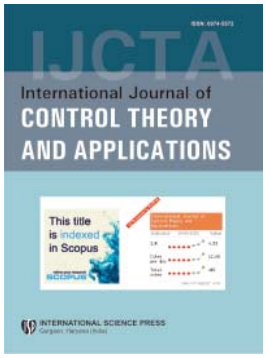

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 40 • 2016

A Predictive System to Predict the Number of Views and Viewer Retention on Videos Through Historical Data

Aman Dubey[a] Lokendra Singh[a] S. Karthick[b] and M.Uma[b]

[a]*B.Tech, Department of Software Engineering, SRM University, Chennai*

E-mail: amandubey199@gmail.com, lokendras395@gmail.com

[b]*Assistant Professor, Department of Software Engineering, SRM University, Chennai*

E-mail: karthik.sa@ktr.srmuniv.ac.in, umaprabhu78@gmail.com

Abstract: Nowadays, there are an enormous number of videos uploaded on Youtube, but the rate of its viewership is not at the same pace some get billions of likes and views and some with a little number of stats. Along these lines, ought further bolstering raise the individuals perspectives of the trademark What's more make it An additional staggering measure improved of the existing viewers of the channel , An predictive skeleton prerequisite been amassed which Might extend on the sure offers may suspect those people measure for sees besides for viewer backing those people remarks greater part of the information could a chance ought further bolstering a chance to be utilized Besides suspicion examination of the viewers might settle on passed on utilizing ordinary vernacular Processing(NLP) frameworks. Thus,the future fill in might include with stay with working on the routines alternately calculations alternately parameters In light of which exceptional prediction of the see checks Might be specified. The Dataset consists of a various number of features which by feature engineering has been subsetted based on its real statistical significance with the response variable View Counts. Our initial strategy involves usage of Linear Regression which will use feature matrix(collection of elements) which are statistically significant to predict the number of views as the response variable.But there are possibilities where the model won't come up with a better or even less accuracy, so the methods to deal with them involves normalization which is reducing the feature values with its values in range as mean value to be 0 and standard deviation to be 1. The opposite and only prediction includes use for an additional algorithm named Similarly as irregular woods Predictor , which is been viewed as Likewise the algorithm over which there is no such sort of acceptance slip , on in this algorithm fractional instances about information need aid utilized Similarly as testing situated , Also Previously, final one those imply of every one model would urge will provide for the expanded correctness , but still if we need aid not persuaded with the exactness from claiming our model we Might streamline it with consideration for tuning parameters.The comments dataset has been used to predict whether the same viewer would visit that particular channel to view that video again. Based on these classification results there would be a better serving of the existing viewers to retain their subscription and enhance the views and virality of the content.

Keywords: *Python, View Count Prediction, Machine Learning, Linear Regression, Random Forest, NLP, Naïve Bayes Classifier.*

1. INTRODUCTION

Features looking into Youtube bring a considerable measure for significance will the individuals channel managers whose ultimacy point will be to transfer elements from claiming distinctive genres Also guaranteeing that reality that the viewers need aid expanding around nonstop intervals , At it is really fundamental that ID number of the viewers which are hugely paramount to the channel are retained , Subsequently , will judge the viewers taste a predictive framework need to be been based In light of specific Characteristics like loves , dislikes, remarks , appraisals, kind need been used to anticipate those number of perspectives Also on anticipate their relationship it will be fundamental to arrange their sentiments In view of those content reviews which need to be been provided for of the particular provided for feature.

1.1. Dataset Description

The Dataset comprises about variables in feature id al-adha, Uploader, period of the feature, class, length, perspectives, rate, appraisals, remarks, basic number about remarks. Those Predictive model initializes for information importing in the python mediator , emulated by which it is exceptionally vital with investigate those information clinched alongside its cleanest manifestation to which information cleaning procedure need been instantiated which comprises of looking of any sort of forgetting qualities or labels , or those outlier qualities which Might aggravate a negative sway in the model building or tuning methods.. The other procedure involves Data Exploration part where some actionable insights has been visualized and analysed like counting the number of likes based on the number of likes or ratings . Pandas & Matplotlib Library is being used to cover up the exploration and statistical signifance between the predictor variables and the response variable view count . The next phase is the feature engineering phase which is considered to be the most scalable phase, since, this phase is important in terms for feature selection for robust model building . The Feature Engineering phase tests the variable signifance with the view count variable through the use of scatter plots, boxplots, Histograms, etc. Then the Part of Model Building where model is fitted on predictor variables present in the feature matrix and the response variable view counts .

