YouTube India

Data Exploration

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

(IIT Bombay Course Project)

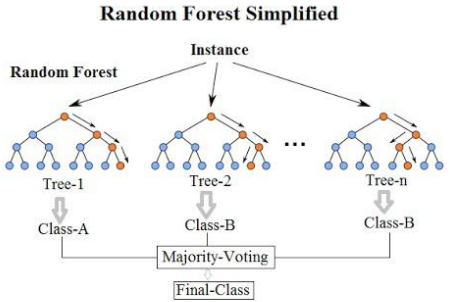
Goals of the Project

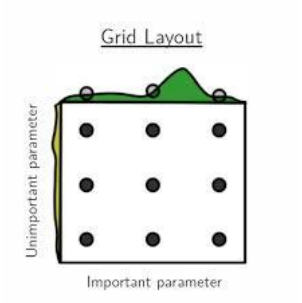
- Predicting YouTube video likes and views using Machine Learning models.
- Analyzing fundamental variables that affect virality of video.
- Exploring new non-trivial parameters like timing of publishing video, channel name, categories, average time interval, waiting period w.r.t pattern for trending.

Literature

- **Random Forest**

  Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.





- **Grid Search**

  Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

Dataset Attributes

- 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

Applied Libraries

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

Data preprocessing

- **For Training the Model:**
  - Datetime format of Trending date and Publish time
  - Removing Column *'Description'*
  - Addition of Column *'Category'*
  - Addition of column '*Days before trend'*
  - Addition of column *'Views per day'*
  - Replacing missing values for category columns by 'Nonprofits & Activism'

- **For Visualization Section:**
  - Dislike percentage
  - Number of words with all uppercase in title
  - Distribution of basic parameters
  - Removing duplicates

Experimented Approaches

- **SVM with RBF Kernel:**
  - Poor Accuracy [60%]
- **Clickbait Score:**
  - Not much correlation found between Clickbait Score and View Count
- **Random Search for Hyperparameter Tuning:**
  - Narrowed down approach to few parameter and finally selected using Grid Search.

Predictions

Models Used in Project

- We choose ***Random Forest Regressor*** for modeling the like and view count predictor.
- It is an ensemble method where large number of individual decision trees are trained on 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

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**

Github Repository - https://github.com/RohitLearner/Youtube_India_Data_Exploration/

