YouTube Videos - Likes and Views Prediction

Vikas Ranjan

DSC680, Summer 2021

Bellevue University, NE

Abstract

YouTube is an online video sharing and social media platform. It is the world's second largest search engine and second most visited site after Google. In fact, 37% of all mobile internet traffic belongs to YouTube. It is the second most popular social media platform with almost 1.9bn users. More than 500 hours of video are uploaded to YouTube every minute. We watch over 1 billion hours of YouTube videos a day, more than Netflix and Facebook video combined. It has not only transformed the music industry, but on a broader level it has given power to its views. Most content is generated by individuals, including collaborations between YouTubers and companies that sponsor them. The contents on YouTube range from music videos, video clips, short films, feature films, documentaries, audio recordings, corporate sponsored movie trailers, live streams, vlogs, as well as content from popular YouTubers. Content creators (YouTubers) are being paid to create and upload videos which is based on the number of subscribers they have. To determine the year's top-trending videos, YouTube uses a combination of factors including measuring user's interactions such as number of views, shares, comments and likes. It is remarkable to note that they are not the most-viewed videos overall for the year. Top videos on the YouTube trending list are music videos, celebrity and/or reality TV performances.

Method:

The project is carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains 6 steps that will be followed throughout the project.

- **Business Problem** - YouTube has not only become a great alternative to traditional media but with billions of hours of content and countless number of corporations and groups, YouTube has also transformed into a tool for social impact. These interesting facts

and social changes promoted me to take this topic as my project 1 for this course. I'll be looking for the trends in the data related to the uploaded videos US and see how various attributes are correlated. I would also perform basic text analysis on 'Description' of the videos and build a word cloud to see which are most used words in the description of the uploaded videos. I'll also be building a model to predict the number of likes a video might get and another model to predict the number of views an uploaded video might get. This analysis and modeling are intended to help an influencer or content creator access their efforts effectively.

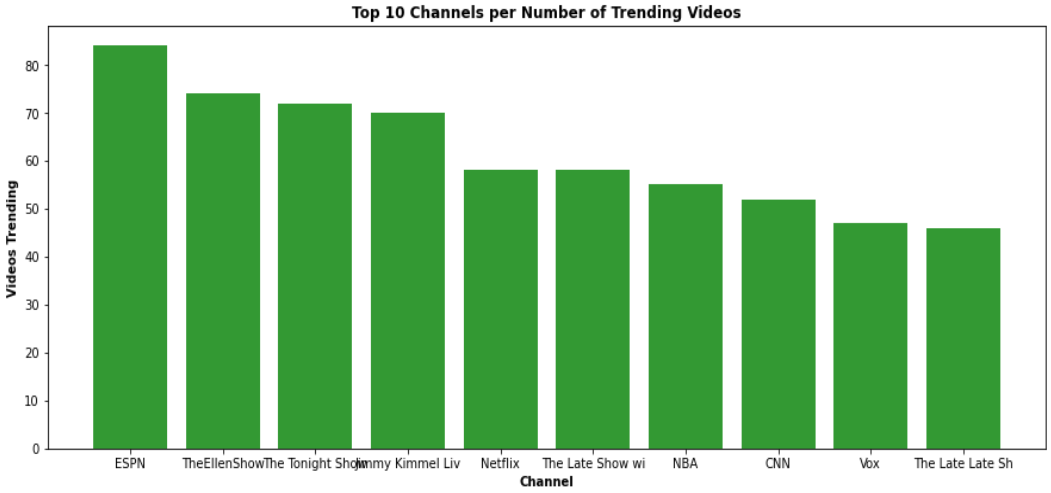
- **Data Understanding** - The dataset contains details of the videos uploaded on YouTube in US during 2018 and 2019. The dataset was extracted from Kaggle from the URL - https://www.kaggle.com/datasnaek/youtube-new

  This dataset includes 40950 rows and 16 feature variables. Each row corresponds to a unique video, and includes the following attributes:

