YouTube Data Analysis and LikeCounts Prediction using Machine Learning

Ayush Singh
ayushkumarsingh97@gmail.com

1. Introduction

Youtube is one of the largest video-sharing website with humongous amount of video on it .The site allows users to upload, view, rate, share, add to favorites, report and comment on videos.There is a huge possibility for analysing data present on YouTube and getting useful insights out of it.This is a report for predicting YouTube Like Counts using Machine Learning Techniques.It contains the details for various processes used for the task which include Data Collection/scraping,Data Cleaning,Data analysis,Feature engineering,Feature slection and Modelling.

2. Dataset

Youtube API and web scraping were the two important tools which were used for the data collection.Youtube being an enormous base of video it was the first challenge to decide the time frame for which the video base needs to be created.
A base of video IDs were created ranging over last 7 years (2010-2016) and collecting approx. 22,000-24,000 videos ids for each category(Youtube has 15 video categories) totalling a base of ~3,50,000 video ids.
Following are the most relevant data attributes related to the video that were collected for all the videos using API and scraping.

Data	**Description**	**Type**
Like count	No. of likes	Number
Dislike count	No. of dislikes	Number
Comment Count	No. of Comments	Number
Category Id	Category of video(total 15)	Number
Duration	Duration of video	ISO 8601
Published at	When was the video published	ISO 8601
Video Title	Title of video	Text
View Count	No. of views of video	Number
Definition	Video quality(HD/SD)	Text
Dimension	2D/3D	Text
Licenced Content	Is video licenced(T/F)	Text
Embeddable	Is video embeddable(T/F)	Text

Caption	Does video have caption(T/F)	Text
Privacy Status	Public/Private	Text
Tags	List of tags assigned by the publisher	Text
Audio Language	Lang. assigned by publlisher	Text
Video Description	Description of video	Text
Subscriber Count	No. of subscribers of channel	Number
Channel View Count	No. of views of channel	Number
Channel Video Count	No.of videos in channel	Number
Channel Published at	When was channel published	ISO 8601
Channel Comment Count	No. of comment of channel	Number
Channel Description	Description of channel	Text
Channel Title	Name of channel	Text
Country	Location of publisher	Text
Social Links	No. of social links mentioned of the channel page.	Number

3. Feature Engineering

3.1 Deriving features

After creating a base with the video data further some relevant features were derived/engineered from the attributes present

- **Title Length** : Length of video title.The title should be informative,crisp and short ,to be easily found out.
- **No .of Tags** : No. of tags assigned by the publisher to video.Tags help in increasing the search rank.
- **Description length** : Length of description provided by publisher.It should be not to long so that it could show up in search results.
- **No. of tags in title** : How many tags/keywords are present in the video title.Relevant tags/keywords help increase the search ranking.
- **No. of tags in description** : No. of tags present in the video desription.
- **No. of links(http) present in the video description** : Its a good practise to include links to other websites before detailed description of video. “http” keyword was found in the description for this.
- **Video Month old** : How old is video(in months) was calculated from the published at data for every video.
- **Channel Video Month old** : How old is channel(in months) was calculated from the 'published at' data for every video.
- **Day of upload** : The day of upload – M,T,W,Th,F,Sat,Sun
- **VC/VM** : Ratio of “Video View Count” and “Video Month old”
- **SC/CVC** : ratio of “Channel subscriber Count” and “Channel Video count”

- **VC/CVM** : Ratio of “Video View Count” and “Channel Month old”
- **VC/T** : Ratio of “Video View Count” and (“Tags in title”+”Tags in description”)
- **CV/CVC** : Ratio of “Channel View count” and “Channel Video Count”

A Set of final features for further Data Analysis was formed after extracting relevant attributes from the initial base created and taking the derived ones.

3.2 Data Exploration/Analysis and Cleaning

After addition of some more features the Data was further analysed.**Visualize_ML**[1] (a self made library) is used for **Uni-variate,Bi-variate exploratory analysis** and **Visualization** for this task.
A total of **33** relevant features were divided into two sets for analysis, *Categorical* and *Continous* Variables.