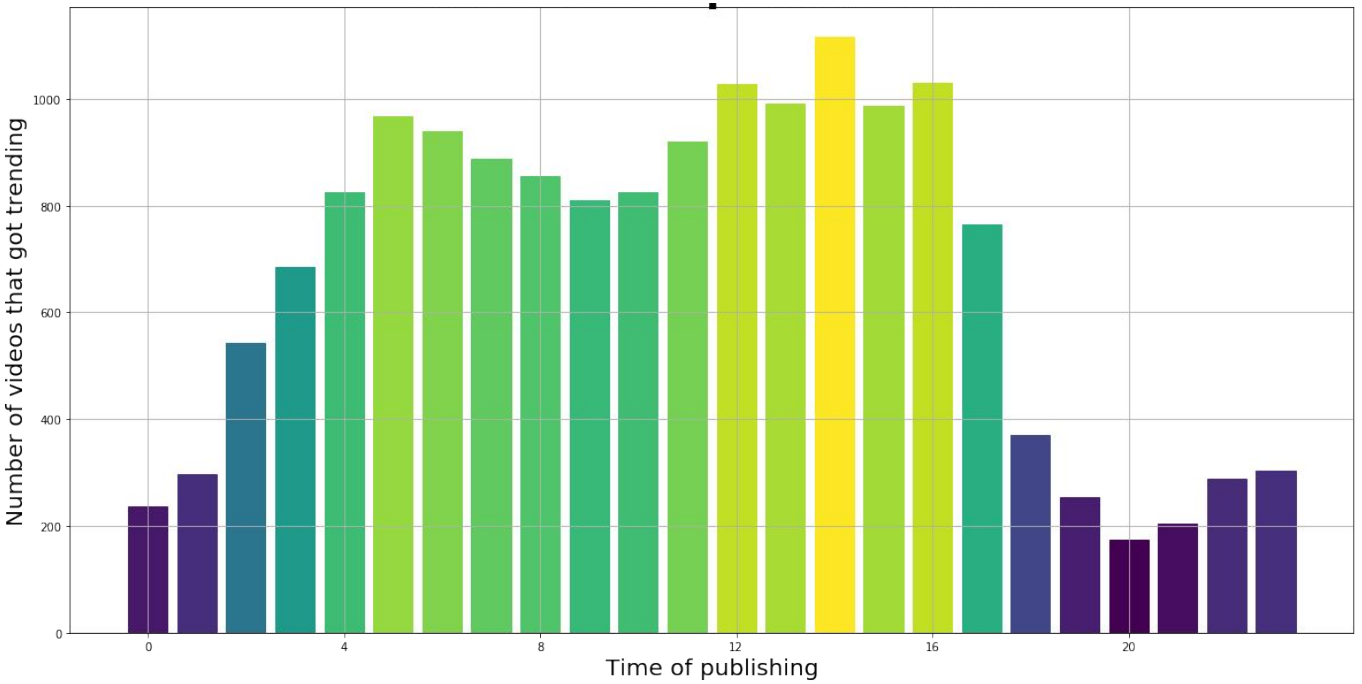
Conclusion

- On the basis of final result, we conclude that youtube data is highly temporal which can be proved by our data visualization.

