e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTIVE ANALYSIS OF YOUTUBE USING MACHINE LEARNING

Punam Wake[*1], Nita Borkar[*2], Swati Dakhore[*3], Mr. Vijay R. Wadhankar[*4], Mr. Abhishek Kumar[*5]

[*1,2,3]Department Of Electronics & Communication Agnihotri College Of Engineering Nagthana, Wardha, India.

[*4]Assistant Prof., HOD, Electronics & Communication Department, Agnihotri College Of Engineering, Nagthana, Wardha, India.

[*5]Assistant Prof. (Principal), Electronics & Communication Department, Agnihotri College Of Engineering, Nagthana, Wardha, India.

ABSTRACT

As we all know using and watching YouTube videos is a crucial a part of our everyday lives. Most people try to create their influence, income, and impact with YouTube and online video. In nutshell, most are trying to be a YouTube influencer. It will be nice if a YouTube influencer can get an idea of how the view count goes to be before making and finalizing the video. In here we tried to make a model which will help influencers to predict the amount of views for his or her next video.

In this work, we have a tendency to propose a regression technique to predict the recognition of an internet video measured by its range of views. we have a tendency to show that predicting quality patterns with this approach provides additional precise and additional stable prediction results, principally because of the non-linear character of the projected technique additionally as its strength. We have a tendency to prove the prevalence of our technique against the education purpose video victimization datasets containing videos from YouTube. We have a tendency to conjointly show that victimization visual options, like the outputs of deep neural networks or scene dynamics' metrics, are often helpful for quality prediction before content publication. moreover, we have a tendency to show that quality prediction accuracy are often improved by combining early distribution patterns with social and visual options which social options represent a way stronger signal in terms of video quality prediction than the visual ones.

Predicting web page quality is a crucial task for supporting the planning and analysis of a good vary of systems, from targeted advertising to effective search and recommendation services. we have a tendency to here gift 2 easy models for predicting the long run quality of web page supported historical info given by early quality measures. Our approach is valid on datasets consisting of videos from the wide used YouTube video- sharing portal. Our experimental results show that, compared to a education baseline model, our planned models cause vital decreases in relative square errors, reaching up to 20% reduction on the average, and bigger reductions of up to 71% for videos that have a high peak in quality in their time period followed by a pointy decrease in quality.

Keywords: Machine Learning, Linear Regression, Polynomial Regression, K-Nearest Neighbors, Decision Tree Regression.

I. INTRODUCTION

Over the past five years YouTube has paid out quite $5 billion to YouTube content creators. widespread YouTuber PewDiePie created $5 million in 2016 from YouTube alone, not as well as sponsorships, endorsements and alternative deals outside of YouTube. With additional and additional corporations turning to YouTube influencers to capture the time period audience, obtaining individuals to look at your videos on YouTube is turning into progressively moneymaking. As YouTube becomes one in every of the foremost in style video-sharing platforms, YouTuber is developed as a replacement sort of career in recent decades. YouTubers earn cash through advertising revenue from YouTube videos, sponsorships from corporations, merchandise sales, and donations from their fans. So as to keep up a stable financial gain, the recognition of videos become the highest priority for YouTubers. Meanwhile, a number of our friends square measure YouTubers or channel house owners in different video-sharing platforms. This raises our interest in predicting the performance of the video. If creators will have a preliminary prediction on their videos' performance, they'll modify their video to

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

realize the foremost attention from the general public. YouTubers concern concerning what percentage individuals watch their videos the foremost. Therefore, the videos are often sorted into in style and non-popular supported the amount of views. Audiences' feedbacks are vital for YouTubers, as a result of the feedbacks mirror the preference of audiences. Therefore, among the popular videos, the videos are any divided into overwhelming praises, overwhelming unhealthy views, and neutral videos based on the feedback. In order to predict the performances of videos, the videos' properties {title, like, dislikes, views, comments, subscriber} are selected as the inputs of the machine learning algorithm. The multi-classification algorithms {Linear regression, multiple-linear regression, , polynomial regression, decision trees, random forest, support vector, K-nearest neighbors (KNN)} are used to output the predicted class. In order to optimize the cost of algorithms, feature selection algorithm is used to select the most efficient combination of features.