Variable Description :

Table 1

Video ID	*An 11-digit string, which is unique*
Uploader	A string of the video uploader's username
Age	An integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
Category	A string of the video category chosen by the uploader
Length	An integer number of the video length
Views	An integer number of the views
Rate	A float number of the video rate
Ratings	An integer number of the ratings

1.2. Data Summary and Exploration

Variable Description :

Table 2

Type of Variable	*Data Type*
Predictor Variable	Object/character
Video Id ,Uploader	Video Id ,Uploader, Category
Age, Category ,Length	Numeric
Rate, Ratings, Comments	Age, views, Rate
Target Variable	Ratings, Comments
Views	

Table 3

	Age	*Length*	*Views*	*Rate*	*Ratings*	*Comments*	*Category_ label*
Count	202,000000	202,000000	202,000000	202,000000	202,000000	202,000000	202,000000
Mean	743,455446	337,777228	20745,693069	4,427228	150,504950	100,658416	4,950495
Std	1,221896	285,477317	54295,006972	0,703536	213,3223386	132,036544	2,892351
Min	738,000000	3,000000	229,000000	1,540000	1,000000	0,000000	0,000000
25%	743, 000000	115,000000	2121,000000	4,252500	53,000000	28,000000	2,000000
50%	744, 000000	273,500000	11627,000000	4,730000	100,000000	69,000000	6,000000
75%	743, 000000	521,25000	20172,500000	4,870000	172,250000	119,000000	8,000000
Max	745, 000000	1444,000000	668112,000000	5,000000	1846,000000	1006,000000	11,000000

2. DATA MINING

2.1. Feature Selection and Model Building

At first appraisals , remarks , length need been utilized Concerning illustration those and only characteristic grid which need the most extreme signifance with the see tallies , thus to catch up those straight relapse model need been actualized on the information which need been divided under preparing and testing situated , thereabouts that the model Might make fitted on the preparing information Furthermore its exactness Might a chance to be computed on the testing information. The Linear Regression method has been imported and instantiated, and model is built which gives us a particular equation formula which would predict the number of views.

2.2. Implementation of Linear Regression

Should execute whatever straight model , those principal step is with search for features What's more its legitimacy on be connected in the model fabricating. Taking after are those guidelines vital to model building. Features are also known as predictors, inputs, or attributes. The response is also known as the target, label, or output.

1. **"Observations"** are also known as samples, instances, or records.
2. In order to build a model, the features must be numeric, and every observation must have the same features in the same order.
3. In order to **make a prediction**, the new observation must have the **same features as the training observations**, both in number and meaning.

Now after implementing the model , the equation for the regression problem needs to be specified , the results summary are as follows :

$$\text{Number of Views} = [\text{Age } (-0.09) + (\text{Length } (-0.05) + (\text{ratings } (0.81) + (\text{comments } (-0.42))]$$

The above equation clearly depicts that the ratings are the ones who plays an immense role for optimized view counts, on doing further variable transformation we use ratings and comments only as the features which gave us a much better accuracy based on the RMSE(Root Mean Squared Error) .The Formula for view counts for using only these features would be as follows :

$$\text{Number of Views} = [(\text{ratings } (0.79)) + (\text{comments } (-0.39))]$$

2.3. Implementation of Bagging(Random Forest)

To build this predictive linear model ,sklearn library has been used , the best part for model building is that we have certain algorithms which could boost up the accuracy of the model like random forests , gradient boosting, XG Boosting .

Random Forests has that powerful characteristic to predict the values for regression problems like view counts . The Accuracy is always maximized in this model as we have 202 observations with 9 columns . If we divide this dataset into training and testing , then training would consist of 151 observations and testing would consist of 51 observations , and we need to specify the following steps :

Create Multiple DataSets : Inspecting may be done with reinstatement on the first information and new datasets need aid structured.

Those new information sets camwood have An portion of the columns and additionally rows, which are for the most part hyper-parameters clinched alongside An packing model.

Build Mutiple Classifiers :

1. Classifiers are built on each data set.
2. Generally the same classifier is modelled on each data set, predictions are made.

Combine Classifiers : Those predictions for every last one of classifiers need aid joined utilizing a mean, average or mode worth contingent upon the issue proclamation.

The joined qualities would for the most part a greater amount strong over a absolute model.

3. SENTIMENT ANALYSIS OF VIEWERS

3.1. Import Text Data

Now the Natural Language Processing (NLP) consists of various stages of text cleaning stages like removal of stopwords , Lexicon Normalization which consists of sub process like Lemmatization and Stemming followed by which Text Data has been converted into features using Syntatical Parsing And Topic Modelling Techniques . So initial process involves the partition of dataset into training and test set of the text data To Apply Machine Learning on the text it is necessary to convert that into numerical features to classify the sentiments of the viewers based on their comments.

3.2. Convert Text Into Tokens

So the next step involves the vectorising of text using Count Vectorizer function from the feature extraction package . Through the use of this package , the text data has been converted into labels with the columns as the word tokens consisting of every present feature value .A sample Document – Term Matrix is given in the below picture which could be further stated as **corpus of documents** .