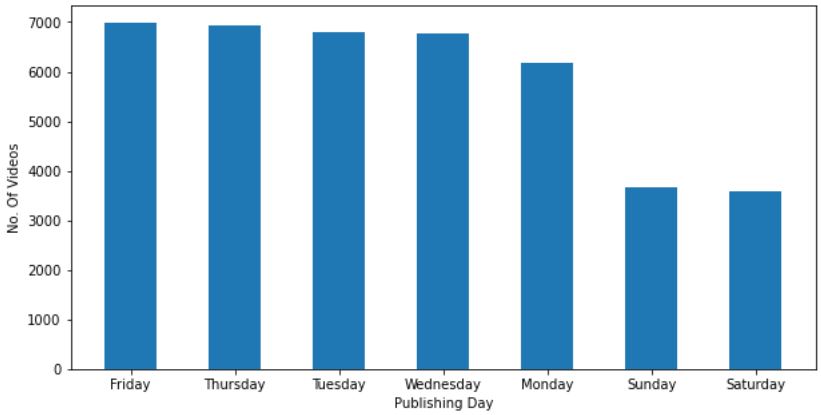
  1. Video_id
  2. Trending_date
  3. Title
  4. Channel_title
  5. Category_id
  6. Publish_time
  7. Tags
  8. Views
  9. Likes
  10. Dislikes
  11. Comment_count

12. Thumbnail_link
13. Comments_disabled
14. Ratings_disabled
15. Boolean field for the error with the video.
16. Description

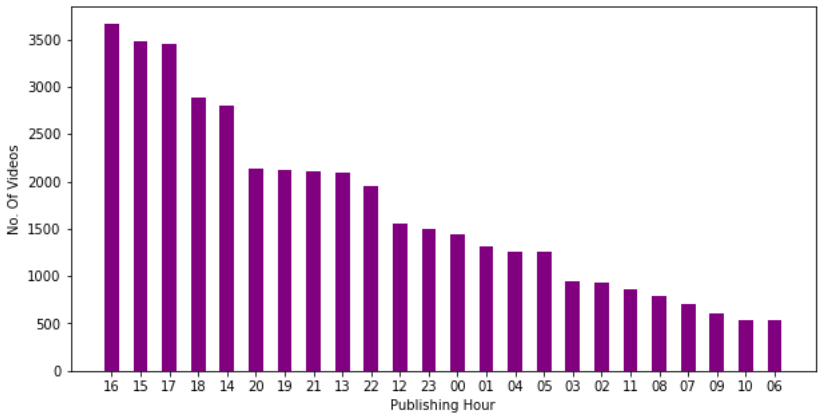
- **Data Preparation** – First step of data preparation was data cleanup.
    - Load the Category json file - Category ID and their respective description was there in a separate json file. Load the json file in the dataframe and subsequently added the Category description in the main dataframe.
    - Thumbnail_link attribute is of no value for this project, therefore removed it from the dataframe.
    - Description field had 570 null values which were replaced with blanks.
    - There were a few duplicates which were removed as well.
    - Formatted the trending_date and publish_time fields in the dataset to be in the desired format.
    - Performed cleanup on tags field and moved tags to a separate dataframe.
- **Exploratory Data Analysis**
    - Here we can see the days of the week which had the largest numbers of trending videos. There is a trend that on weekends, there are lesser videos being uploaded.