II. LITARATURE REVIEW

This paper explains us about concerning many machine learning algorithms are used to predict the performance and backward search is used on features to select out the foremost relevant options. Being a latest kind of job, YouTubers earn cash through the promotional advertisement and bonus from videos. Hence, the recognition of videos is that the prime priority of YouTuber. This project tries to predict the working of the videos that are going to be uploaded to YouTube. Predicting quality of social media contents has attracted wide attention in recently years. The dimensions of gigantic database make machine learning become a sturdy tool to deal with the matter. This project explores the way to use machine learning algorithm to predict the video performance for YouTuber. Once mistreatment several algorithms, model enhancements, and backward search on features.

Thus, in this paper we have a tendency to formalize the matter of predicting trends and hits in user generated videos. Also, we have a tendency to describe our research and analysis methodology on approaching this problem. To the simplest of information, our work is novel in that specializes in the problem of predicting popularity trends complementary to hits. Moreover, we have a tendency to intend on evaluating effectiveness of our results not solely based on common applied statistical error metrics, however conjointly on the attainable online advertising revenues our predictions can generate. Once describing our proposal, we have a tendency to here summarize our latest findings regarding to (1) uncovering common popularity trends,

(2) measuring associations between UGC features and recognition trends, and (3) assessing the effectiveness of models for predicting quality trends.

III. RESEARCH METHOD

In this experiment, our main objective is to predict the potential total view count of a YouTube video as accurate as possible based on several influential factor. To do so, we have decided to apply one of the predictive modeling techniques, which is the regression technique in our experiment. We will be using regression technique to model the mathematical correlation between our independent variables, which are the attributes in the dataset we have acquired, and the dependent variable, in this case it is the amount of viewership of a YouTube video. After figuring out the pattern and the relationship between the said variables, we are then able to predict the future value of the dependent variable.

One of the key benefits that regression analysis offers is that it indicates the strength of impact of multiple independent variables on a dependent variable. This will allow us to compare the outcome when a variable has its value changed. For example, the result of this experiment will show how the channel subscriber count will affect the total view count of a YouTube video.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

IV. SYSTEM WORKFLOW

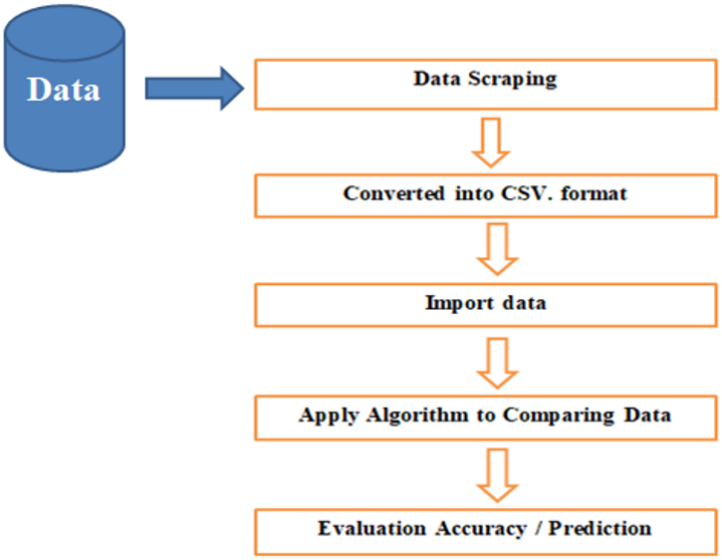

Figure 1: System Workflow

V. DATA SCRAPING

- We were unable to search out an appropriate dataset, therefore we have a tendency to scraped our own. we used Parse Hub that contains 32 GB value of pre-labeled information categorized by numerous genres (i.e. Sports, Fashion, Movies). we have to filtered out all the info with labels that give us 37,353 videos. we have to scraped the subsequent options for every video:
- Likes
- Dislikes
- Views
- Comments
- Subscriber

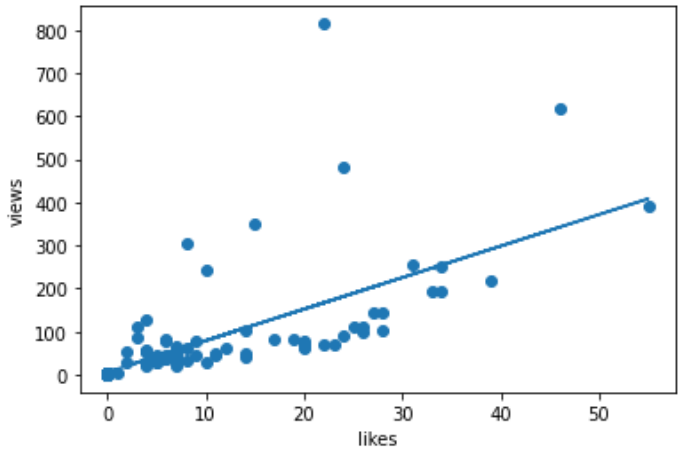

Figure 2: Data Visualization using Views, Likes graph

VI. DATASET

Dataset that have chosen for this project is the trending YouTube Video Statistics:

(https://www.kaggle.com/datasnaek/youtube-new) from daily statistics for trending YouTube videos in Kaggle Kaggle is a website that provides dataset for data scientists and machine learners. It allows us to download data sources in both CSV and JSON format. This dataset includes several months of data on daily trending YouTube videos from regions such as the USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India. Data from each region is stored in a separate file. Data includes the video title,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

channel title, publish time, tags, views, likes and dislikes, description, and comment count. The excerpt of the dataset is shown in Fig. 3

	Video id	Channel title	category id	publish time	views	likes	dislikes	subscribers	comments count
0	78qgdv2aYel	gk duniya	Education	27-Jan-21	3	0	0	1000	0
1	JXEmLLHYvqk&t=23s	gk duniya	Education	27-Jan-21	4	0	0	1000	0
2	nlfNNwn6xzU	gk duniya	Education	27-Jan-21	0	0	0	1000	0
3	oYU3lmqMdlQ	gk duniya	Education	27-Jan-21	0	0	0	1000	0
4	mXpi_SRc2bM	gk duniya	Education	27-Jan-21	0	0	0	1000	0
...	...	...	...	...	...	...	...	...	...
108	SfOyAYvnjJA	gk duniya	Education	18-Dec-18	143	27	0	1000	0
109	7QxP_oXNZXg	gk duniya	Education	17-Dec-18	391	55	1	1000	0
110	etxm60K_VEQ	gk duniya	Education	17-Dec-18	192	33	0	1000	0
111	RAJvMiNvMc4	gk duniya	Education	17-Dec-18	219	39	0	1000	0
112	BkZrUM145nI	gk duniya	Education	17-Dec-18	618	46	0	1000	0

113 rows × 9 columns

Figure 3: Dataset

VII. REGRESSION

Regression is that the construction of an efficient model to predict the dependent attributes from a bunch of attribute variables and the output variable is either real or a continuous value i.e salary, weight, area, etc.

We can also define regression as a statistical means applied mathematics implies that is used in applications like housing, investing, etc.

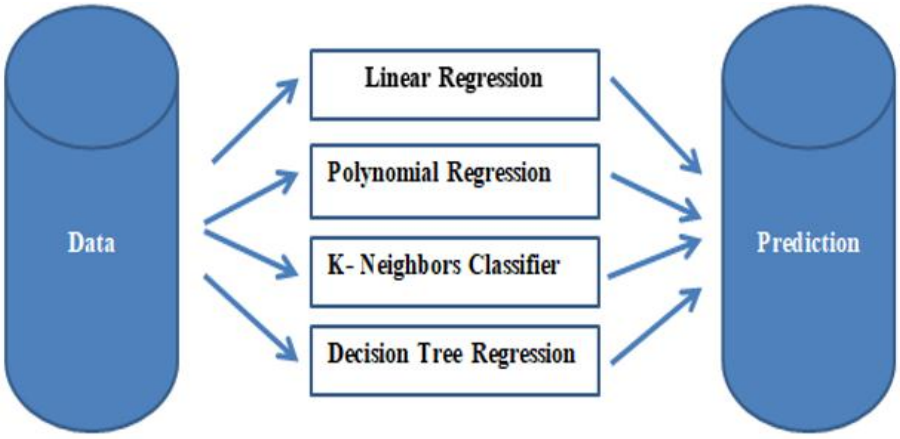

Figure 4: machine learning algorithms

1. Linear Regression

Linear Regression is an attractive model because the representation is so simple. The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

Now that we understand the representation used for a linear regression model, let's review some ways that we can learn this representation from data.

2. Polynomial Regression

In simple linear regression algorithm only works when the relationship between the data is linear but suppose if we have non-linear data then Linear regression will not capable to draw a best-fit line and it fails in such conditions. consider the below diagram which has a non-linear relationship and you can see the Linear regression results on it, which does not perform well means which do not come close to reality. Hence, we introduce polynomial regression to overcome this problem, which helps identify the curvilinear relationship between independent and dependent variables.

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

Suppose we have X as independent data and Y as dependent data. Before feeding data to a mode in preprocessing stage we convert the input variables into polynomial terms using some degree.

