vide	Newsflare	7/1/2019 14:00	900744	6846	698	1104	0:18:24	News & Politics	603
mwENYk66q6M	Hilarious Cat Vir	Newsflare	5/7/2019 10:08	12233245	119978	5392	557	0:09:17	News & Politics	7389
uQl-r3pqnnQ	Top 100 Viral Vic	This Is Happeni	12/30/2018 16:48	41228504	223496	22796	1844	0:30:44	Entertainment	19836
xQGniTrrf1E	PART 3 | VIRAL \	Raffy Tulfo in A	12/4/2019 13:22	1628757	32617	1438	1824	0:30:24	News & Politics	11940
l4a7MBfz-sw	50 Best Viral Vid	Newsflare	8/18/2019 17:00	5375783	52638	4993	898	0:14:58	News & Politics	3464
FC31lxet5B8	11 Most Unusua	#Mind Wareho	8/18/2019 18:01	6207175	54418	2828	719	0:11:59	Education	3871
VFn3HxKmJEY	She CAN'T Get It	America's Funn	11/30/2019 14:00	256073	1954	117	625	0:10:25	Comedy	88
1asKDNnjZlc	Dogs hilarious re	Newsflare	11/22/2019 16:30	586690	16855	372	81	0:01:21	News & Politics	1173
cWiPROxS2HA	Top 100 Viral Vic	This Is Happeni	12/23/2017 16:00	65694818	437043	34372	1757	0:29:17	Entertainment	30411
Z9jnQdmllMI	PART 2 | VIRAL \	Raffy Tulfo in A	12/3/2019 15:10	3498769	48797	2094	882	0:14:42	News & Politics	14305
WxRw73QWzuv	Top 50 Viral Vide	FailSnare	10/9/2019 15:40	150015	601	64	602	0:10:02	Entertainment	49
2_NzEpRVGxQ	Top 40 Viral Vide	Newsflare	12/22/2018 19:52	13001037	80260	6248	487	0:08:07	News & Politics	3392
k8-S7VkQpT4	The 100 BEST Pr	America's Funn	1/9/2019 14:00	12815758	72277	6749	718	0:11:58	Comedy	2447
68ORjuHT1og	Top 50 Funniest	America's Funn	9/27/2018 15:15	10058776	22109	3048	495	0:08:15	Comedy	718
iCBPf4alAFM	Top Viral Videos	Newsflare	4/28/2019 18:45	1751992	11307	1103	804	0:13:24	News & Politics	1107
mFLoerQIFO0	Top 5 Hot Viral \	Shams Studio	1/7/2019 16:30	58107	162	86	360	0:06:00	Comedy	3894
Qk4acvdyyqo	Top 10 Viral Vide	WatchMojo.co	12/12/2014 15:30	13774051	91807	6143	850	0:14:10	Entertainment	14146
yDTk6u2oVvo	TOP VIRAL VIDE(	Mr Kuchbhi Tv (	2/15/2019 17:21	40712	70	27	100	0:01:40	Entertainment	8
ys_uWgg8OY0	Dressed For FAIL	America's Funn	11/29/2019 14:00	566850	2948	311	635	0:10:35	Comedy	129
oSdUTQM_YH8	Organization's C	RED BD	10/1/2017 17:36	358165	404	138	430	0:07:10	People & Blogs	556
b-1l-Am555Y	Trump calls Trud	CityNews Toro	12/4/2019 22:50	360	41	4	188	0:03:08	News & Politics	1833

Data Cleaning

There are a few videos that have missing data for Likes and Dislikes columns. So, as to handle this missing data, we did Listwise deletion. We scraped the number of comments using XPath. We used regular expressions to clean this data that consisted of the word comments with the numeric value. We detected some outliers in the data, and we are including them in the data set to answer a few questions.