Table 4

	Hello	*Call*	*Me*	*Let*	*Tomorrow*	*Him*
0	0	1	0	0	1	1
1	1	1	1	0	0	0
2	0	1	1	2	0	0

Now After converting our text based comments data into tokens, there are plenty of algorithms like Logistic Regression , Support Vector Machines(SVM) , Naïve Bayes Classifier , since , Naïve Bayes works on the principal of Conditional Probability , thus , it works pretty well in terms of model's accuracy .

3.3. Naïve Bayes 'S Equation

1. P(C|X) = (P(x|c) * P(c)) / P(x).
2. P(c|x) may be those posterior likelihood for class (target) provided for predictor(attribute).
3. P(c) will be those former likelihood from claiming class.
4. P(x|c) will be those probability which will be those likelihood from claiming predictor provided for population.
5. P(x) is the former likelihood of predictor. EQUATION TO INTERPRET VIEWER RETENTION
6. P(V|R) = (P(r|v) * P(v)) / P(r)
7. P(V|R) is the posterior probability of viewer revisit given predictor(comments).
8. P(V) is the prior probability of viewer revisit to the channel .
9. P(R|V) is the likelihood which is the probability of giving comment given class.
10. P(R) is the prior probability of comments (predictor).

Now the functioning of this classifier initially depends on the document – term matrix which has been developed from the X _ train feature data ,and is applied on the Response data . The Response Variable has been labelled as 0 and 1 .1 representing the viewer's revisiting as 1 and non – revisiting as 0 . After the model has been fitted , there would be class predictions made on the testing data of features present as document – term matrix . After the Class predictions has been made , the accuracy of the model would be tested from the predicted values to the actual values being present in the vectorised format of data . Now to cover more insights it is necessary to look at this problem in a more actionable way by generating confusion matrix , which gives us a clear understanding of our predictions where we could match up with the prediction values being true or false with the actual values being true or false. Thus, concepts like True Positive which tells that the number of predictions which has been classified as the viewer retention as yes , it correctly matches up with the true actual values . True Negatives Confirms to the fact that the predicted values which says that viewer would again visit this page incorrectly classified it as true ,where actually viewer wouldn't visit this page. Vice – Versa to these concepts of False Positives & False Negatives are also being used to come up with actionable insights .

Confusion Matrix Diagram : Array ([[1203, 5], [11, 174]])

3.4. Confusion Matrix Interpretation

1. **Accurate Positives (TP):** These would instances over which we predicted yes (they will return to the channel). Also they would bring visited. (1,1).
2. **Genuine inconsistency Negatives (TN):** We predicted no, What's more they didn't visit. (0,0).
3. **False Positives (FP):** We predicted yes, Yet they don't really visited the channel. (Also known as a "Type i slip. ")(0,1).
4. **False Negatives (FN):** We predicted no, Be that they really do bring visited those channel. (Also known as a "Type ii slip. ")(1,0).

3.5. List of Insights From Confusion Matrix

Accuracy: Overall, how often is the classifier correct?

1. (TP+TN)/total = (174+1203)/1393 = 0.98.
2. **Misclassification Rate:** Overall, how often is it wrong?
3. (FP+FN)/total = (5+11)/1393 = 0.011.
4. Also known as "Error Rate".
5. **True Positive Rate:** When it's actually yes, how often does it predict yes?
   a) TP/actual yes = 174/185 = 0.94.
   b) Also known as "Sensitivity" or "Recall"
6. **False Positive Rate:** When it's actually no, how often does it predict yes?
   a) FP/actual no = 5/1208 = 0.004.
7. **Specificity:** When it's actually no, how often does it predict no?
   a) TN/actual no = 1203/1208 = 0.99
   b) Equivalent to 1 minus False Positive Rate
8. **Precision:** When it predicts yes, how often is it correct?
   a) TP/predicted yes = 174/179= 0.97
9. **Prevalence:** How often does the yes condition actually occur in our sample?
   a) Actual yes/total = 185/1393 = 0.13.

3.6. Compute the Accuracy of the Model

3.6.1. Accuracy score

After the model has been built and confusion matrix has been analysed , the next step is to measure the accuracy of the fitted naïve bayes model. since, our main target is the true positives, who are the main visitors who are going to revisit the channel post video upload process, thus, the model's accuracy needs to be justified in order to interpret that what

Amount of percentage our model is accurate . If the model's accuracy is above 75 % , then it is considered to be a better fit

To interpret the accuracy score , we would use metrics package and accuracy_score function to interpret the model's accuracy. Below is the snapshot of the code line to measure the accuracy score for our fitted model .

```
In [61]: # calculate AUC
         metrics.roc_auc_score(Y_test, y_pred_prob)
Out[61]: 0.98664310005369615
```

3.6.2. Roc scores and area under the curve

After the implementation of confusion matrix and gathering insights through its positives and negatives ,there is a need to quantify the accuracy of the classifier . The parameters required for the ROC score are as follows :