Consider an example my input value is 35 and the degree of a polynomial is 2 so I will find 35 power 0, 35 power 1, and 35 power 2 And this helps to interpret the non-linear relationship in data. The equation of polynomial becomes something like this.

3. K-Neighbors Classifier

K-nearest neighbors could be a straightforward algorithmic program that stores all accessible cases and classifies new cases supported a similarity measure k-nearest neighbors' algorithmic program (k-NN) could be a non-parametric technique used for classification and regression. In each case, the input consists of the k-nearest coaching examples within the feature area. The output depends on whether k-NN is employed for classification or regression: In k-NN classification, the output could be a category membership. An object is assessed by a plurality vote of its neighbors, with the item being appointed to the category commonest among its k nearest neighbors (k could be a positive whole number, usually small). If k = 1, then the object is solely appointed to the category of that single nearest neighbor. In k-NN regression, the output is that the property price for the item. This price is that the average of the values of k nearest neighbors. K-NN could be a variety of instance-based learning, or lazy learning, wherever operate is simply approximated regionally and every one computation is deferred till classification. The k-NN algorithmic program is among the only of all machine learning algorithms. Both for classification and regression, a helpful technique may be accustomed assign weight to the contributions of the neighbors, so that the nearer neighbors contribute additional to the common than the additional distant ones. As an example, a typical weight theme consists in giving every neighbor a weight of 1/d, wherever d is that the distance to the neighbor.

4. Decision Tree Regression

In this algorithm at a specific node, a split of knowledge happens. Thus, the most effective attribute to separate is to be known. Once the split for every price, a child node is formed. For every child node if the set is pure then it might stop otherwise a algorithmic split happens. This algorithm is prime down and goes on in dividend conquer manner. There are three varieties of nodes in an exceedingly call tree specifically root node, branch node and leaf node. Leaf node represents a category. Partitioning of knowledge is stopped once the subsequent conditions are satisfied.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

- Once all samples for a given node belong to identical category
- No attributes ought to be remained for. Further partitioning and so as classify leaf adopt for majority balloting is employed.
- No information samples ought to be left. Once the choice tree is made mistreatment that the testing set. It's foretold by traversing across the tree every, for every sample and selecting suitable price at each node.

Decision Tree Regression: Evaluation the Accuracy

- To evaluate the accuracy of the model.

Output:

```
In [66]: model = DecisionTreeRegressor()

In [67]: model.fit(xin,xout)
Out[67]: DecisionTreeRegressor()

In [68]: model.score(xin,xout)
Out[68]: 0.9998106095380962

In [ ]:
```

Figure 5: Decision Tree Regression Accuracy

CODE ANALYSIS

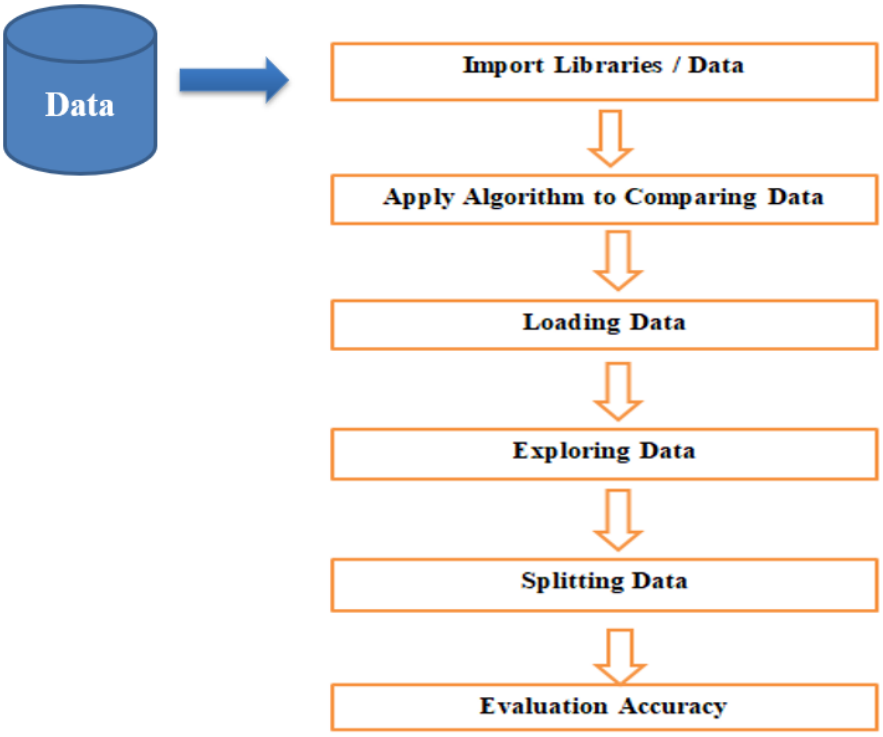

Figure 6: Code Analysis

VIII. FUTURE SCOPE

There are many things we are able to do increase model performance and those can be listed as follows.