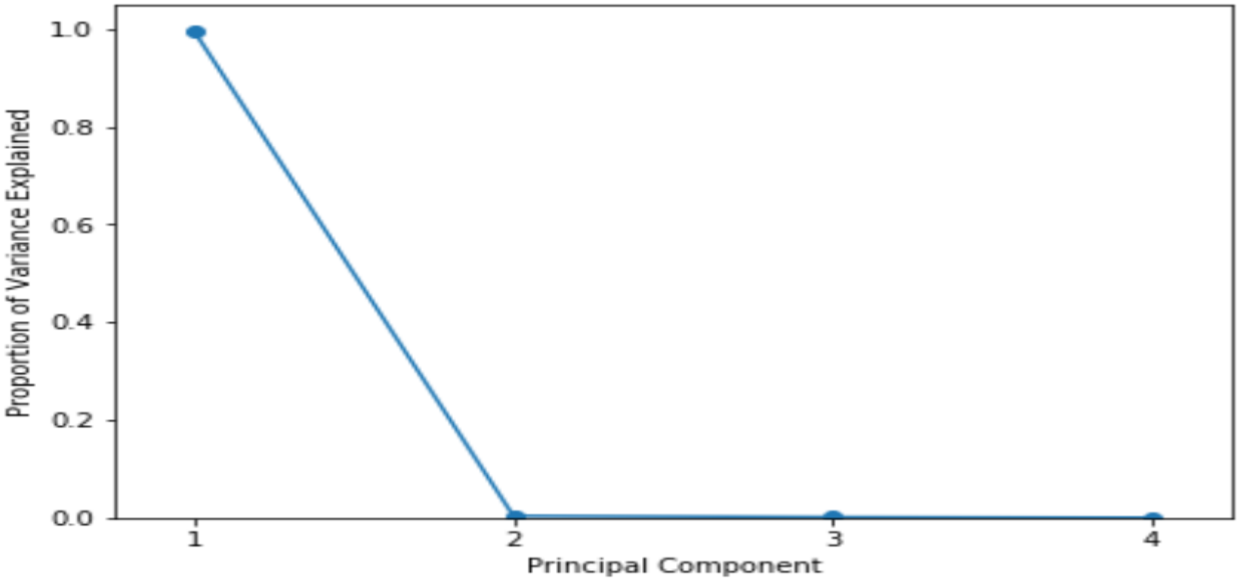
Data Transformation

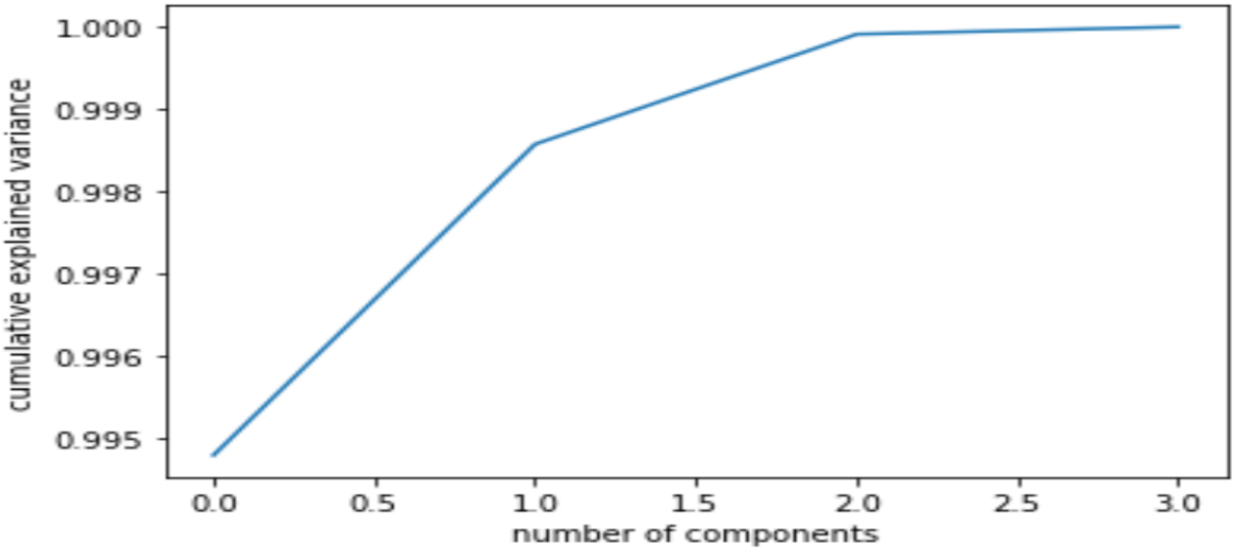
For Logistic regression, we have converted the categorical variable 'Category' to numeric. The rest of the variables which were required for the regression were already in numeric format. The further transformation was not required on the dataset as it was in the usable format to implement modeling. There was no need to apply any log transformations or normalize the data.

Data Reduction

Principal Components Analysis (PCA) is an unsupervised statistical method to reduce the dimensions in the data. It takes the most important information from all the variables in the data and projects the information to orthogonal principal components, following the principle of diminishing marginal return. That is, the first principal component captures the most information from the data and the residual information is captured the next component and so on. Each component is uncorrelated with other components. These components can be used as independent variables in building regression models and as clustering variables. We have used Principal Component Analysis (PCA) as a dimension reduction technique. We have done Principal Component Analysis on the 4 numeric variables 'Likes', 'Dislikes, 'No of comments' and 'Length in seconds'. After running the analysis, we found that the first principal component captures most of the information and the remaining 3 components almost have no value. This agrees with the intuition that a video generally has a high correlation between the likes, dislikes, and comments on a video.

From the image below, we can see that choosing a single principal component is enough.