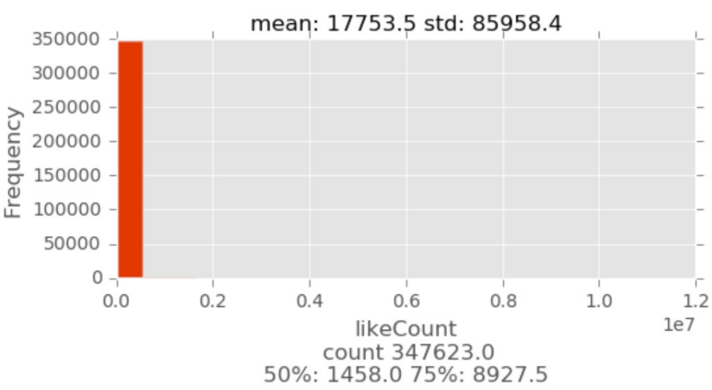
Continous Variables :

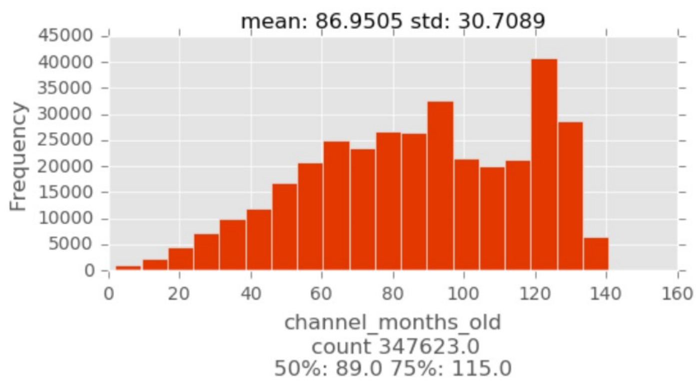
Likes Count, Comment Count, Dislike Count, Duration, ViewCount, Channel View Count, Channel Comment Count , Channel Subscriber Count, Channel Video Count, Description Length, Http in descp, Tags in descp., Video Title Length, Tags in Title, No. of tags, Channel Description Length, Video Months Old, Channel Months Old, Social Links ,VC/VM , SC/CVC , VC/CVM, VC/T , CV/CVC

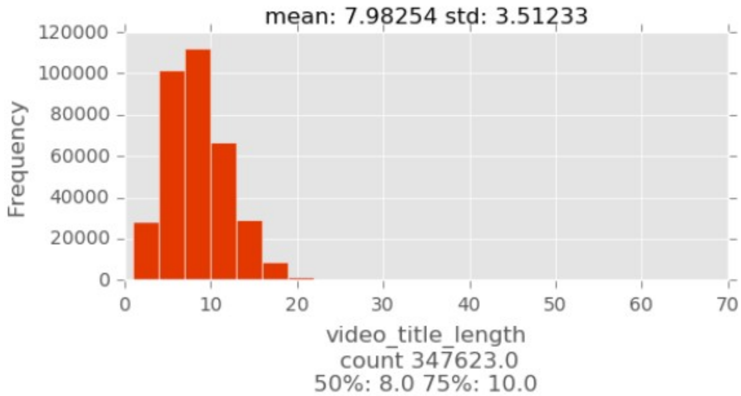
Categorical Variables :

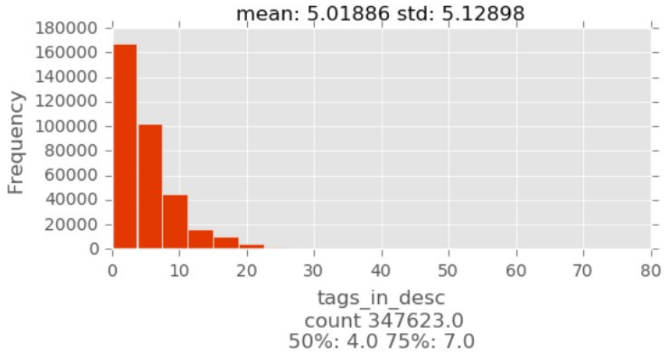
Caption, PrivacyStatus, LicencedContent, Embeddable ,Dimension, Definition, CategoryId, Day Uploaded,Country

Note : *Description, Tags, Title,PublishedAt,ChannelDescription, ChannelPublishedAt, ChannelTitle* were removed from the DataBase as they further didnt have any further relevance for analysis.