1. This data set doesn't contain channel specific data (subscriber count, etc) it's performance is low. We are able to add them and train once again.

2. Current view count continuously depends on the previous video data. have to be compelled to embrace the number of views get to the video previously uploaded.

3. Since we have a tendency to predicting future views it will be nice to own a dataset with the most recent

time, like 2019 and 2020.

4. Can scrape additional data and integrate them to form a better data set than the prevailing one.

5. Current data set include detail about trending video data, not sensitive to a small number of views predictions. Higher to scrape data related to normal videos and include them additionally. At last, what have to be need to confine mind is we used a data set of trending data for creating the model. Owing to that predicting view count only be perform well for channels that targeting trending. For normal channels and recent channels won't be able to get an honest performance here. But we can improve this model to that level

IX. CONCLUSION

- Henceforth, we conclude that, the work done fulfills objective of YouTube Views Prediction by involving multiple strategies like Linear Regression, Multi-Linear Regression, Polynomial Regression, KNN Classifier, Decision Tree Regression.
- Different strategies applied to process inconsistent data. Another insights provided featured graphs for acquiring better knowledge of the dataset and look at the relationships shared by the featured of the dataset .
- Although there is scope for improvement which would be gained taking help from the work done.

X. REFERENCES

[1] Yuping Li1, Kent Eng1, Liqian Zhang1 1Department of Civil and Environmental Engineering, Stanford University, Stanford, CA 94305..

[2] Allen Wang, Aravind Srinivasan, Kevin Yee and Ryan O'Farrell.

[3] https://github.com/nciganovic/Youtube-subscribers-prediction

[4] Henrique Pinto, Jussara M. Almeida, Marcos A. Gonçalves ComputerScience Department Universidade Federal de Minas Gerais, Brazil {hpinto, jussara, mgoncalv}@dcc.ufmg.br

[5] YouTube, "Press statistics," https://www.youtube.com/yt/press/statistics. html, 2015, [Online; accessed 19-October-2015].

[6] Facebook, "Company info," http://newsroom.fb.com/company-info/, 2015, [Online; accessed 06-October-2015].

[7] Instagram, "Press," https://instagram.com/press/, 2015, [Online; accessed 06- October-2015].

[8] Twitter, "Company info," https://about.twitter.com/company, 2015, [Online; accessed 06-October-2015].

[9] Adage.com, "Facebook 85 users creating content," http://adage.com/ article/digital/facebook-85-users-creating-content/236358/, 2015, [Online; accessed 06-October-2015].

[10] Twitter, "What fuels a tweet engagement," https://blog.twitter.com/2014/ whatfuels-a-tweets-engagement/, 2015, [Online; accessed 16-October2015].

[11] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in Proceedings of ACM SIGCOMM Conference on Internet Measurement, 2007.

[12] TechCrunch, "2015 ad spend rises to $187b, digital inches closer to one third of it," http://techcrunch.com/2015/01/20/2015-ad-spend-rises-to-187b-digitalinches-closer-to-one-third-of-it/, 2015, [Online; accessed 19-October-2015].

[13] N. Techblog, "Its all a/bout testing: The netflix experimentation platform," http://techblog.netflix.com/2016/04/its-all-about-testing-netflix. html, 2016, [Online; accessed 10-March-2016]. 49

[14] Intelligence, "Using dark posts to a/b test videos on facebook," http://intelligence.r29.com/post/130204487611/using-dark-posts-to-ab-testvideos-on-facebook, 2016, [Online; accessed 10-March-2017].

[15] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," Communications of the ACM, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[16] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. L. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos." Performance Evaluation, vol. 68, no. 11, pp. 1037–1055, 2011.

[17] R. Bandari, S. Asur, and B. A. Huberman, "The Pulse of News in Social Media: Forecasting Popularity," CoRR, vol. abs/1202.0332, 2012. [Online]. Available: http://arxiv.org/abs/1202.033