Data Dictionary

Attribute Name	Description	Data Type	Source
VideoId	It is the alphanumeric unique id of the individual videos	String	https://www.youtube.com/
Title	Name of the video	Char(30)	https://www.youtube.com/
Uploader	Name of the channel that published the video	Char(30)	https://www.youtube.com/
PublishedDateTime	The time and date the video was uploaded	Date time format as String	https://www.youtube.com/
Viewcount	The number of users who viewed the video	Integer	https://www.youtube.com/
Likes	Users who liked the video	Integer	https://www.youtube.com/
Dislikes	Users who disliked the video	Integer	https://www.youtube.com/
Length_in_sec	Length of the video in secs	Integer	https://www.youtube.com/
Duration	The duration of the video in HH:MM:SS format	String	https://www.youtube.com/
Category	Category of the video	String	https://www.youtube.com/
No of Comments	Count of the comments on the video	Integer	https://www.youtube.com/

Selecting Modeling Technique

We have implemented four modeling techniques in our project.

Logistic Regression: The goal of selecting the logistic regression technique for our dataset was to find the dependency between the variables viewcount, Likes, Dislikes, No of comments, length in seconds and category of the video. Logistic Regression assumes that the viewcount is determined by the type of category of the videos. When the regression was applied to the data, we were hoping to find that the likability of the video could be determined by its category. But this was not the case that we found. The results were contradictory to what we expected.

Regression Tree: The goal of selecting this modeling technique is to know which variable is the most important in predicting the target variable. Initially, we assumed that Likes would have more impact on the target variable ‘viewcount’. After performing this technique, we found that Dislikes is having more impact on the target variable.

Neural Networks: The goal of selecting this model is to find out the accuracy of how well the model fits the data. It didn’t fit well because the sample size of our data is small. Neural networks require large datasets for training and testing.

Linear Regression: The goal of selecting this modeling technique is to find out which independent variables are significant in predicting the target variable ‘viewcount’. Our assumption was still the same that Likes would be the most significant in determining our target variable. We found out that Likes, Dislikes and No of comments are significant in determining the target variable.

Data Splitting and Sub-Sampling

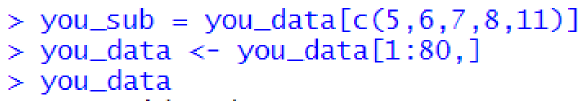
A model’s training performance is usually an optimistic metric that will overestimate its test performance. This inherent tendency of the model will leave us unsure of whether to deploy it in production. So, we need to ensure that its test performance is good.

Holdout sample

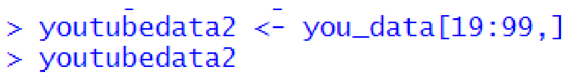
We have divided the data into 2 parts, the first one contains 80% of the data and the second one is given 20%. The idea behind the splitting is that we build the model on the training data and use it to evaluate its performance on the test data. This test performance is a more accurate estimate of its true usability.

There are 100 records in our data set, so we have performed splitting and sub-sampling to select 80% of records. We have selected the numeric columns excluding all other variables.

Code Snippets:



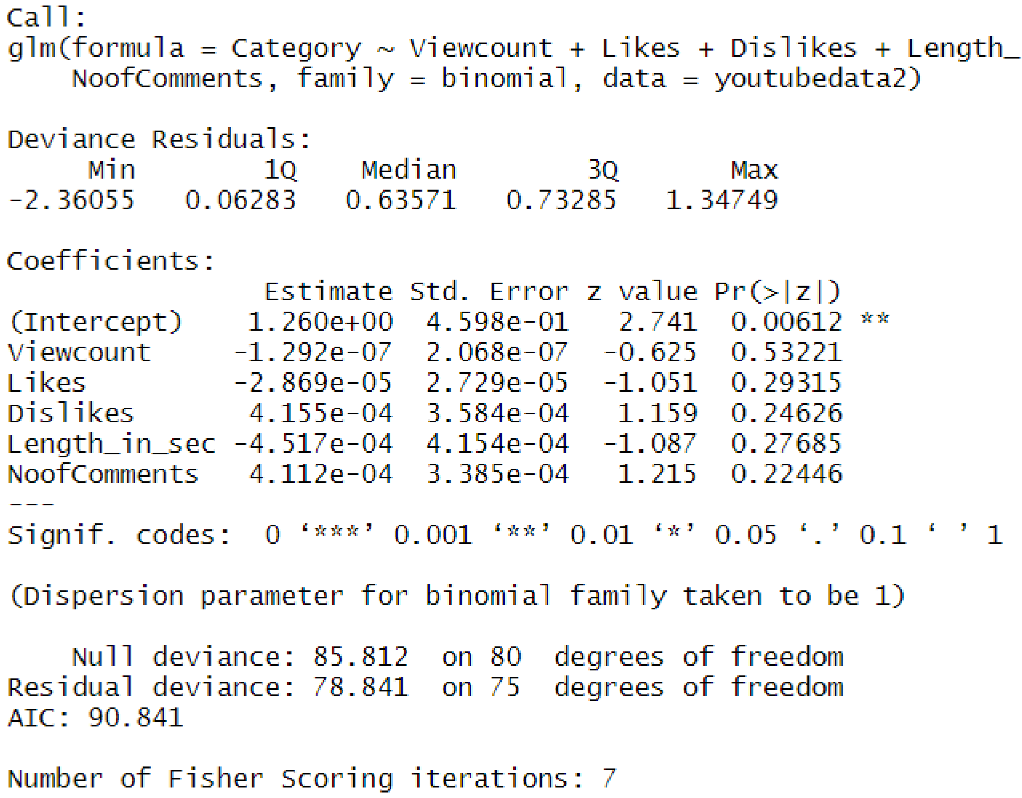
```
> you_sub = you_data[c(5,6,7,8,11)]
> you_data <- you_data[1:80,]
> you_data
```



```
> youtubedata2 <- you_data[19:99,]
> youtubedata2
```