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

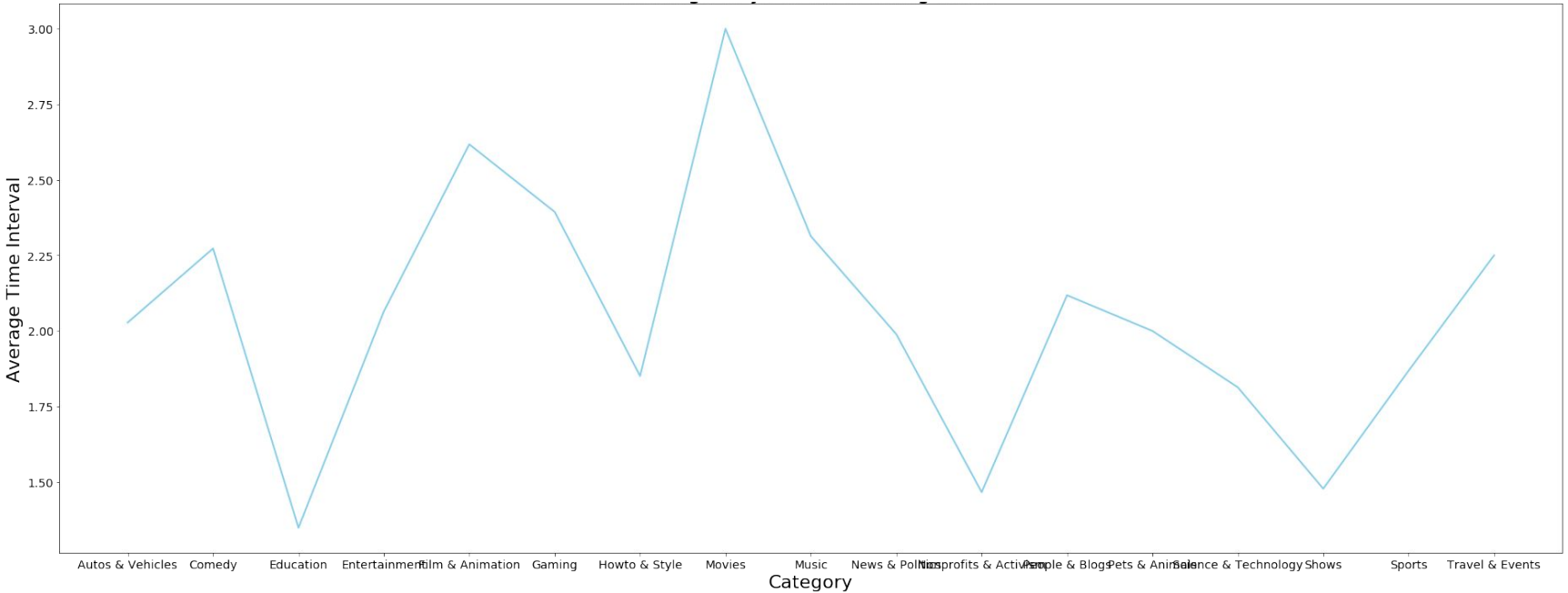
Data Visualization

Best Time to Publish Video



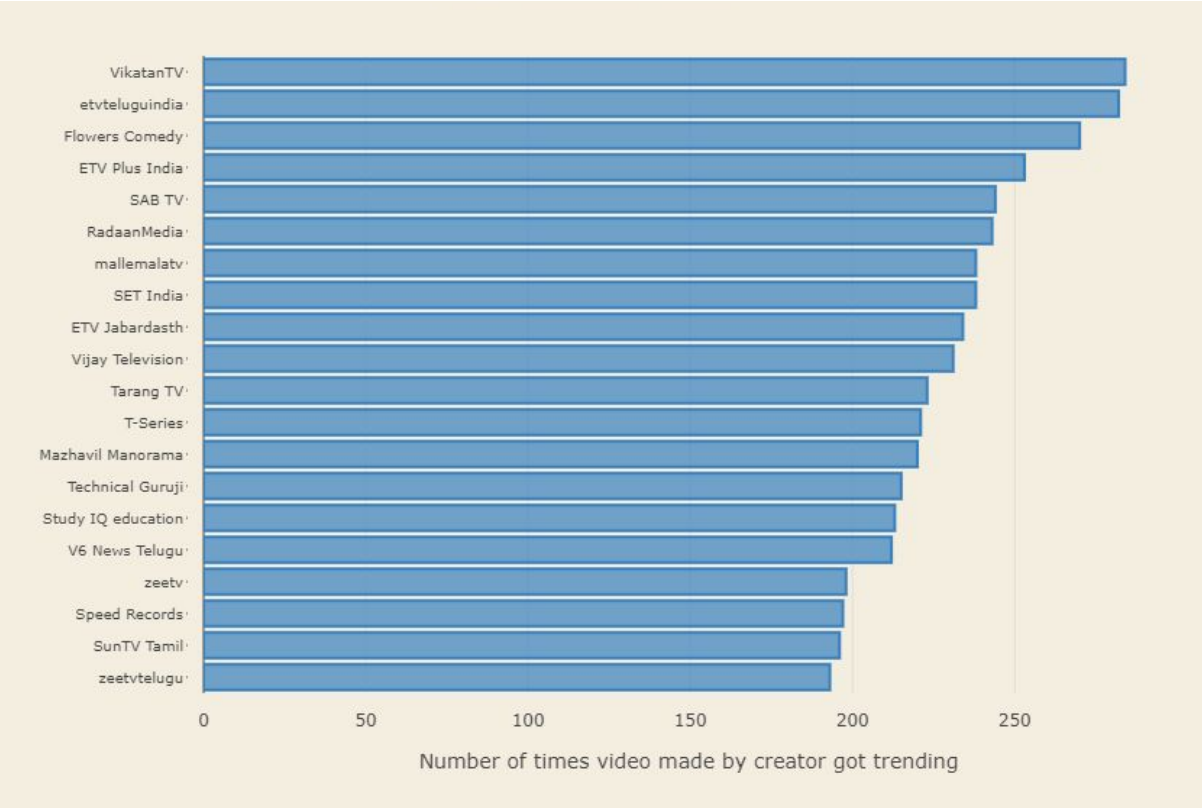
Correlation between Dataset Attributes

Average Time to Get Trending (by Category)

