Foundations of Machine Learning
Project Report for
YouTube India Data Exploration

Anand Bhararia - 193050077
Rohit Kumar Singh - 193050069
Basant Kumar Bhala - 19305R006
Jayabrata Das - 193050085

November 24, 2019

Contents

1 Goals of the project

This is a report for predicting YouTube Like & View Counts using Machine Learning Techniques. This report also allow us to analyze the fundamental variables that affect the virality of the video.
It also contains the details for various processes used for the task which include data collection and scraping, data cleaning, data analysis, feature engineering, feature selection and other techniques.

While development of the project, we also took in-depth analysis of the various non-trivial parameters like timing of publishing video, creator's identity, categories, average time interval, waiting period, dislikes, tags with respect to pattern for trending videos and tried to represent the work in user friendly graphical representation.

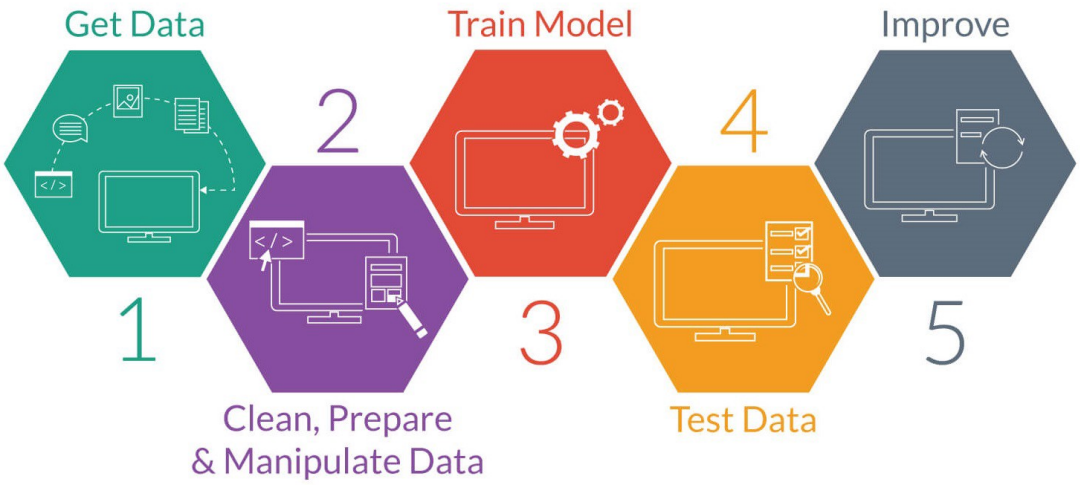

Figure 1: Step by Step Project Development

2 Related Literature

Youtube is one of the largest video-sharing website with humongous amount of videos on it. The site allows users to upload, view, like/dislike, share, add to favorites, report and comment on videos. There is a huge opportunity for analysing data present on YouTube and getting useful insights out of it.
The data-set includes several months (and counting) of data on daily trending YouTube videos with up to 200 listed trending videos per day. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

DATASET INFO
Youtube V3 API and web scraping were the two important tools which were used for the data collection. Following are the most relevant data attributes related to the videos that were collected for all the videos using the API and web scraping.

- Video id
- Trending date
- Title
- Channel title
- Category id
- Publish time
- Tags
- Views
- Likes
- Dislikes
- Comment count
- Thumbnail link
- Comments disabled
- Ratings disabled
- Video error or removed
- description

RELATED WORK
There have been several studies conducted on YouTube due to the fact that it is (one of) the most popular video sharing website(s). These studies have focused on different characteristics of videos and user profiles. Some of them include :-
The YouTube Social Network
Author: Mirjam Wattenhofer, Roger Wattenhofer, Zack Zhu

Trending Videos: Measurement and Analysis
Author: Iman Barjasteh, Ying Liu, Hayder Radha

3 Approaches tried

1. We firstly tried to build this project with high bias model i.e SVM with RBF kernel. The result of the SVM model was quite poor with 60% accuracy. The time consumption for training and tuning the SVM model was huge, so we switched to a different model with lower bias.

2. We started with an extracted feature from the title and thumbnail:
   Clickbait Score
   We were interested in seeing if YouTubers used varying levels of "clickbait-y" titles on their videos. So we calculated the difference in clickbait scores across videos. Then after some analysis, we noticed that there is not much correlation between the view count and the clickbait score implying that clickbait probably isn't a prerequisite for virality.

3. We tried Random Search for hyperparameter tuning and used different attributes at initial stages. But after proper analysis by producing heatmap of correlation table, we narrowed down our approach to few useful parameters and ends up selecting Grid Search for hyperparameter tuning.