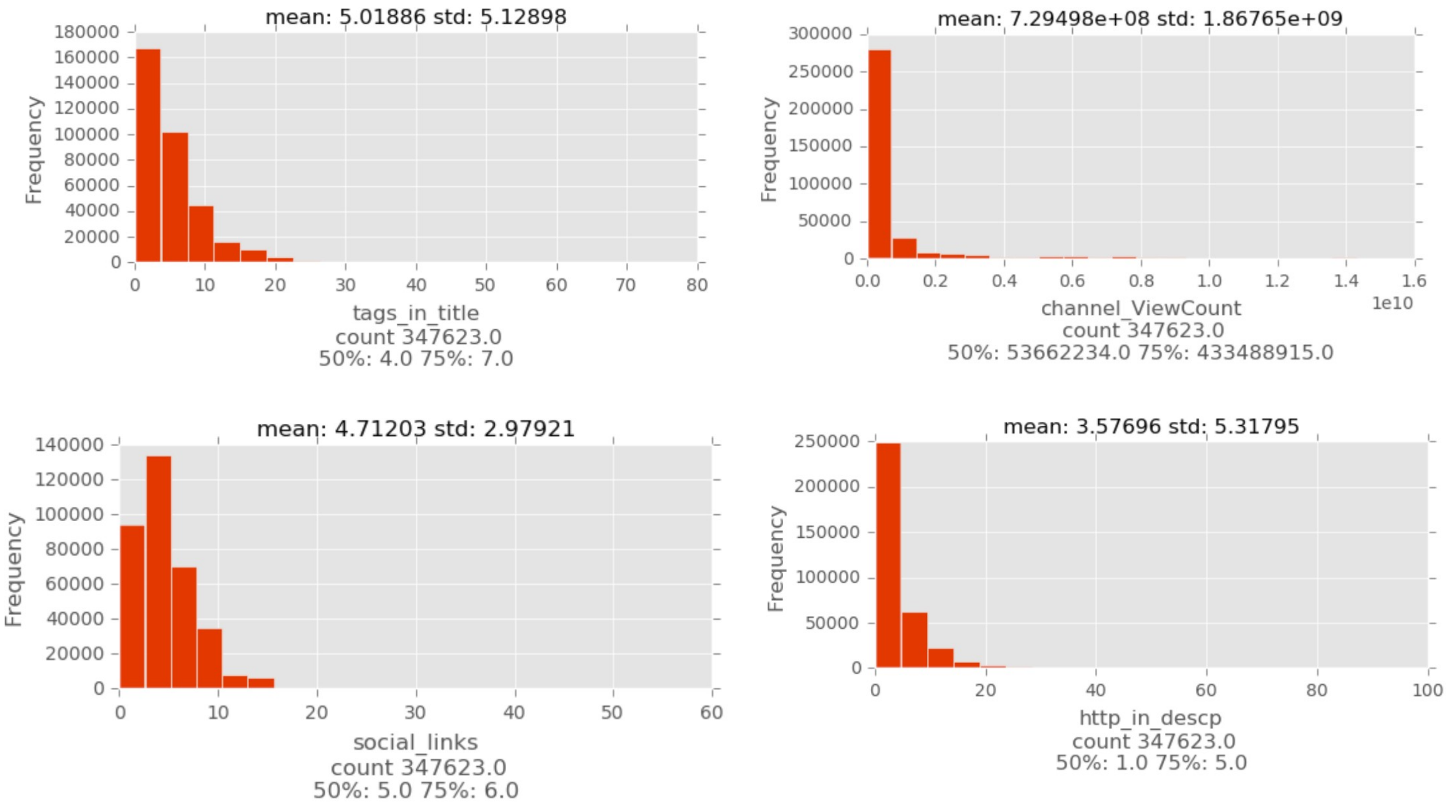


Fig1. Above are the univariate analysis plots of some of the continous features

	likeCount	dislikeCount	viewCount	commentCount	VC/VM	VC/T	SC/CVC	CV/CVC	VC/CVM
count	3.476230e+05	347623.000000	3.476230e+05	347623.000000	3.476230e+05	3.476230e+05	347623.000000	3.476230e+05	3.476230e+05
mean	1.775348e+04	583.892251	2.339295e+06	1902.862355	1.117118e+05	2.358704e+05	2900.721632	8.606676e+05	3.489580e+04
std	8.595840e+04	5298.228342	1.870051e+07	9232.638090	1.126983e+06	2.792814e+06	11858.969187	4.530922e+06	3.111542e+05
min	0.000000e+00	0.000000	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
25%	2.620000e+02	7.000000	5.283400e+04	34.000000	1.308000e+03	3.712000e+03	30.000000	1.595700e+04	6.221832e+02
50%	1.458000e+03	42.000000	2.422860e+05	212.000000	6.662000e+03	1.726900e+04	178.000000	8.116100e+04	2.986296e+03
75%	8.927500e+03	243.000000	1.099496e+06	1176.000000	3.827400e+04	7.897400e+04	1386.000000	4.010740e+05	1.446045e+04
max	1.112990e+07	938894.000000	2.110166e+09	938761.000000	1.960912e+08	8.198530e+08	436539.000000	2.891623e+08	5.537205e+07

Fig2. Summary Statistics of some of the Continous features

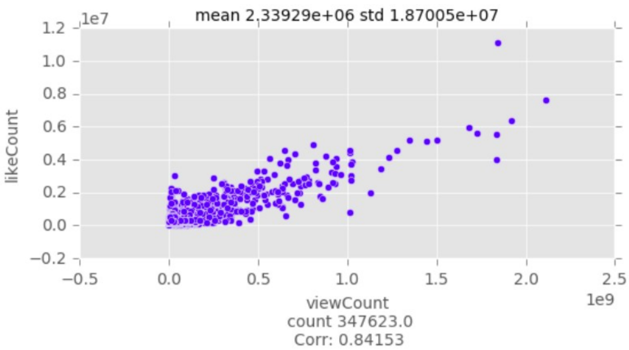
	categoryId	defaultAudioLanguage	definition	dimension	embeddable	licencedContent	privacyStatus	projection	caption	day
count	347623	63218	347623	347623	347623	347623	347623	347623	347623	347623
unique	15	89	2	2	2	2	2	2	2	7
top	1	en	0	0	1	1	0	1	0	5
freq	25778	39684	249789	347591	344605	222385	347618	347487	310277	57329

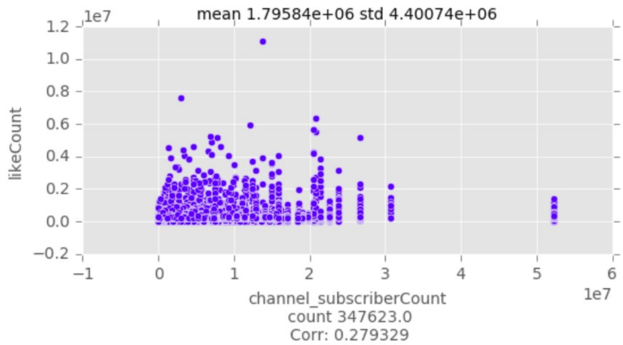
Fig3. Summary Statistics of Categorical features

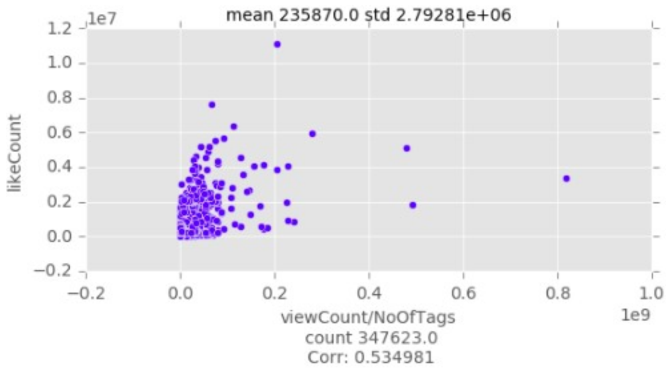
After the analysis features like *dafaultAudioLanguage* and *Country features* are removed due to large amount of missing data.

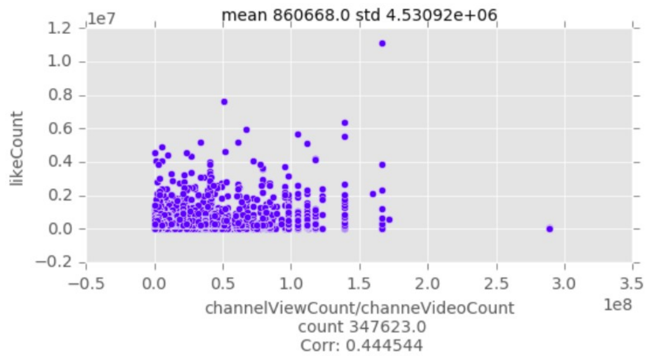
2.3 Feature selection

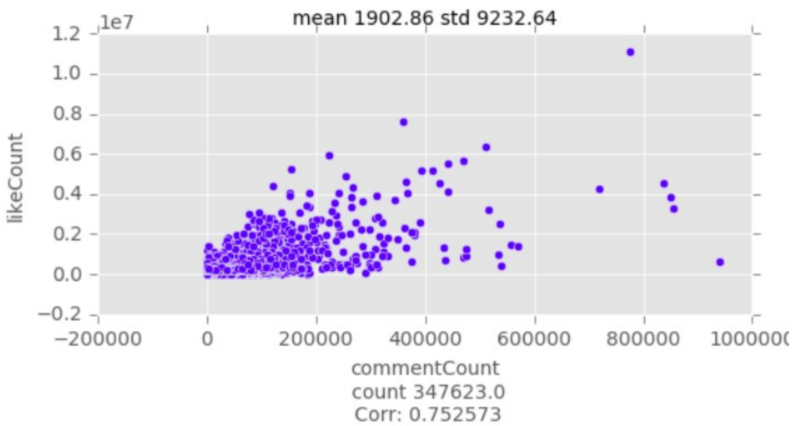
After the Univariate data analysis, Bivariate analysis was done between the the **Target variable-“Likes Count”** and the **predictors** to see the extent of correlation between them using Visualize_ML.Further after extracting the top correlated features from analysis, **RFE(Recursive Feature Elimination)** technique with **Random Forest Regressor** was used to extract final features to train the model.