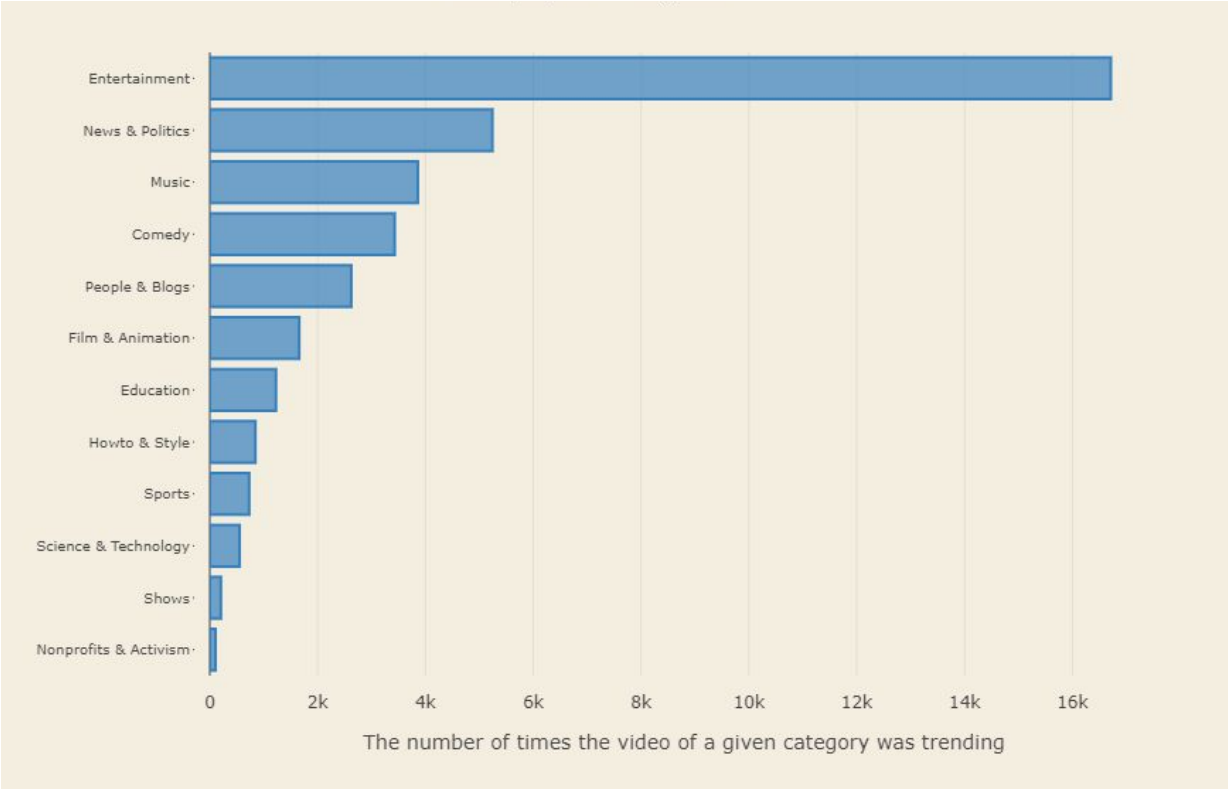


Sensational Trending

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

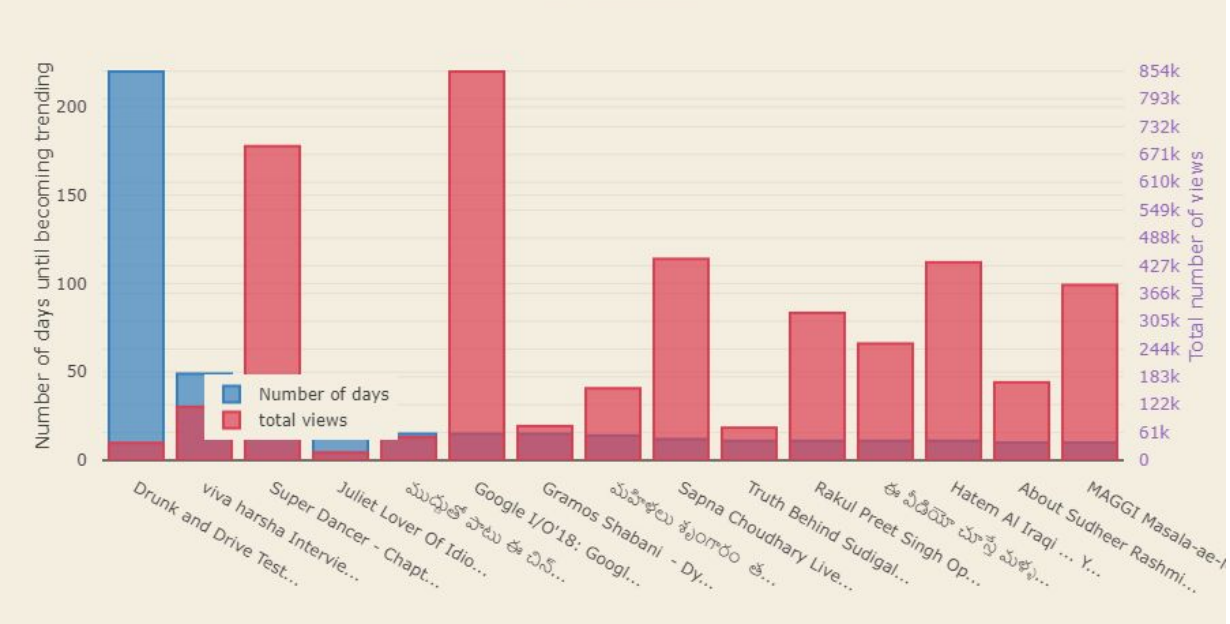


Most influential creators (By Channel)