1. True Positive Rate(TPR) = TP/Actual Yes
2. False Positive Rate(FPR) = FP/Actual No
   a) TPR = 174/185 = 0.94
   b) FPR = 5/1208 = 0.004

Thus, the area under the curve (AUC) is 0.94 , which seems to be a pretty robust classifier ,and the threshold value is around 0.5 , above 0.8 the model's accuracy optimizes to its maximum range .


```
In [61]: # calculate AUC
         metrics.roc_auc_score(Y_test, y_pred_prob)
Out[61]: 0.98664310005369615
```


4. ANALYSIS AND INSIGHTS

Table 5

Category	*Age*	*Length*	*Views*	*Rate*	*Ratings*	*Comments*
Autos and Vehicles	743,000000	900,000000	12973,000000	4,960000	81,000000	28,000000
Comedy	743,560000	275,080000	10763,040000	4,412400	140,800000	121,160000
Entertainment	743,774194	343,955484	12836,483871	4,539355	107,709677	72,161290
Film and Animation	743,500000	631,911765	24071,529412	4,877941	154,117647	75,558824
Gadgets and Games	742,000000	259,666667	224653,333333	4,093333	737,000000	223,000000
Hot to and DIY	744,000000	615,000000	7207,166667	4,748333	136,333333	71,666667
Music	743,593750	229,406250	13397,281250	4,508125	142,031250	76,031250
News and Politics	743,200000	293,666667	20277,733333	4,206667	230,066667	236,333333
People and Blogs	743,111111	280,777778	11325,555556	4,129722	130,277778	115,777778
Pets and Animals	743,428571	56,285714	21427,857143	3,298571	100,857143	67,714286
Sports	743,363636	125,272727	61989,636364	4,348182	151,272727	53,272727
Travel and Places	744,000000	212,000000	2535,000000	4,740000	137,000000	150,000000

5. RELATED WORK

Several approaches have been proposed to make the View count predictions more robust are as follows :

5.1. Development of Classification system for virality content

Time Series Data has been collected and analyzed , and especially to look at the viral part , the virality is measured on per hour basis , and view counts and meta data has been used as features to automatically classify the video to be viral or fixed population .

5.2. Data Exploration according to virality and popularity

There are immense amount of metadata extracted from the youtube data API , which could be used as features for model building but the target mainly is to classify the video category according to its view count dynamics and virality , thus ,visualizations would be used to interpret the results and insights.

5.3. Implementation of Six-Biological Models

Utilizing an expansive situated about experimental data, that those view-count to 90% of features to Youtube might Undoubtedly a chance to be cohorted with no less than a standout amongst these models, with An mean lapse which doesn't surpass 5%. We infer programmed approaches for classifying those view-count bend under a standout amongst these models Also from claiming extracting those The greater part suitableness parameters of the model. We investigation observationally the effect of videos' Notoriety Furthermore classification on the Development of its view-count. We At last utilize the over arrangement alongside the programmed parameters extraction in place on foresee those Development for videos view count.

5.4. Using Multiple State-of-the-art-techniques

In this system ,there are six diffusion models been implemented to classify the virality of the content and population growth. But in this paper, initially, Linear Regression was used ,and after using best set of parameter or feature engineering ,Root Mean Squared Error(RMSE) which gives out the accuracy of the model. On the another front to improve the accuracy of the model , bagging algorithms like random forest has been implemented to maximize the accuracy of the fitted model to predict the view counts on a particular video .

5.5. Using NLP Techniques for analysis of Text Data

The dataset comprises of video id variable , which could be identified by the comments and hashtags they have used for reviewing video ; thus,collecting those text-based data would enable us to analyze them and come up with sentiment polarity in the form of positive , negative and neutral sentiments.Hence, in this way we would come to know about the emotional tone of the viewers towards the channel which would act as a recommender for the viewers .

5.6. Feature Building and Model Optimization

Those altered target populace property happens clinched alongside exactly feature classifications clinched alongside Youtube Concerning illustration news, sport What's more motion pictures. Indeed, features done these classifications scope fast those top of the Ubiquity et cetera inside a short chance the dissemination dies out and those view-count doesn't further increment. Those second paradigm in the order worries those structural virality.

6. PROPOSED METHODOLOGY

6.1. Data Cleaning

1. Initially Data is collected and scraped from youtube Data API , beautiful soup in python and is being stored in Mongo DB.
2. After storage of data,it might be possible to detect the outliers and see the values which are inconsistent or missing.
3. Missing Data could be replaced by the values according to the problem statement, namely ,by measures of central tendency , Mean,Median,Mode,Inter - Quartile Values .