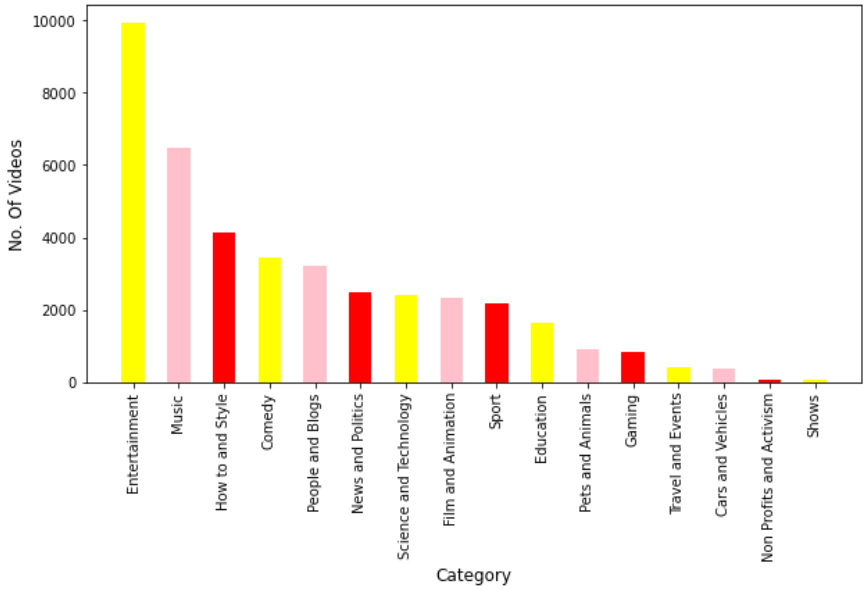
- Data and plot shows the trend that most of the videos are being uploaded on the weekdays, and over the weekends there is a significant drop in videos being published.



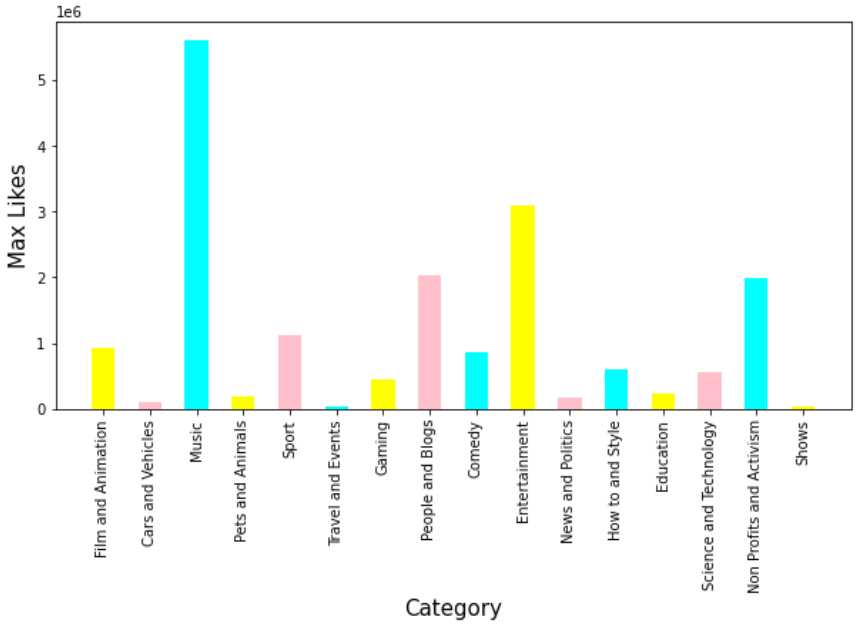
- Below plot shows the trend that most of the videos are being uploaded between 14:00 and 16:00, highest number of videos being uploaded are around 16:00.



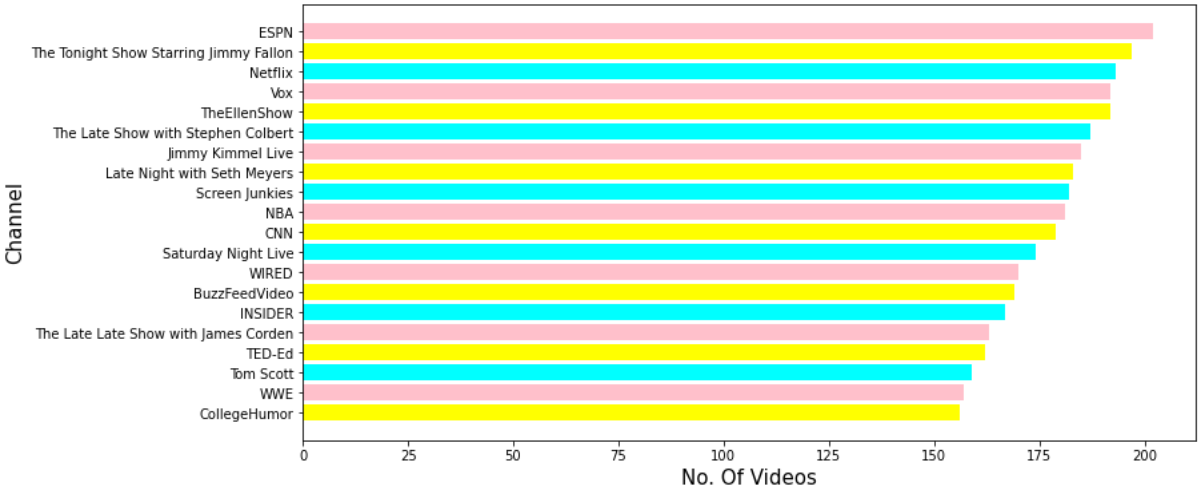
- Entertainment category contains the largest number of trending videos with around 10,000 videos, second is Music category with around 6,200 videos, followed by How to & Style category with around 4,100 videos.