Model used in Project

Random Forest Regressor was chosen as the learning algorithm for modelling the Like & View counts predictions. It is an ensemble method where multiple base estimators(tree) are trained on sub-samples of input data and give output after averaging the result of all estimators. Considering the size of the dataset,computational power available and ability of estimator to fit data, this model was considered.
The parameters of an algorithm always have an effect on it's performance. Grid Search and Cross Validation were used to tune the parameters for the model.
The final tuned parameters were :

n_estimators	200
max_depth	25
min_samples_split	15
min_samples_leaf	2

4 Experiments

4.1 Code Description

Language and Environment
Python is our core language in which everything is build upon from ground. We exhaustively used different libraries for project.

- numpy
- pandas
- scipy
- seaborn
- matplotlib
- plotly
- sklearn
- pickle

Hardware Requirements: Any 4 (or Better) Core Processor, 3 GB or More RAM , No Specific Graphical Requirement

Software Requirements: Kaggle (for preprocessing and visualization), Python3 , Google API Key (for web scrapping)

How many lines of code? Nearly 250 lines of code.

What code did we start from? We learned from the various kernel available from this link. Trending YouTube Video Statistics Kernel list

Link for Code Repository - Youtube India Data Exploration

4.2 Experimental platform

Project is run on kaggle notebooks for preprocessing and standalone Linux platform for training & prediction.

4.3 Results

Likes for old dataset	0.93
Views for old dataset	0.91
Likes for new dataset	0.80
Views for new dataset	0.75

*All the numbers are in R^2 score in the above list

- On the basis of final result, this implication is clear that youtube data is highly temporal and there is a paradigm shift in pattern of trending videos which can be proved easily.