6.2. Data Exploration

6.2.1. Univariate amd Bivariate Analysis

To find the relationship between two continous variables , we usually do univariate(only one variable) & bivariate(more than one variable) analysis through different statistical tools and methods like Analysis Of Variance(ANOVA),Hypothesis Testing,P-value,and to find the strength of relationship we use Pearson's Coefficient .

1. Use {Plotting Mechanisms like Scatter Plot, Box Plot, Histogram, Ternary Plots, etc. to visually demonstrate the strength of relationship among each other.
2. Bivariate Analysis[Categorical & Categorical Data Types]
3. Categorical Data - Chi- Square Test
4. Categorical & Continous - Boxplot , 3d-plots , Contour plots .

6.3. Feature Engineering

Following are the two steps involved in this mechanism :

1. **Variable Transformation :** Existing Variables could be used to create a new set of variables which could be lately used as a feature for model buliding.
2. **Variable/Feature Creation :** This phase is about the features which could be used after deriving the information of the mentioned variables , like for example in this module we could divide the number of likes to the number of View Counts to come up with a new feature termed as impressions.

6.4. Analysis and Modelling

1. Our Ultimate aim is to predict the number of views based on metadata & the other statistics collected from different other sources.
2. The next step is to come up with a prediction system which would be optimum enough to tell about the accuracy of the model.
3. The Roadmap specifies the fact that firstly we try off with Linear Regression whose coefficients and intercepts are those values which would be intrepreted as the deciding parameter for feature selection & model building .
4. The other methods which would be relevant or could be used for future reference are like Random Forest Classifier, XGBoost ,Gradient Boosting, etc.
5. The main reason behind using these advanced ML methods is that it involves multiple partitions of dataset ,which in turn is calibrated into a single model of the classifier or regressor model which would be the most efficient model as compared to others.

7. PERFORMANCE EVALUATION

7.1. Results and Discussions

1. **Existing Approach:** Fig.1, demonstrates the relationship between the different kinds of insights being generated from the fitted classification model .
2. **Proposed Approach:** Fig.2, is the regression coefficients and its parameters which seems to play an influential role in prediction of the views like ratings, comments, likes, dislikes, etc.
3. **Comparison:** Fig.3, gives us a comparative study of videos in different genres , where different parameters like ratings ,comments , likes,dislikes ,views , etc. has been used and Gadgets & Games seems to have won by immense margin.
4. **Variable Selection :** Fig 4-7 tells us about the statistical relationship between the attributes and their relationships between each other ,since, based on these statistical tests and visualizations, features would be subsetted for implementing a machine learning algorithm .
   a) **Fig 4:** Relationship between views and ratings
   b) **Fig 5 :** Relationship between views and age of video
   c) **Fig 6 :** Relationship between comments and views.
   d) **Fig 7 :** Relationship between videos of different genres and the video length .

After the data has been visualized and each variable signifance has been tested through different statistical tests like ANOVA has been used to subset the features for model building .

Table 6
7 Days Scenario

Model Type	*Distribution (%)*	*Hard window*	*Hard bounded (%)*	*Soft window*	*Soft bounded (%)*
E	66.9	m: 0.6	13.5	m: 0.66	15.3
ME	0.9	m: 0.75	17.9	m: 0.82	20
G	10.5	m: 0.42	7.8	m: 0.47	8.8
MG	7.9	m: 0.54	10.9	m: 0.61	12.7
S	11.2	m: 0.97	32.9	m: 1	33.8
MS	2.6	m: 0.81	18.8	m: 0.84	19.5
All	100	m: 0.63	15.1	m: 0.68	16.6

Table 7
15 Days Scenario

Model Type	*Distribution (%)*	*Hard window*	*Hard bounded (%)*	*Soft window*	*Soft bounded (%)*
E	62.7	m: 0.55	10.7	m: 0.59	12.1
ME	1.3	m: 0.9	22.2	m: 0.94	23.3
G	8.4	m: 0.44	7.5	m: 0.47	7.9
MG	16.5	m: 0.7	13.4	m: 0.77	15.6
S	8.1	m: 1.1	38.1	m: 1.11	38.9
MS	3	m: 0.94	23	m: 0.98	24.5
All	100	m: 0.63	13.6	m: 0.67	15

Table 8
30 Days Scenario

Model Type	*Distribution (%)*	*Hard window*	*Hard bounded (%)*	*Soft window*	*Soft bounded (%)*
E	52.3	m: 0.5	9.5	m: 0.55	10.5
ME	2.2	m: 0.79	17.3	m: 0.87	19.8
G	5.2	m: 0.45	8.5	m: 0.48	9.1
MG	30.6	m: 0.68	11.6	m: 0.76	14
S	5.8	m: 1.1	39.6	m: 1.14	40.6
MS	3.9	m: 0.89	20.8	m: 0.92	21.7
All	100	m: 0.61	12.5	m: 0.66	13.9

INSIGHTS

View Prediction

Age of the video & length seems to be inversely proportional to the number of views,thus,old videos and lengthy videos are one of the

reasons for view reduction.