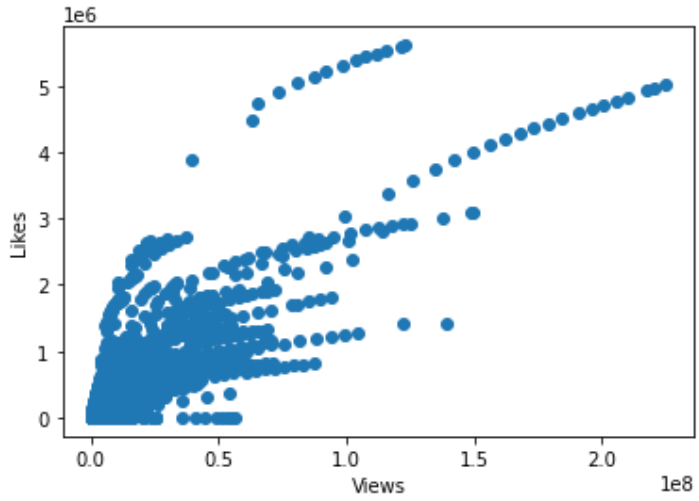
- As we can see in below plot, Music and entertainment videos have the most number of likes.



- ESPN seem to have the greatest number of videos, followed by late night shows, mostly in entertainment category.



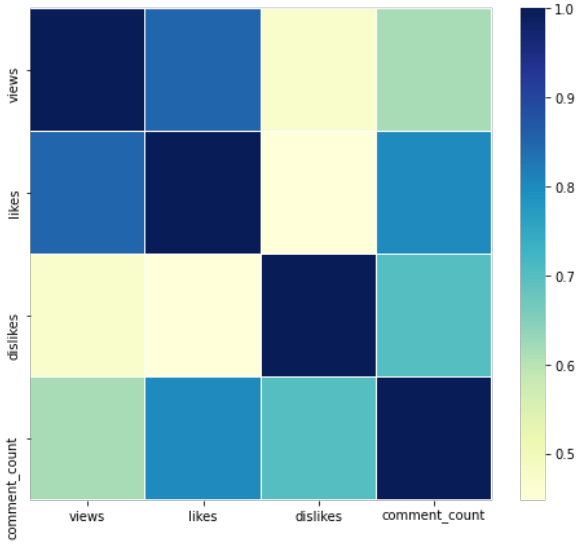
- Based on the below scatterplot, there seems to be positive correlation between the number of views and the no of likes of a video.



- Correlation matrix between views, likes, dislikes and number of comments on the video.

	views	likes	dislikes	comment_count
views	1.00	0.85	0.47	0.62
likes	0.85	1.00	0.45	0.80
dislikes	0.47	0.45	1.00	0.70
comment_count	0.62	0.80	0.70	1.00

- Heatmap of the correlation matrix:



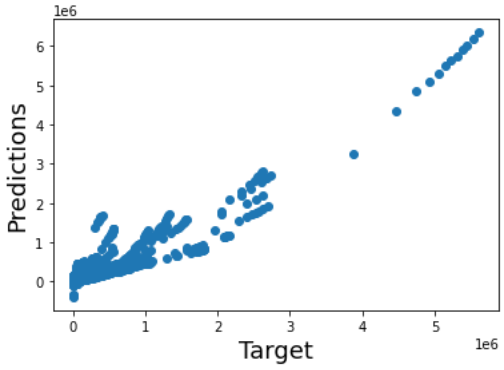
- Word cloud for the Description field, showing the most used words in the description.