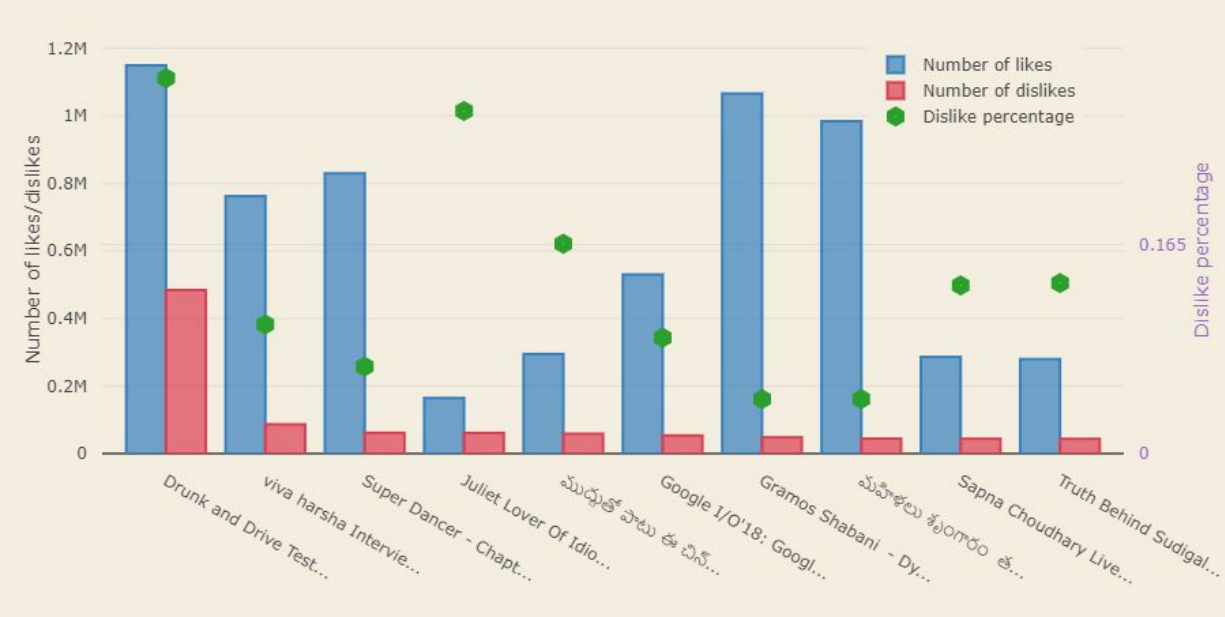


Most Popular Categories

Late Bloomers

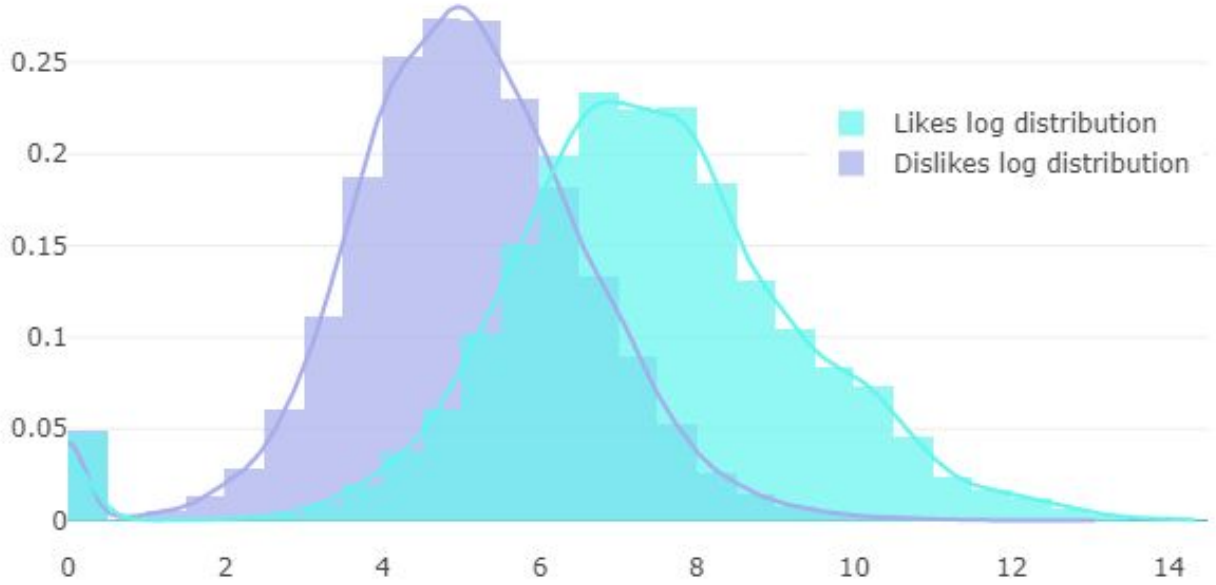


Like vs Dislike



Word Cloud for Tags

Like and Dislike Distribution



Future Scope

Some more things that we could have tried if we had more time would include: -

- Applying sentiment analysis on comments to create a more robust “user profile” that could be used as a feature.
- Using sentiment analysis on comments to create a robust “reception” feature, (similar to like/dislike) which could then be predicted.
- Using complex lower bias neural network based models to predict.
- Training a CNN on the thumbnail images.

Thank You