Build the Models

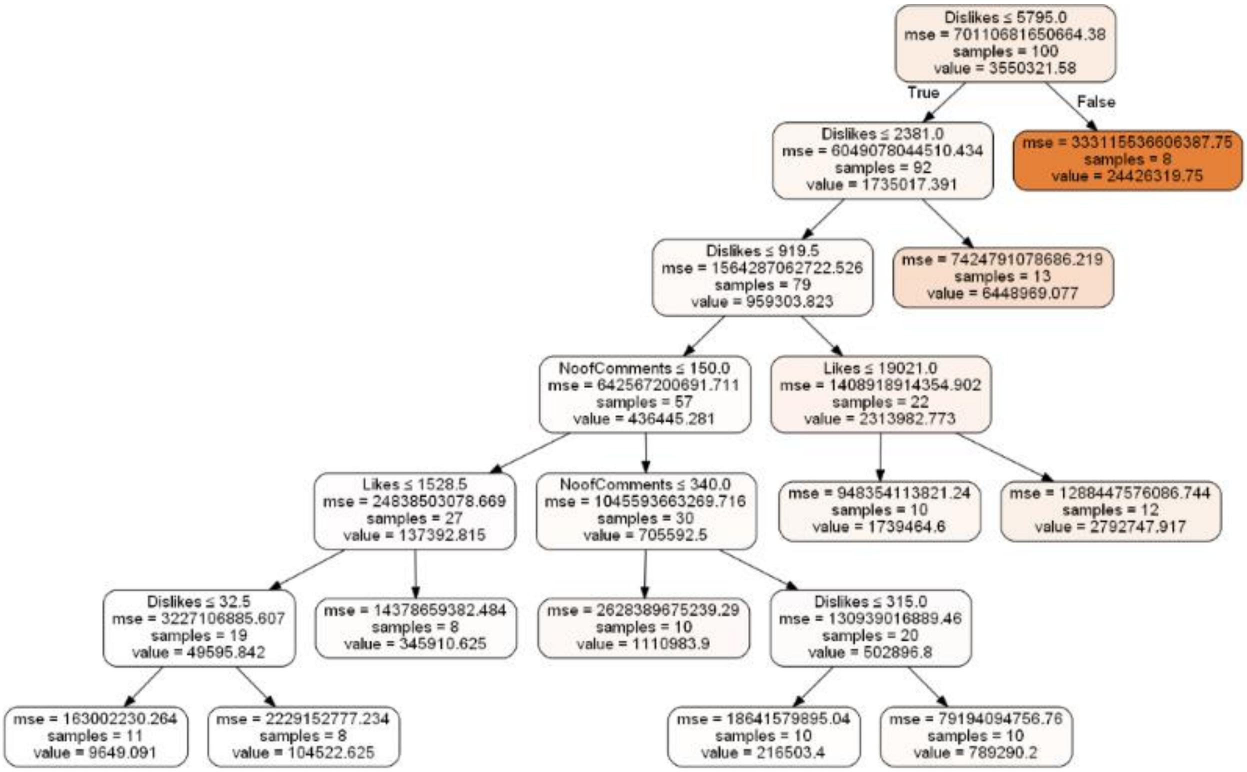
- Logistic Regression



```
Call:
glm(formula = Category ~ Viewcount + Likes + Dislikes + Length_in_sec +
    NoofComments, family = binomial, data = youtubedata2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.36055   0.06283   0.63571   0.73285   1.34749

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.260e+00  4.598e-01   2.741  0.00612 **
Viewcount     -1.292e-07  2.068e-07  -0.625  0.53221
Likes         -2.869e-05  2.729e-05  -1.051  0.29315
Dislikes       4.155e-04  3.584e-04   1.159  0.24626
Length_in_sec -4.517e-04  4.154e-04  -1.087  0.27685
NoofComments   4.112e-04  3.385e-04   1.215  0.22446
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 85.812  on 80  degrees of freedom
Residual deviance: 78.841  on 75  degrees of freedom
AIC: 90.841

Number of Fisher Scoring iterations: 7
```

The above image is the output of a logistic regression model with ‘Category’ being the target variable and ‘Likes’, ‘Dislikes, ‘Length in seconds’ and ‘No of comments’ as the explanatory variables. From the model, none of the variables are significant at 5% level. That follows the intuition as the category of video can’t be explained by the fame of the video. We just tried to see if there is any relation between the views and likes and comments of a video vs the category it belongs to. But we found out that there is no relation. Videos from all categories are being viewed and commented on and liked at the same level.