- **Modeling & Evaluation** – As part of modeling, the intent is to build a model to predict number of likes and number of views on a video. Since both these values are continuous variables, therefore regression models are to be applied.

  Predict number of Likes:

I'm using linear regression model to predict the number of likes on a published video on YouTube. Validation of this model would be performed on its accuracy based on the r-square and mean absolute error.

```
Mean Absolute Error on Training Set:  30058.4653286869
Mean Absolute Error on Testing Set:  32831.88943833925
R-Squared Score on Training Set:  0.8793238222682671
R-Squared Score on Testing Set:  0.887092843540879
```



Looking at above results for predicting likes using linear regression, R-Squared score of .89 is fairly accurate.

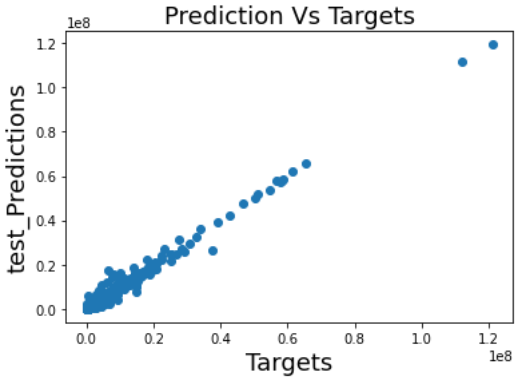
Predict number of Views:

Applied Linear regression to predict number of views, and the R-Squared score of .76 is not a very good score to go with.

```
Mean Absolute Error on Training Set:  979261.5911413658
Mean Absolute Error on Testing Set:  1017251.9907693673
R-Squared Score on Training Set:  0.7632860415673052
R-Squared Score on Testing Set:  0.7626145637025763
```

Linear regression results shows that r-squared score of 0.76 which is not accurate enough and had scope for improvements. Therefore, we will apply Random Forest Regressor model and evaluate if it is a better fit to predict number of views.

Random Forest Regressor model to predict number of views has a mean absolute error (MAE) of 220924. That means that 220924 is the amount of error in your measurements. It very much represents the difference between the measured value and "true" value. The score for Random Forest Regressor model is 0.985 which is a good score to go with.



Conclusion

- Linear Regression is a good fit when modeling to predict number of likes.
- To predict the number of views, Random Forest Regressor has better score than linear regression.
- Positive correlation between the number of views and the number of likes of a video. Likewise is the case with the number of views and the no of comments on a video.

Acknowledgements

We are indebted to the communities behind the multiple open-source software packages on which we depend. We would like to thank our families for their understanding of our time in this endeavor. Last by not least, thanks to Prof. Catherine Williams and our classmates for their guidance and feedback.

Appendix:

1. J, M. (2019, June 3). *Trending YouTube Video Statistics*. Kaggle. https://www.kaggle.com/datasnaek/youtube-new.
2. Srinivasan, A. (2017, December 17). Youtube Views Predictor. Medium. https://towardsdatascience.com/youtube-views-predictor-9ec573090acb.
3. Dulanjani, Y. (2021, May 18). YouTube View Prediction with Machine Learning. Medium. https://medium.com/analytics-vidhya/youtube-view-prediction-with-machine-learning-fdd4f40f352d.
4. Wikimedia Foundation. (2021, June 24). YouTube. Wikipedia. https://en.wikipedia.org/wiki/YouTube.
5. Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda
6. Machine Learning with Python Cookbook, Chris Albon
7. 57 Fascinating and Incredible YouTube Statistics. Brandwatch. (n.d.). https://www.brandwatch.com/blog/youtube-stats/.
8. Denison, C. (2020, April 23). YouTube turns 15: How it Has Changed the World Forever. Digital Trends. https://www.digitaltrends.com/features/how-youtube-has-changed-the-world-in-15-years/.