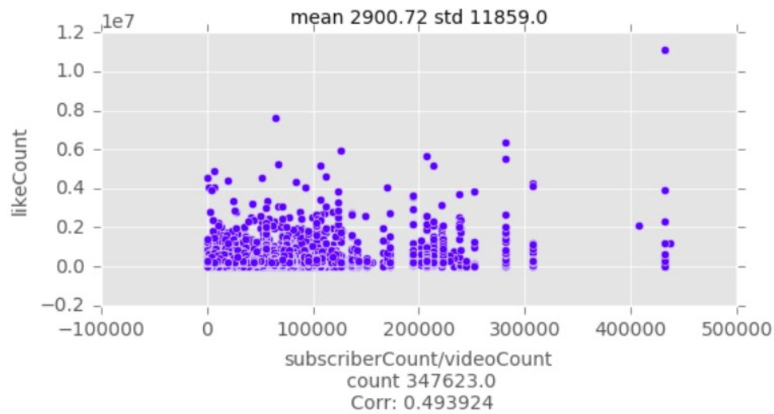


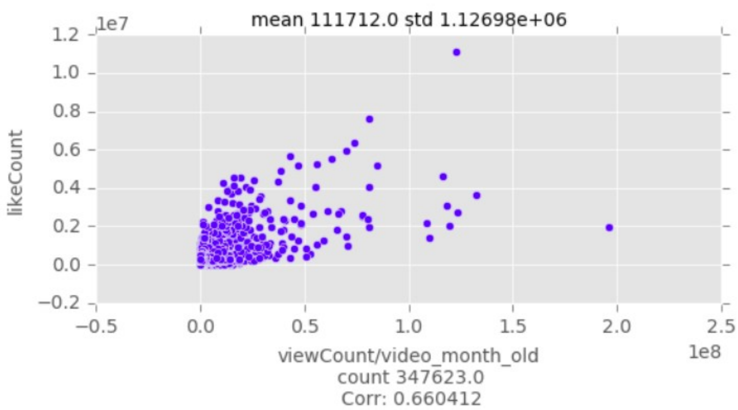


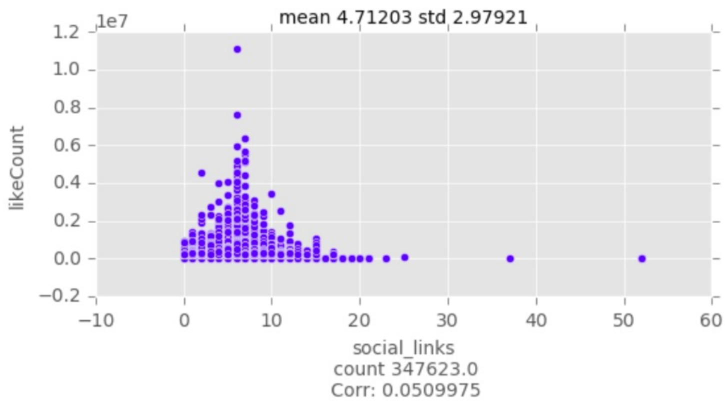












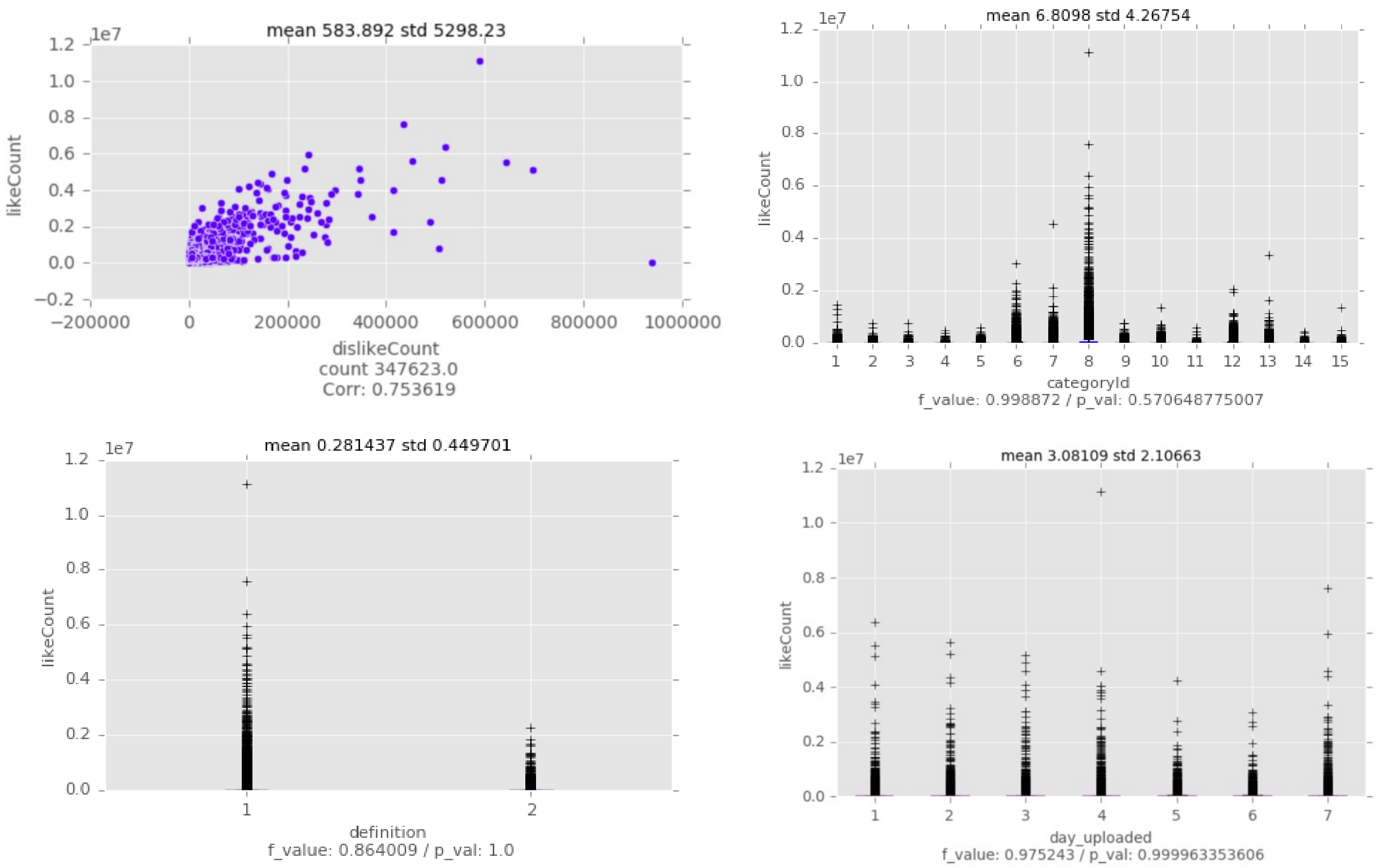


Fig4. Above are some of the the Bivariate analysis plots of continous and categorical features.Were Pearson correlation coefficient is calculated for Continous features and P value for catgorical using ANOVA test.

RFE on these features gave the feature rankings and top features were chosen for training of Machine Learning Model.

3. Model

Random Forest was chosen as the learning algorithm for modelling the Like counts predictions.It is an ensemble method were multiple base estimators(tree) are trained on subsamples of input data and give output after averaging the result of all estimators.Considering the size of 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 parametes for the model.

The final tuned parameters were :

n_estimators	200
max_depth	25
min_samples_split	15
min_samples_leaf	2

Final Features trained on : *ViewCount, CommentCount, DislikeCount, ViewCount/VideoMonthOld, SubscriberCount/VideoCount*

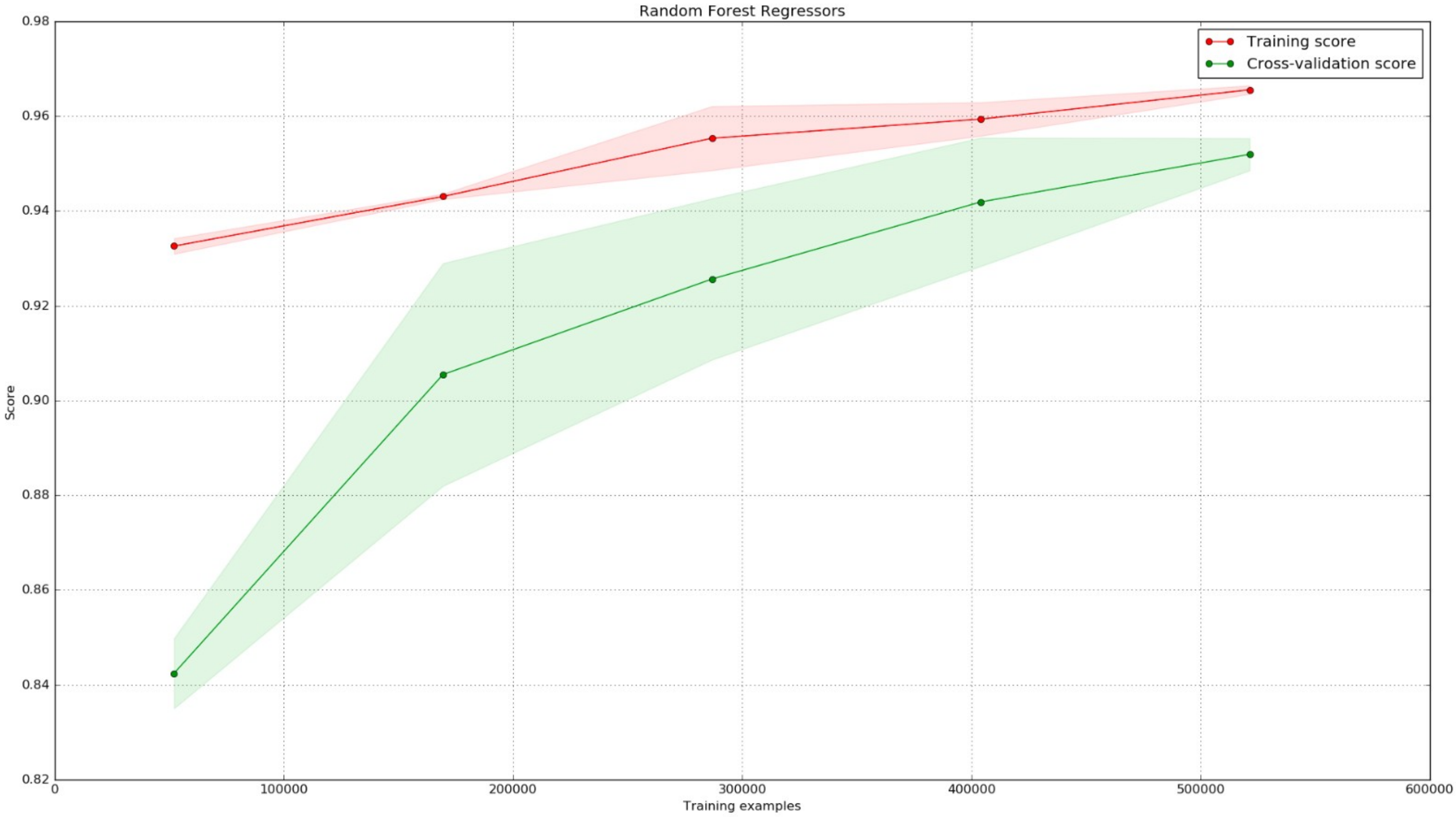


Fig 5. Gpraph shows the variation of R^2 Training and Cross-validation score with the training data over 2 epochs ~6,00,000 training examples

Evaluation metric : R^2 score
Cross-Validation score : 0.950340650248
Training Score : 0.96903254357

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2}$$

y_i : True value $\hat{y}_i$: Predicted value

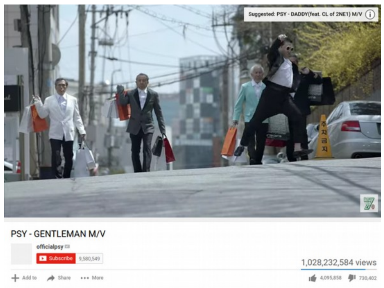
4. Conclusion

It was seen that Random Forest has fairly performed for this task of predicting Like counts for youtube videos.As per to my knowledge there hasn't been much research in this area and hence there is no benchmarks to validate the result againt.
Given more data and computational power this can have better resuts if trained on models like SVM(support vector machines) or Deep Neura Networks.

Prediction

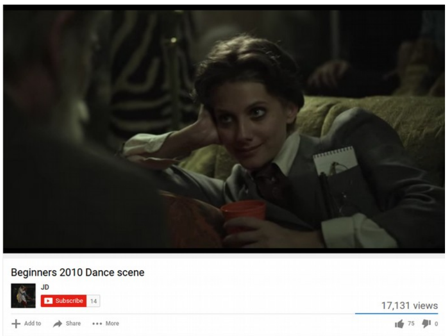
The model was run for predictions on some unseen data below are some of the predictions.

Id : dOyJqGtP-wU
True : 158014
Pred : 163751
Error : 3.630691



Id : ASO_zypdnsQ
True : 4095830
Pred : 3843383
Error : -6.163513

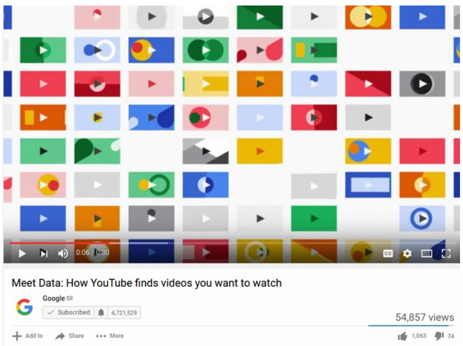
JdnuqdqLq-A 1063.0 1297.0 234.0 22.013170



Id : R5IzIUR3KP4
True : 75
Pred : 94
Error : 25.333333

Id : gAfFNMohv68
True : 10012
Pred : 11260
Error : 12.465042

Id : KQ6zr6kCPj8
True : 4350204
Pred : 4520800
Error : 3.921563



Id : KQ6zr6kCPj8
True : 4350204
Pred : 4520800
Error : 3.921563