- Regression Tree



This is the output of our decision tree model with the continuous target variable (Viewcount). From the image, we can see that our final tree has 10 root nodes. The parameters min_sample_split and min_sample_leaf is both set to 8. The output reveals Dislikes is the most important variable followed by 'Noofcomments' and 'Likes' for splitting.

Terminal node 1: Left side, Dislikes <= 32.5. It has a total of 19 samples which is further divided into 11 and 8 samples based on the number of Dislikes. We have 11 samples when Dislikes <= 32.5 and 8 when Dislikes > 32.5. The 11-samples set has an average value for Viewcount 9649.091 and the 8-sample set has an average value for Viewcount 104522.625. The value is high as the Views on videos are in millions.

Terminal node 2: Right side, Dislikes <= 315.0. It has a total 20 samples which are further divided into 10 and 10 samples based on the number of Dislikes. We have 10 samples when Dislikes <=

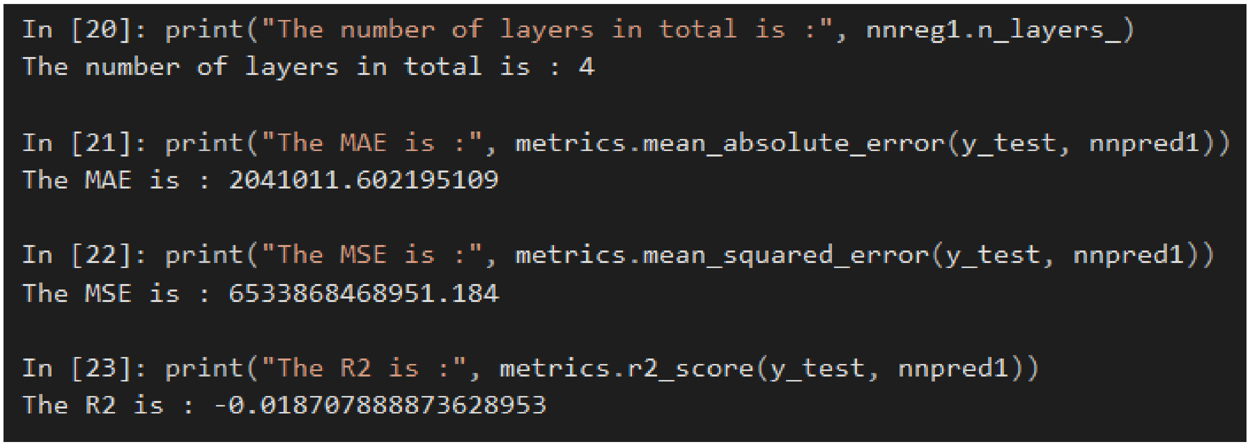
315.0 with the average value for Viewcount 216503.4 and 10 when Dislikes > 315.0 with an average value for Viewcount 789290.2.

Terminal node 3: Right side, Likes <= 19021.0. It has a total of 22 samples which is further divided into 10 and 12 samples based on the number of Likes. We have 10 samples when Likes <= 19021.0 and 12 when Likes > 19021.0. The 10-samples set has an average value for Viewcount 1739464.6 and the 12-sample set has an average value for Viewcount 279277.917.

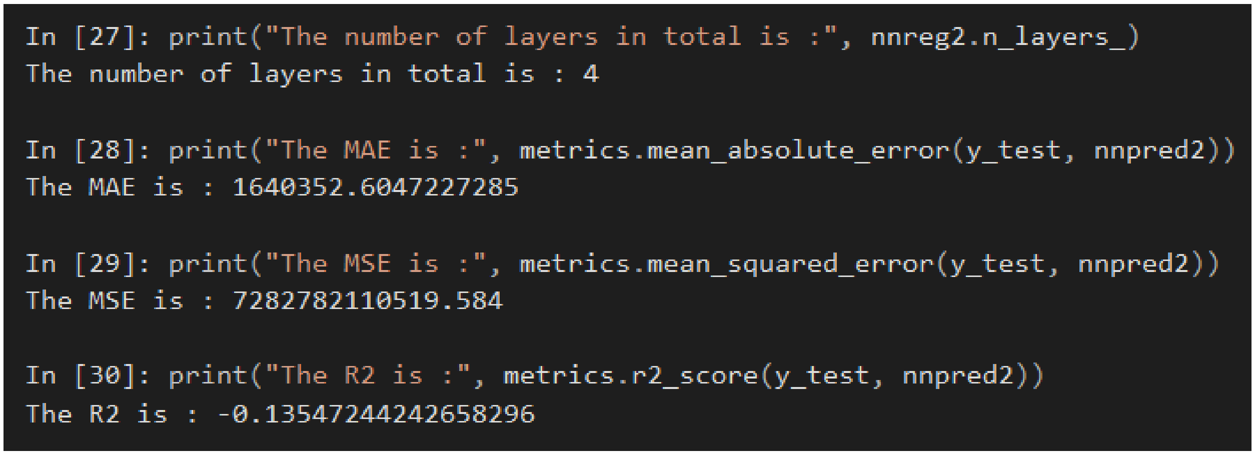
Terminal node 4: NoofComments <= 340.0. It has a total of 30 samples which is further divided into 10 and 20 samples based on the number of NoofComments. We have 10 samples when NoofComments <= 340.0 and 20 when NoofComments > 340.0. The 10-samples set has an average value for Viewcount 1110983.9 and the 20-sample set has an average value for Viewcount 502896.8.

- Neural Networks

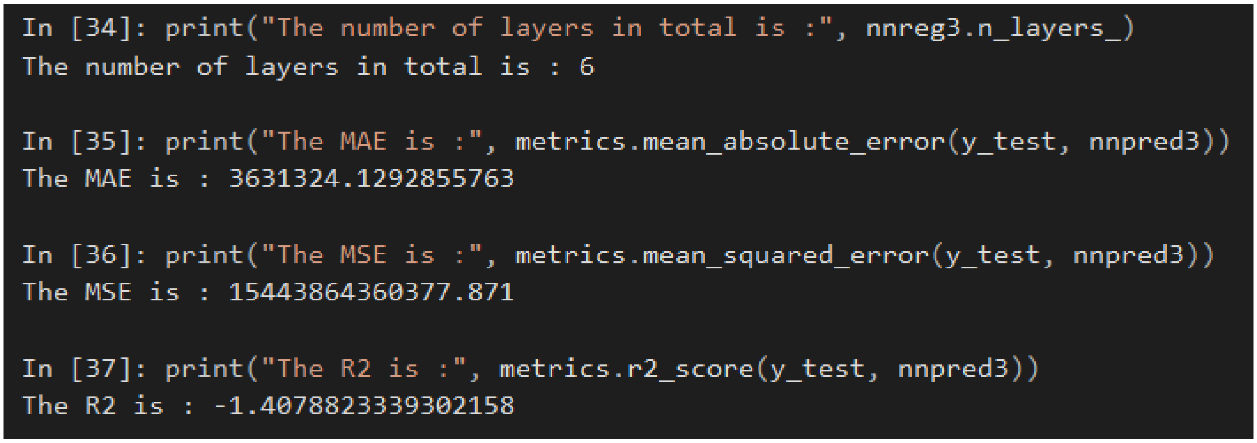
Artificial neural networks (ANN) are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron that receives a signal then processes it and can signal neurons connected to it.



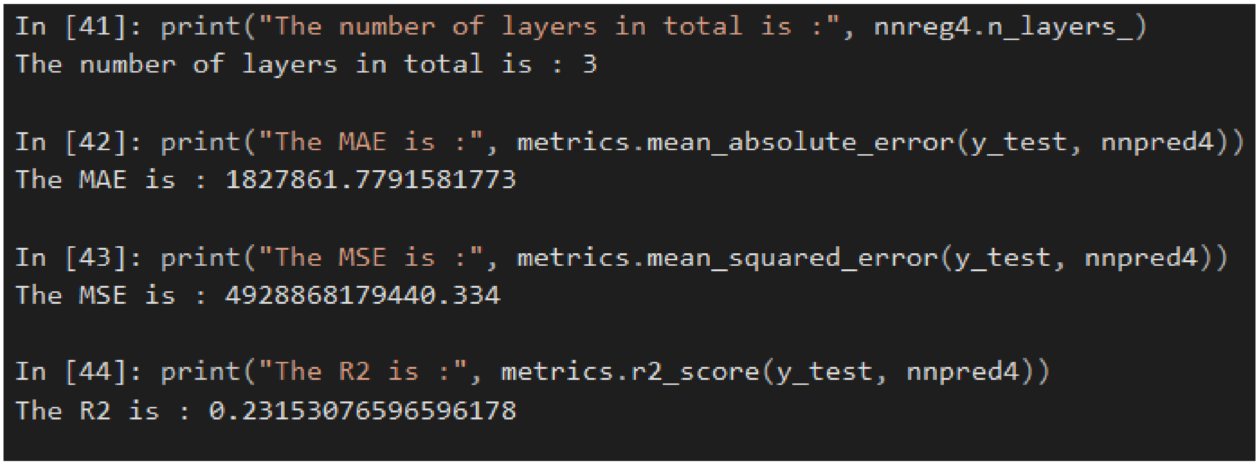
```
In [20]: print("The number of layers in total is :", nnreg1.n_layers_)
The number of layers in total is : 4

In [21]: print("The MAE is :", metrics.mean_absolute_error(y_test, nnpred1))
The MAE is : 2041011.602195109

In [22]: print("The MSE is :", metrics.mean_squared_error(y_test, nnpred1))
The MSE is : 6533868468951.184

In [23]: print("The R2 is :", metrics.r2_score(y_test, nnpred1))
The R2 is : -0.018707888873628953
```



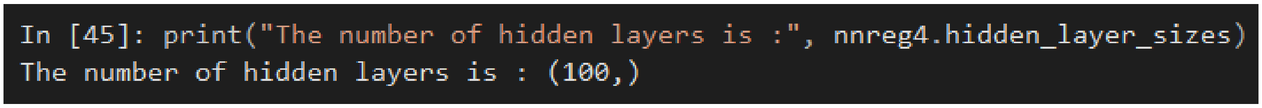
```
In [27]: print("The number of layers in total is :", nnreg2.n_layers_)
The number of layers in total is : 4

In [28]: print("The MAE is :", metrics.mean_absolute_error(y_test, nnpred2))
The MAE is : 1640352.6047227285

In [29]: print("The MSE is :", metrics.mean_squared_error(y_test, nnpred2))
The MSE is : 7282782110519.584

In [30]: print("The R2 is :", metrics.r2_score(y_test, nnpred2))
The R2 is : -0.13547244242658296
```



```
In [34]: print("The number of layers in total is :", nnreg3.n_layers_)
The number of layers in total is : 6

In [35]: print("The MAE is :", metrics.mean_absolute_error(y_test, nnpred3))
The MAE is : 3631324.1292855763

In [36]: print("The MSE is :", metrics.mean_squared_error(y_test, nnpred3))
The MSE is : 15443864360377.871

In [37]: print("The R2 is :", metrics.r2_score(y_test, nnpred3))
The R2 is : -1.4078823339302158
```



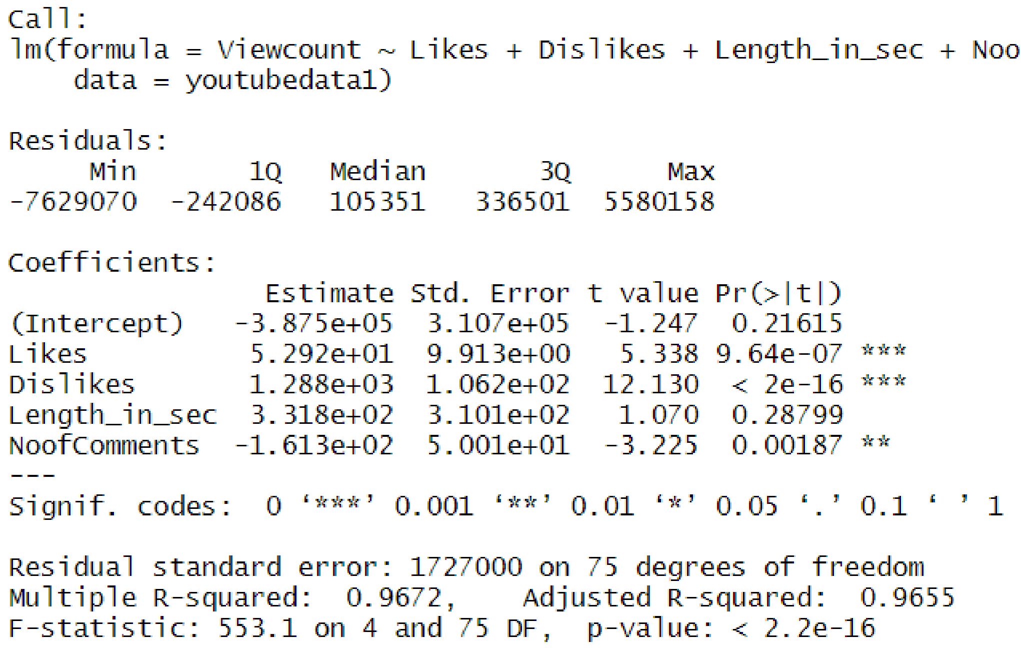
```
In [41]: print("The number of layers in total is :", nnreg4.n_layers_)
The number of layers in total is : 3

In [42]: print("The MAE is :", metrics.mean_absolute_error(y_test, nnpred4))
The MAE is : 1827861.7791581773

In [43]: print("The MSE is :", metrics.mean_squared_error(y_test, nnpred4))
The MSE is : 4928868179440.334

In [44]: print("The R2 is :", metrics.r2_score(y_test, nnpred4))
The R2 is : 0.23153076596596178
```



```
In [45]: print("The number of hidden layers is :", nnreg4.hidden_layer_sizes)
The number of hidden layers is : (100,)
```

We built 4 NN models in this project with several layers and several hidden units. Each NN model is built to predict the view count of a video based on the 4 other numeric variables ‘Likes’, ‘Dislikes’, ‘Length in secs’ and ‘Noofcomments’. NN models don’t have any specific assumptions regarding the distribution of variables like regression models. They, however, would work only with non-missing values and will exclude any observations with missing values. Each of the 4 numeric variables is standardized using standard scalar before inputting to the model. The final goal of the model is to predict the value of viewcount for each video. The one compromise we do in NN models is that we lack the power of interpretability of the model for the sake of accuracy. ANN model, when properly trained is often much more accurate than any statistical model such as regression. But we don’t have anything to explain about the model because the NN model works as a Blackbox. The ANN model with 3 hidden layers and 100 units in each layer is the best model among the 3 models that we built. It has an R^2 of 23.1%. That means, 23.1% of the variability in the target variable is being explained in the model. This model fails in comparison with other models such as linear regression and decision trees. This is because ANN models need a lot of data and just 100 data points are not enough to fully take advantage of the complexities of ANN models.

- Linear Regression



```
Call:
lm(formula = Viewcount ~ Likes + Dislikes + Length_in_sec + NoofComments,
    data = youtubedata1)

Residuals:
     Min       1Q   Median       3Q      Max 
-7629070  -242086   105351   336501  5580158 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.875e+05  3.107e+05  -1.247  0.21615    
Likes          5.292e+01  9.913e+00   5.338 9.64e-07 ***
Dislikes       1.288e+03  1.062e+02  12.130  < 2e-16 ***
Length_in_sec  3.318e+02  3.101e+02   1.070  0.28799    
NoofComments  -1.613e+02  5.001e+01  -3.225  0.00187 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1727000 on 75 degrees of freedom
Multiple R-squared:  0.9672,	Adjusted R-squared:  0.9655 
F-statistic: 553.1 on 4 and 75 DF,  p-value: < 2.2e-16
```

Now, we have built a linear regression model with ‘Viewcount’ as the target variable and ‘Likes’, ‘Dislikes, ‘Length in seconds’ and ‘No of comments’ as the explanatory variables. The main model is significant at 5% level. From the model, ‘Likes’, ‘Dislikes’ and ‘Noofcomments’ are significant at a 5% level. ‘Length in seconds’ is not significant at 5% level. The adjusted R2 is 96.55%. That means that 96.55% of the target variable’s variance is being explained by the independent variables.

‘Dislikes’ is the most important variable, with the highest t-value followed by ‘Likes’ and ‘Noofcomments’. This is an interesting finding. More important than the likes, the number of dislikes a video gets is more explanatory of the views a video would get. That means a video with a high number of views would have more dislikes proportionally than the likes. This is an interesting observation, and this could have something to say about the psychology of the people.

Assess the Models

Strengths and weaknesses of the models used:

Linear Regression model:

Strengths: It has good statistical backing, and it provides a very clean and understandable output and each variable's significance is explained in the regression model and easy to understand and provides good prediction models.

Weaknesses: A Linear Regression model has certain assumptions about the distribution of the data.

1. Each variable needs to be linearly correlated to the target variable.
2. The independent variables need to be uncorrelated with each other.
3. The independent variables should have a near-normal type of distribution where having extreme outliers will affect the data drastically.
4. The residuals of the model need to be distributed normally with mean value 0 and variance 1.
5. It cannot handle missing values.

However, multiple regression is robust to the violation of the data.

Decision Tree:

Strengths: It is a non-parametric, supervised Machine Learning model which doesn't assume any distribution for the intended models, so decisions are much more robust than regression models when we have non-normal data and the data which is not linearly related to the target variable. The output is even more interpretable because it forms a tree-like structure which makes it easier to understand. It can handle missing values.

Weaknesses: Decision trees are prone to overfitting so they may need regularization.

Artificial Neural Network (ANN):

Strengths: When ANNs are trained properly with large data, results are much more accurate than any other model.

ANN models are just like tree-based models that don’t have any explicit assumptions over distribution for the independent variables.

Weaknesses: ANN models can’t handle missing values and require data to be more normal.

We don’t understand what’s going in the model and we don’t know the reason for the model failure.

Logistic Regression:

Strengths: It is the simplest classification model and it is accurate when the decision boundary is linear.

It provides the probabilities of an event happening and not happening and it is easy for us to select specified cutoff.

Weaknesses:

Data should be normal and should not have any missing values and don’t work well when the decision boundary is not linear.

Conclusion

We consider that Linear regression is the best modeling technique for our data. Because as the target variable is continuous, we compare R^2 of the models to understand how well the model performs. The ANN model provides an R^2 value of 23.1% while the Linear Regression model provides 96.55% value of R^2. Subjectively, the relative importance of each coefficient is well explained in the Linear Regression model. Regression also provides a relation between the independent variable and the dependent variable.

Appendix

```
## Creating the DataFrame ##

list = {'VideoId':id, 'Title':name, 'Uploader':aut, 'PublishedDateTime':pub,
        'Viewcount':count, 'Likes':like, 'Dislikes': dislike, 'Length_in_sec': lent,
       'Duration': dur, 'Category': categ, 'NoofComments': no_of_comments}
youtube_data = pd.DataFrame(list)
```

Reference:

https://www.youtube.com/

https://pythonhosted.org/Pafy/

https://en.wikipedia.org/wiki/Artificial_neural_network