Figure 1

Ratings is one of the feature which keeps on boosting up the number of views , as evident from the prediction model , for every one rating

per view is increased.

Equation for View Counts

Number of Views = [Age *(-0.09)* + *(Length* (-0.05) + (ratings *(0.81)* + *(comments* -0.42))

Figure 2

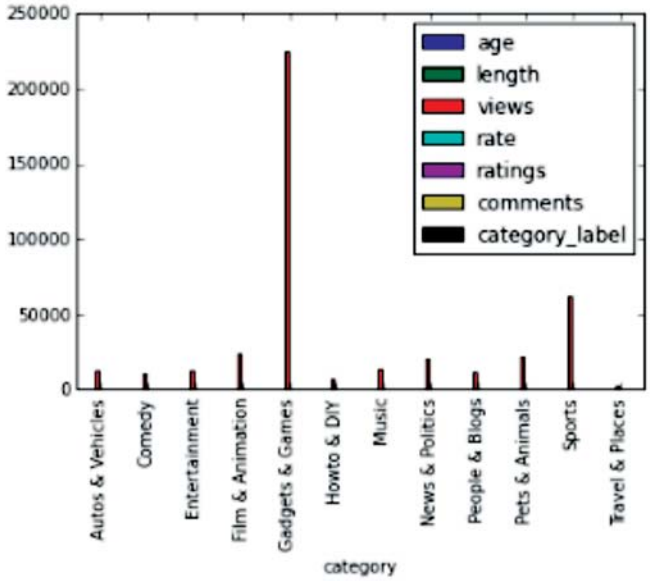

Figure 3

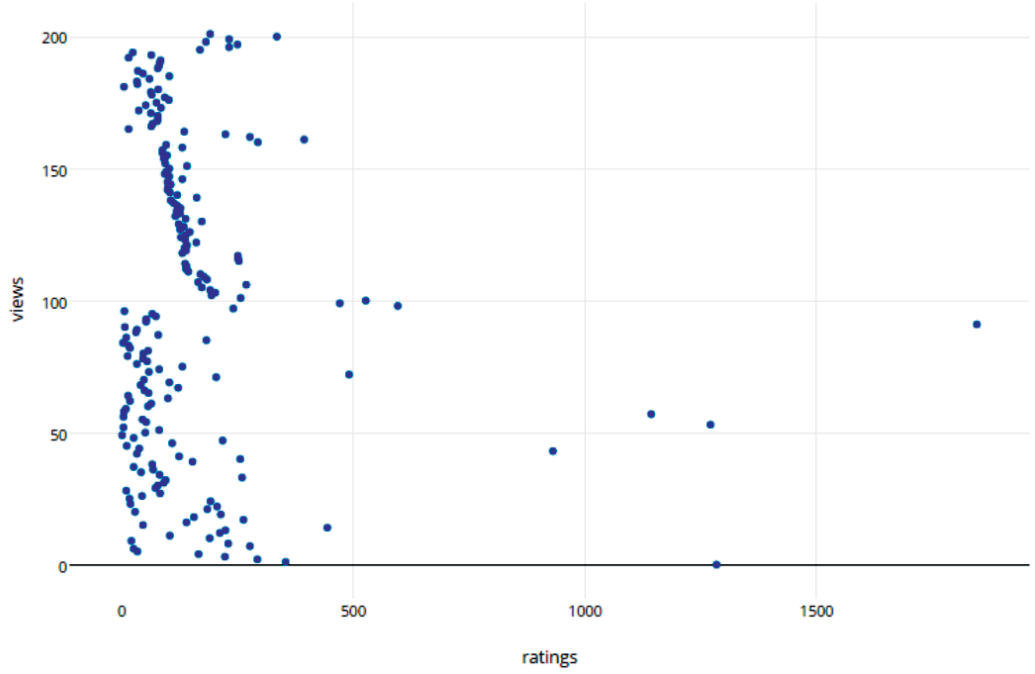

Figure 4

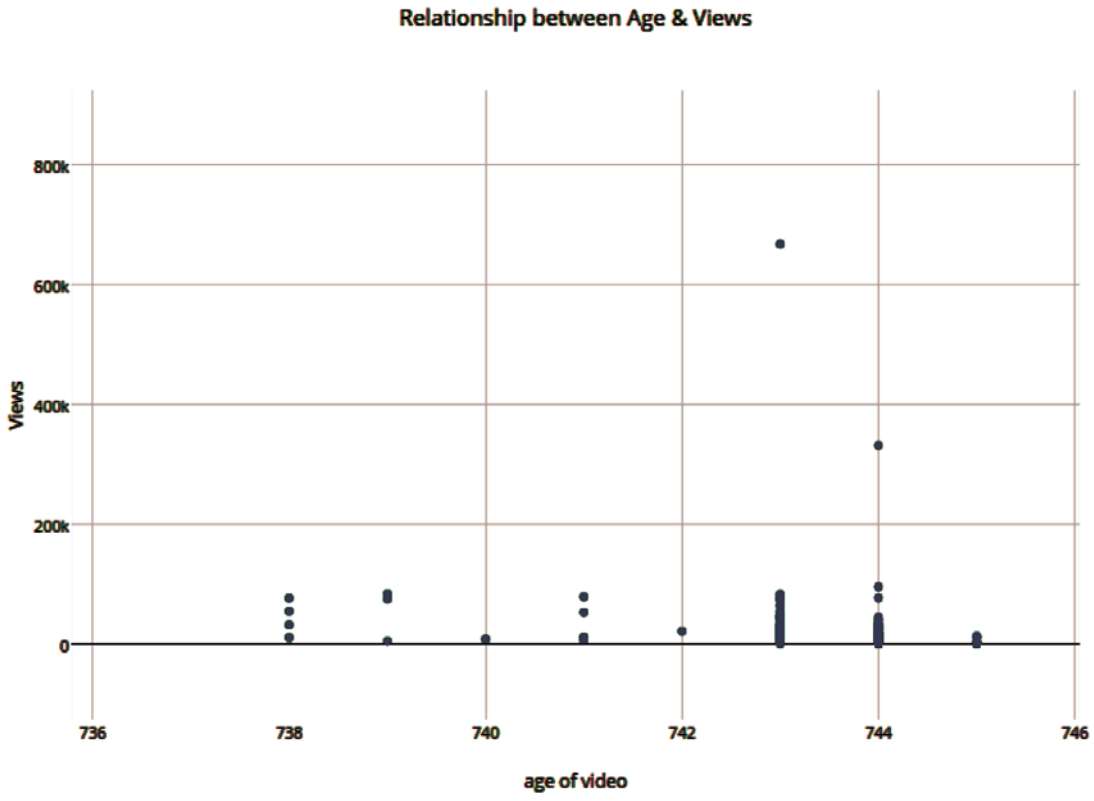

Figure 5

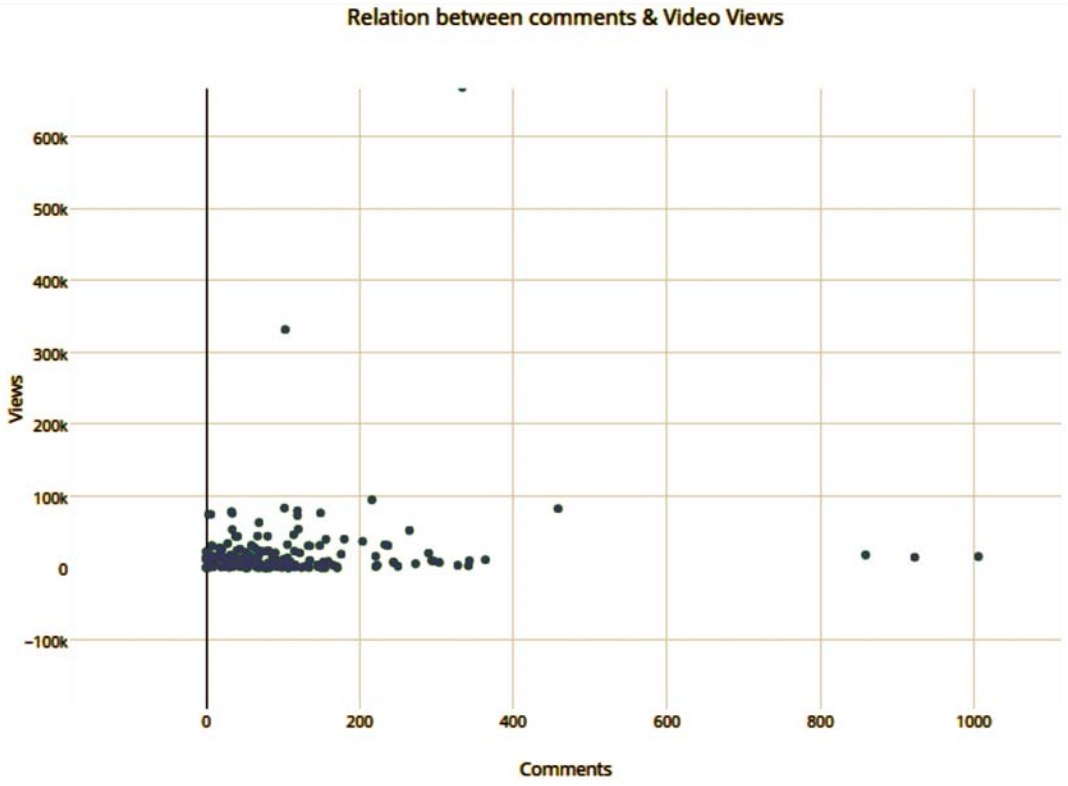

Figure 6

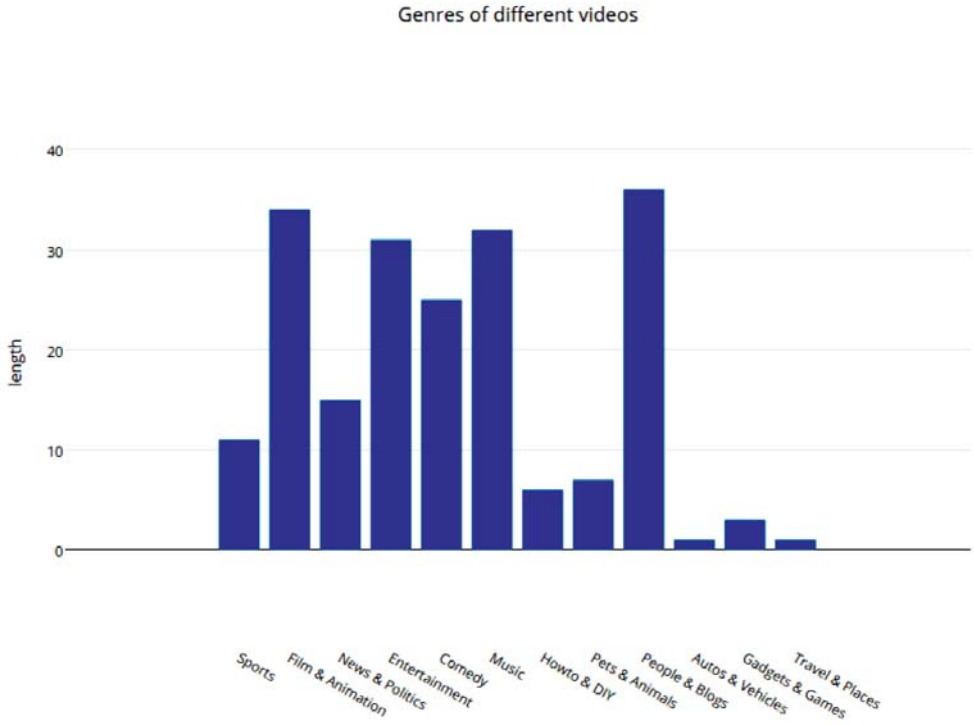

Figure 7

8. CONCLUSION AND FUTURE WORK

In the present scenario we have evolved a method to predict the number of views where we would certain set of parameters to predict that this video would with these features could come up with certain number of views , based on which marketing strategy of the video or content would be formalized. Naïve Bayes classifier has been used to analyze the comments – related text data on the videos, which would classify the viewer by comments whether he or she would