1. Data Visualization

   - Best Time to Publish Video

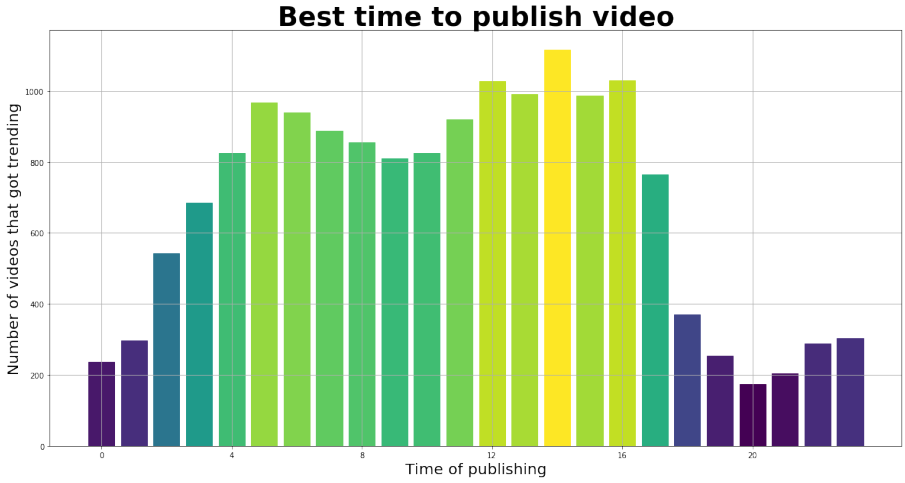

Figure 2: Best time to publish video

- Correlation between Dataset Attributes

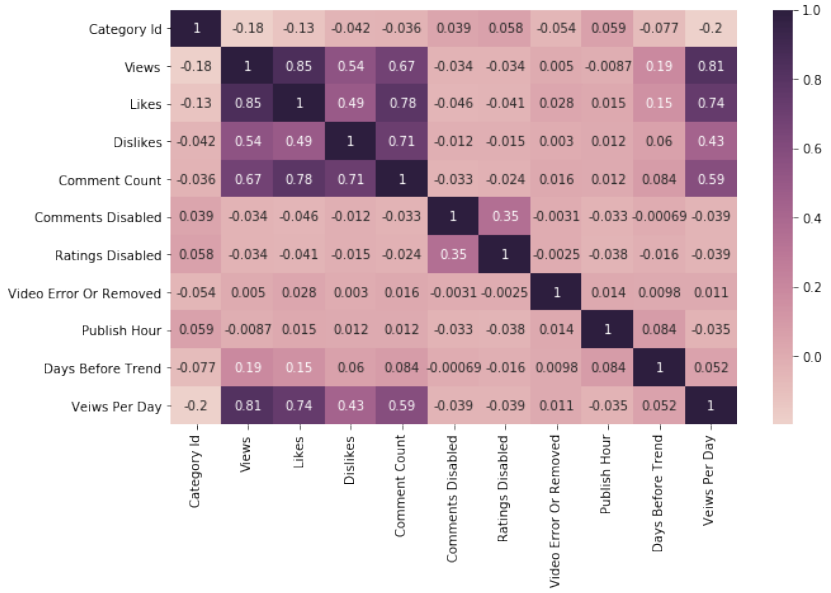

Figure 3: Correlation between dataset attributes

- Sensational Trending

Photo	Channel Name	Title	Category	Publish Date
	FoxStarHindi	Sanju | Official Teaser | Ranbir Kapoor | Rajkumar Hirani	Entertainment	2018-04-24
	Rhythm Boyz	Golak Bugni Bank Te Batua Full Movie (HD) | Harish Verma | Simi Chahal | Superhit Punjabi Movies	Movies	2018-05-31
	FoxStarHindi	Sanju | Official Trailer | Ranbir Kapoor | Rajkumar Hirani | Releasing on 29th June	Entertainment	2018-05-30
	ashish chanchlani vines	Restaurant Sutiyapa | Ashish Chanchlani	Comedy	2018-05-30

Figure 4: Sensational Trending

- Most influential creators (By Channel)

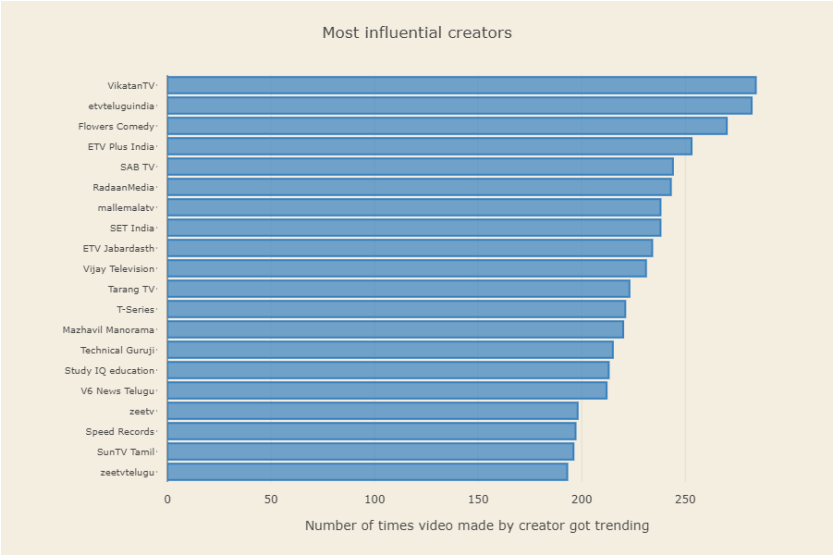

Figure 5: Most influential creators (By Channel)

- Most Popular Categories

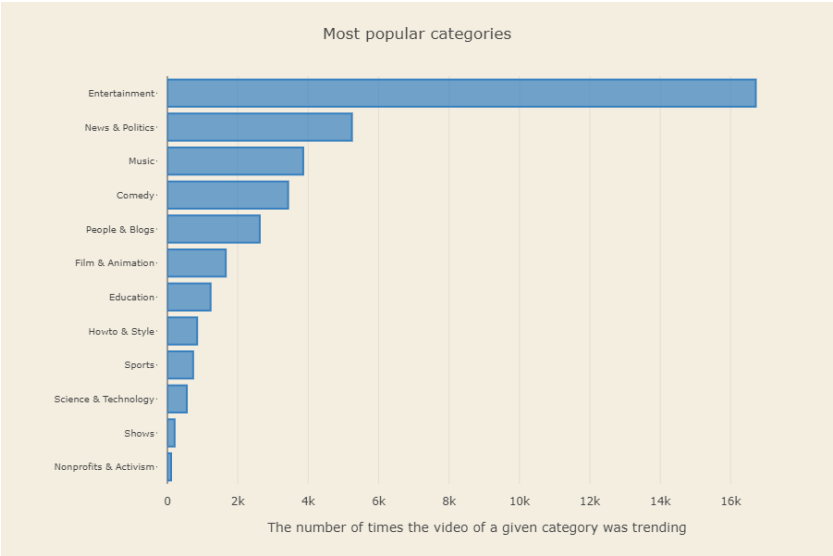

Figure 6: Most popular categories

- Average Time to Get Trending (by Category)

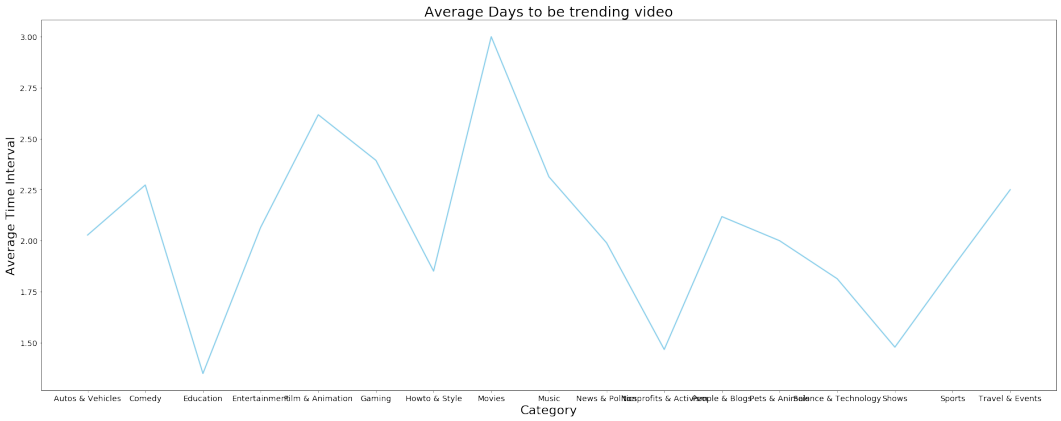

Figure 7: Average time to get trending (by category)

- Late Bloomers

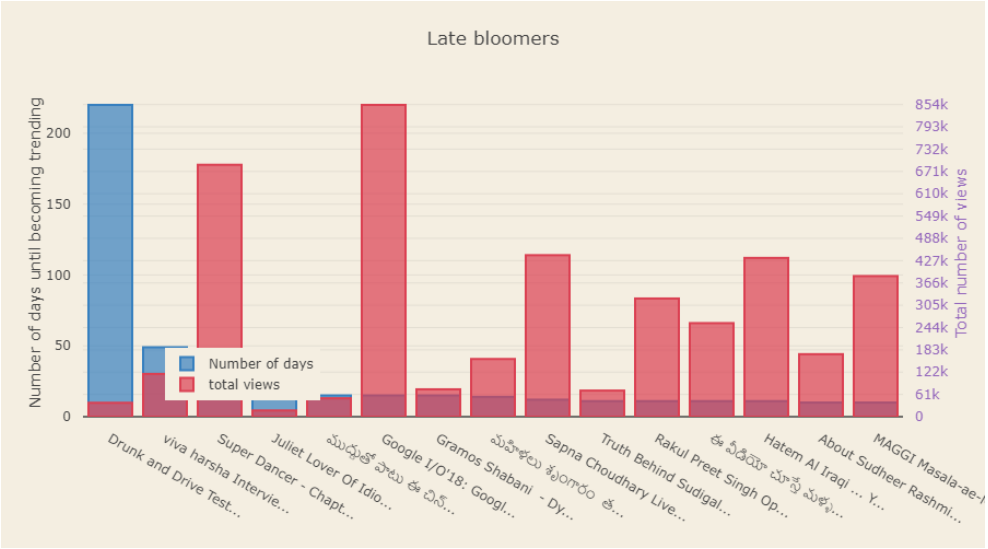

Figure 8: Late Bloomers

- Like vs Dislike

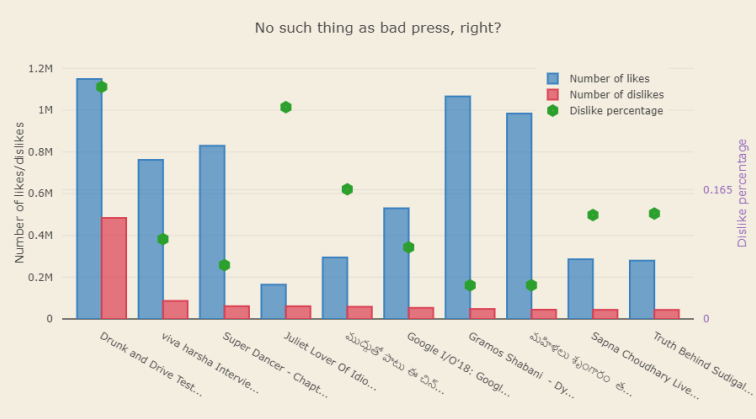

Figure 9: Like & Dislike

- Word Cloud for Tags

Figure 10: Word cloud

5 Effort

5.1 Time distribution

1. Web Scraping - 15%
2. Data Preprocessing - 20%
3. Model Selection & Training - 10%
4. Prediction & Testing on Test Data - 30%
5. Improve Accuracy - 25%

5.2 Challenges faced

1. Selection of machine learning model to suit the requirement of dataset
2. Training & Tuning the hyperparameter of the model
3. Selection of score to describe the accuracy of the model
4. Preprocessing & Refining of the Web Scraped Dataset

5.3 Contribution

1. Anand Bhararia - Development & training of machine learning model.
2. Rohit Kumar Singh - Data Preprocessing & Data Visualization
3. Basant Kumar Bhala - Report Development & Preliminary Data Survey
4. Jayabrata Das - Data Visualisation & Slide Presentation

6 References & Citations

1. Youtube Views Predictor by Aravind Srinivasan
2. YouTube Trending Videos Analysis by Ammar Alyousfi
3. Extensive USA Youtube [EDA] by Leonardo Ferreira