again visit the channel or not, and this concept has been analyzed and detailed insights have been given in the confusion matrix section. Our Prediction accuracy was pretty less when Regressions were applied ,but after better parameter tuning, the random forest turned out to be the better model to implement.

Thus,the future work would involve to keep working on the methods or algorithms or parameters based on which better prediction of the view counts could be specified .

In the Case of Viewer Retention Problem , Much more training accuracy would be the ultimate aim for its classification , so that there is no problem of overfitting and accuracy score is also maximized .

REFERENCES

[1] Modelling View-count Dynamics in YouTube Ce´dric Richier∗, Eitan Altman†, Rachid Elazouzi∗, Tania Jimenez∗, Georges Linares∗ and Yonathan Portilla∗ ∗University of Avignon, 84000 Avignon, FRANCE.

[2] The Application of Regression Analysis in Correlation Research of Economic Growth in Jilin Province and Regional Income Distribution Gap Shaojie Zhang*,Yanhua Wu*,Sch. of Manage., Jilin Univ., Changchun.

[3] Analysis Of a Random Forest Model. Gerard Biau ´ ∗ LSTA & LPMA Universite Pierre et Marie Curie – Paris VI ´ Boˆıte 158, Tour 15-25, 2eme ` etage ´ 4 place Jussieu, 75252 Paris Cedex 05, France. Journal of Machine Learning Research 13 (2012) 1063-1095.

[4] A Review Paper On Algorithms Used For Text Classification.Bhumika1 , International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 3, March 2013 ISSN 2319 – 4847.

[5] An Overview of Use of Natural Language Processing in Sentiment Analysis based on User Opinions. Vikash Singh Rajput. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 4, April 2016 ISSN: 2277 128X.

[6] Text Mining Approach to Classify Technical Research Documents using Naïve Bayes. Mahesh Kini M1 , International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015. ISSN (Online) 2278-1021 ISSN (Print) 2319 5940.

[7] Text Mining with Information Extraction. Raymond J. Mooney and Un Yong Nahm Department of Computer Sciences, Multilingualism and Electronic Language Management: