PREDICTING THE POPULARITY OF YOUTUBE VIDEOS

APURVA GURURAJ KATTI(AGK17C)

TEJAS DATTATREYA HASARALI(TDH17)

DECEMBER 10

FALL 2017

Introduction:

With the widespread online global access of data and the ease with which an online content can be produced, we often are directed to understand the underlying concept of popularity growth on the internet. It is of utmost relevance to a broad range of services like designing an effective caching model, viral marketing strategies, estimation of costs, advertisement campaigns and for the overall improvement for the future content. Our approach involves considering You Tube a, a video sharing portal as case study and predicting the popularity of the given video, or in other terms the prediction of the occurrence of a video on the trending page.

"Trending Videos" are videos that have become popular because they were embedded in the web's most popular websites and a significant number of people viewed the video externally in addition to on youtube.com [3]. Since the dataset which we were successful in collecting consisted of variables in the form of categorical and continuous data, we decided to develop our model using logistic regression. Regression is commonly used for prediction based problems. It concentrates on finding the relationship between the dependent and independent variables. The dependent variable gives us the output and the result of the prediction.

The independent variable represents the input to the model or the causes for the output. Regression model helps in studying the change in the value of the dependent variable according to the varying independent variables. After integrating the datasets, we used two-third of it as our training model and one-third for testing. In our model we predicted the popularity by using either 0 or 1. Given the data from the test set, our model can predict the popularity. Value 0 represents a non-tending video and a value 1 represent that the given video is trending.

Literature Survey:

Previous work on prediction of popularity includes efforts by Cha et al, who studied the popularity cycle of YouTube videos. The findings stated that the trending videos tends to be the one uploaded recently but on an average, that around 80% of the videos watched on a day are older than a month. Figueiredo et al, characterized the popularity evolution patterns and studied the effect of different referrers on these patterns. Rodrigues et al, found out that when the duplicates of the same videos were considered, it produced different popularities. Crane and Sornette identified four main classes when analyzing the popularity evolution pattern.

Accordingly, the results stated that majority of the videos do not experience popularity peak and that it can be explained through a simple stochastic process. These videos were termed as Memoryless. In contrast, there were other which experience peak popularity and can be categorized as Viral, Junk and Quality videos. Viral videos gain popularity through internal propagation through word of mouth. The peak increases slowly and later ceases. Quality videos rise to popularity instantly due to external event such as its upload on the first page of YouTube and slow decay as it gets propagated amongst several users.

Junk videos experience the burst, but it quickly drops after a point in time. Wu et al, proposed on the special type of neural network called the reservoir computing based on popularity data of

previous days. But it fails because of randomization effects, as the hidden layer weights are randomly chosen values and it takes several iterations to reflect the data instead of random weights. Yin et al, proposed a model based on the opinions of users. This can be seen in the form of likes and dislikes representing positive and negative response respectively. By conducting the empirical studies, it was discovered that users can display two personalities: Conformers and Mavericks. Conformers voted in favor of the majority and mavericks in disagreement. It was also seen that any user exhibits both personalities to different degrees. One disadvantage of this model is that it requires the full knowledge about users votes in order to build user profiles.
In the research paper titled Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study by Soufiana Mekouar, Nabila Zrira, El-Houssine Bouyakhf, a new function was designed to decide if the video is popular or not. They have used logistic regression to predict the popularity by using the popularity function. The parameters of the video considered includes uploader, age, category, length, number of views, rate of video, rating of video, number of comments. Our model also uses the logistic regression but since YouTube no longer provides parameters in the form of rate and rating, we have developed our model such that they incorporate the new parameters that the YouTube now offers in the form of likes, dislikes, views, comments, published date, category ID, video ID and so on.

Description of the methodology used:

Data Collection: Data collection is one of the most important steps of data analysis. A model is referred as 'good' when it is capable of producing correct output on diverse datasets. In our model it is the meta data of the YouTube videos. Meta Data is the data which provides information about other data. These metadata can be gathered using the public APIs exposed by YouTube. The meta data contains the information like - id, title, published date, trending date, likes, dislikes, comments, description, tags, channel name of the uploader.

Data Pre-processing: The first step of data pre-processing is converting the collected data into readable format. The collected data was converted from pickle format to CSV format. The obtained CSV format is pre-processed by performing following tasks:

- Missing values are replaced depending on the type of data. This is important because the missing values can affect the performance of the model
- All the gathered data is formatted into a similar form for combining into one file. The Trending data and Non-trending data are in different form, they are converted into similar form to produce one file with all the data
- Removal of stop words, emoticons and punctuations from the data. We remove the stop words as they are frequently used in most of the sentences and produces no meaning. Emoticons and punctuations are removed because they do not have any importance in classification.

Feature Extraction: In feature extraction the formatted data is converted into the form that can be used to train the model. Following tasks are carried out in feature extraction step:

- Conversion of text data into numerical data: For this we employed two methods frequency calculation and ranking. The frequency of text data in the trending videos are calculated and it is used to build a ranking system. Text data is replaced by numerical values using the ranking system developed
- Categorical data is converted by creating dummy variables. Separate columns are created for every category, if a video belongs to a category it is represented by 1 (0 otherwise).
- Creation of new features using two or more features.
- Important features are extracted using L1 norm error Linear SVC with penalty parameter 0.01. Linear SVC gives the most important features from the set of given features by calculating its importance.

Training the Model: The model is trained using Logistic Regression, as the obtained features are continuing and categorical. Logistic Regression is a type of regression analysis used when the dependent variables are binary and independent variables are categorical and continues. Logistic Regression uses sigmoid function to predict the dependent variable. KNN classifier is also used for comparing the accuracy of the Logistic Regression model. The performance of the model is evaluated using ROC curve, Classification report, and confusion matrix.

- ROC Curve-ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. [4]
- Confusion matrix-It is also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. [5]
- Classification report- It is a summary of precision, recall and F1 score for each class of the classification model.
- Precision-precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.
- Recall-recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. [6]
- F1 score-It combines both precision and recall by calculating the harmonic mean between the two

Implementation:

Data Collection: The Trending videos meta data is directly obtained from crowdsourcing platform Kaggle in CSV format. The Trending videos of five regions (United States, Canada, United Kingdom, Germany, and France) are collected for the month of November 2017. The features of the dataset include 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.
The Non-Trending Videos meta data was collected using the public APIs exposed by YouTube. The Video IDs of more than 25000 videos uploaded in the month of November 2017(for the regions United States, Canada, United Kingdom, Germany, and France) was collected by scrapping the YouTube API. Around 8700 unique Video IDs obtained from more than 25000 Video IDs are selected. The meta data of these videos are collected by again querying the YouTube API. The obtained meta data was converted from pickle format to CSV format. The features present in the meta data of the videos include duration, dimension, projection, Video id, caption, license, category Id, channel Id, description, published At, tags, thumbnail, comment Count, dislike Count, like Count, view Count, embeddable, licensed Content, privacy Status, title, definition, default Audio Language. The Trending data contains Channel Name, where as Non-Trending Videos contains Channel ID, one more API call was made to get the dictionary of Channel ID to Channel Name for solving this inconsistency.

Data Pre-Processing and Feature Extraction: The tasks carried out in data pre-processing are:

- The meta data of trending videos is resampled randomly to balance the classes.
- The dates obtained are reformatted to date datatype of pandas to carry out different operations.
- The missing values are filled with either zero or one depending on the type of feature.
- Missing columns like trending, obtained date are inserted into the data.
- The channel IDs in Non-Trending is converted to channel title using the dictionary obtained by scrapping the YouTube API.
- Trending and Non-Trending datasets are reformatted to make them consistent with each other and are combined into one big dataset.
- The category IDs are converted into their corresponding category names using the JSON dictionary of category IDs. After the conversion, dummy variables are created to represent all the categories of data. These are inserted as new columns and value 1 or 0 is interested. The value 1 represents the video belongs to that category and 0 represent otherwise.
- The Title and tags of the trending videos are used to build a ranking system. They are striped and split into individual words after removing the stop words, emoticons and punctuations. The frequency of every word is calculated and used as a ranking system. This ranking system is used to replace the tags and title. Unique words in tags and titles are

extracted after removing the stop words, punctuations and emoticons, average of the frequency of these words are used in building the model.

- The frequency of Channel Title in trending videos is calculated and It is used to replace the Channel titles.
- New features like age of the videos, rate of views, rate of likes per day are calculated using the existing features.
- The features selected for training the model are age, channel title, comment count, dislikes, likes, tags, title, trending, views, rate of views, rate of likes, category autos & vehicles, category comedy, category education, category entertainment, category film & animation, category gaming, category how to & style, category music, category news & politics, category non-profits & activism, category people & blogs, category pets & animals, category science & technology, category shows, category sports, category travel & events.
- The most important features are selected by applying L1 norm error Linear SVC with a penalty of 0.01. The dimensionality was reduced from 27 to 21 with negligible loss in accuracy.

Training the model: The model is trained using Logistic regression after scaling the data. The accuracy of the model is calculated by splitting 1/3rd of the data as testing and remaining for training the model. Confusion matrix is used to calculate the accuracy of the model which is 80-85%. The performance of the model found using both ROC curve and Classification report which contains recall, precision, F1 score. The model is compared with KNN classifier, even though KNN classifier is returning higher accuracy for smaller number of k value, the Logistic regression model is selected as KNN classifier is highly sensitive to data.

Assumptions:

The title ranking, tag ranking and channel title ranking are calculated using the trending videos of our dataset using the frequency of occurrence. To overcome these limitations, we have used both trending and non-trending datasets of the same time period and same region. Such that the trending google searches in that period can be obtained by using tags of trending videos. The same applies to Channel title ranking as well, the channels which are repeatedly appearing in a very small interval trends of time will have high subscriber count, or it should be about a trending topic.

The accuracy of the model can be improved by obtaining the data of all the top channels and ranking them based on the number of subscribers. The title and tag ranking can be calculated using the google search trends to find the trends in that particular time period and ranking them based on it. The accuracy of our model reduces just by 1-3% when title ranking, tag ranking and channel title are removed from the feature set.

Problems Encountered:

Obtaining of Non-Trending video dataset was not easy, as Non-trending videos are not readily available to scrape from YouTube API. The problem was solved by obtaining the videos randomly in the same time period and same region as the trending videos. Since the videos are obtained randomly, it was compared with trending videos to remove incorrect data and convert the random data into Non-trending data.

Experimental Results:

The Accuracy of the Data model is shown in the figure 1 and figure 2. The four features that are removed to check if the model is a good model are Channel Title, Tags, Age of the Video and Video Title. The Figures 3 to 5 shows the ROC curve of Linear Regression model.

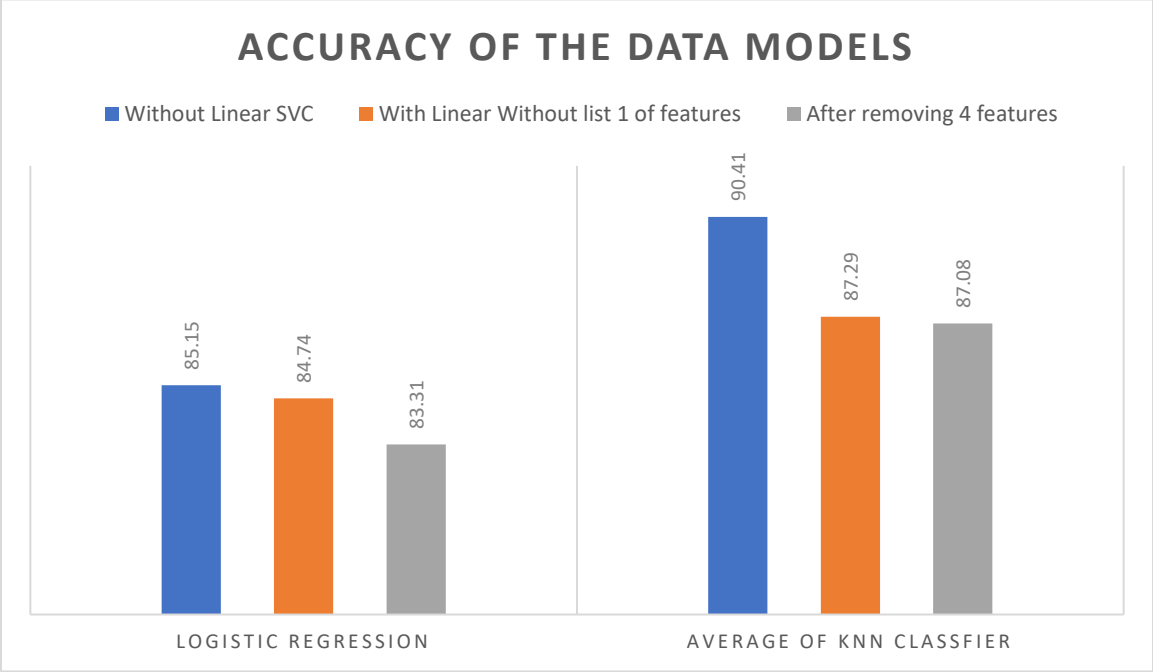

Figure 1

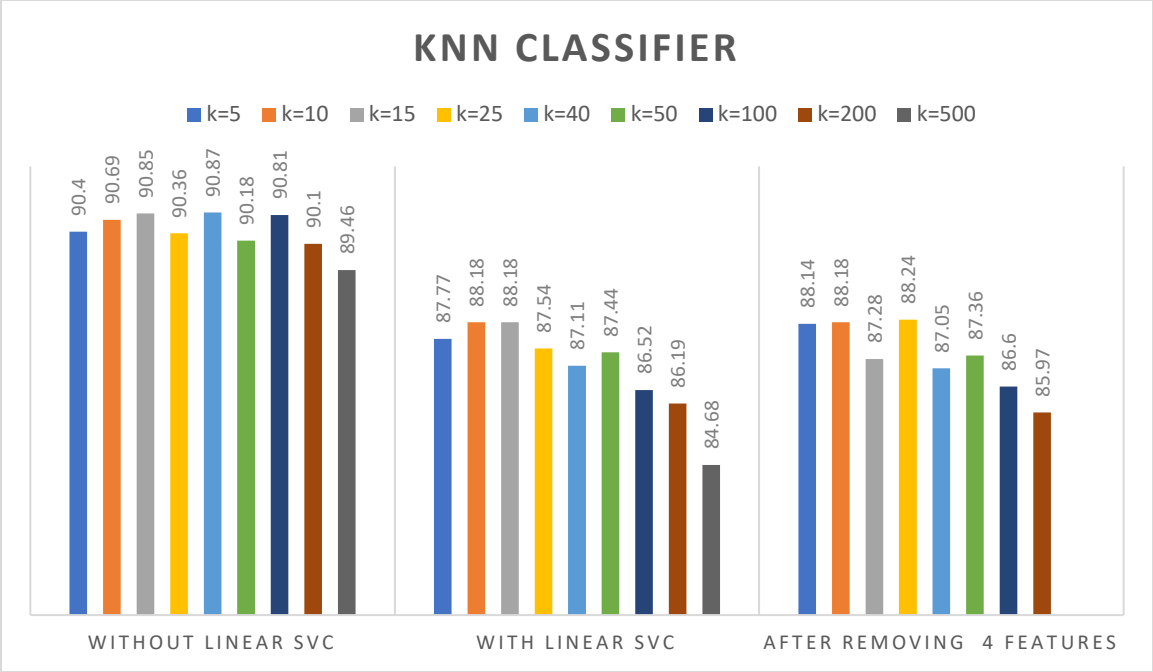

Figure 2

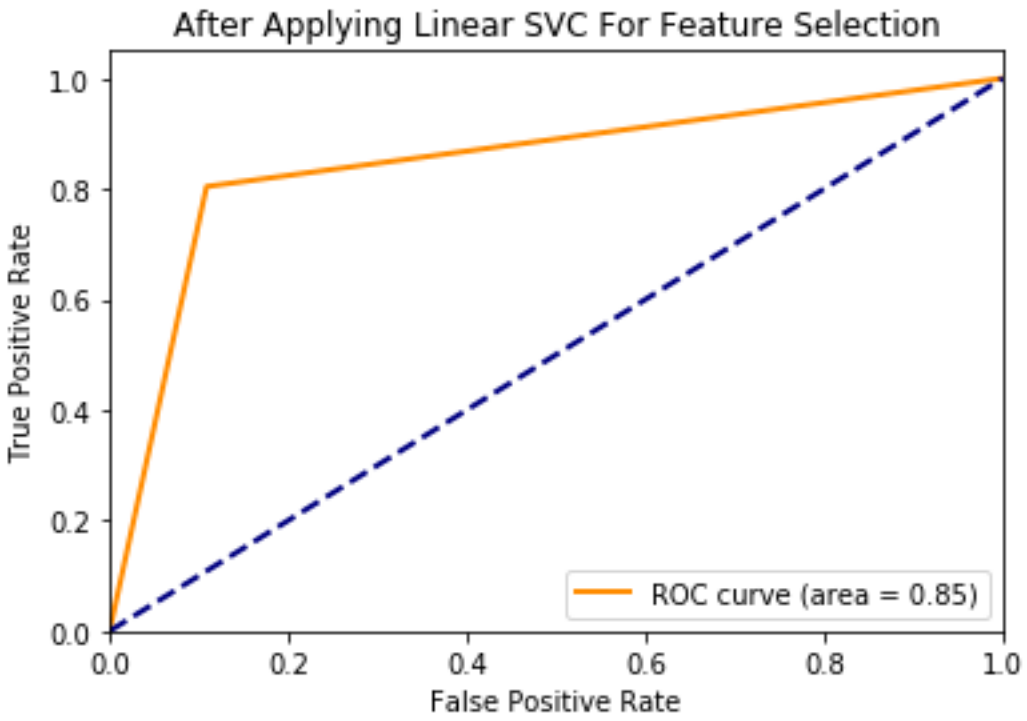

Figure 3

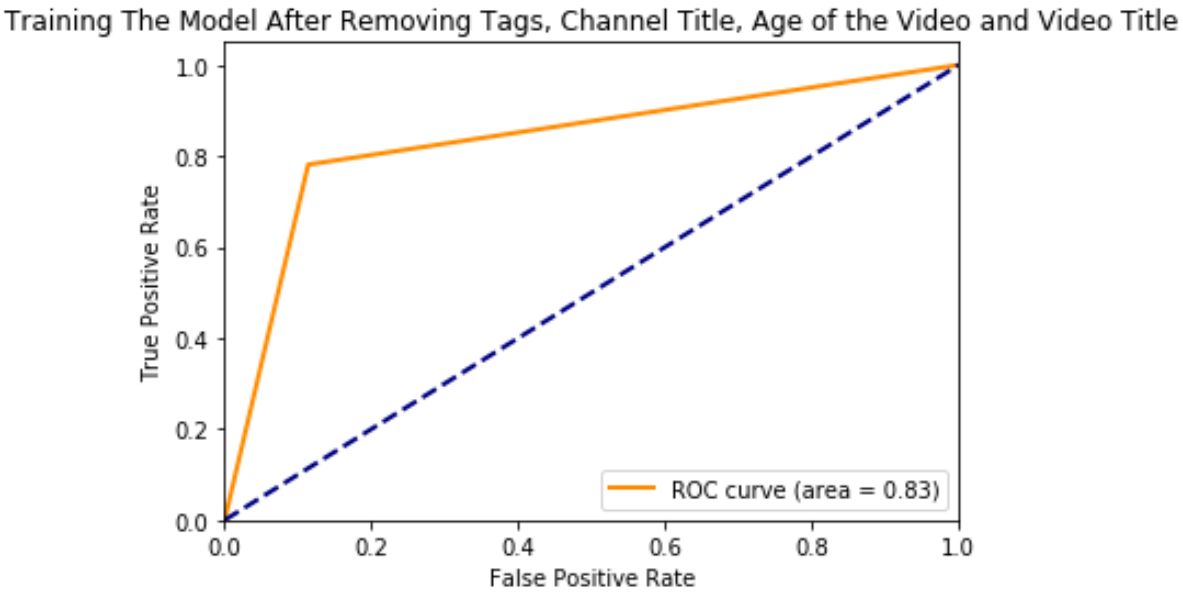

Figure 4

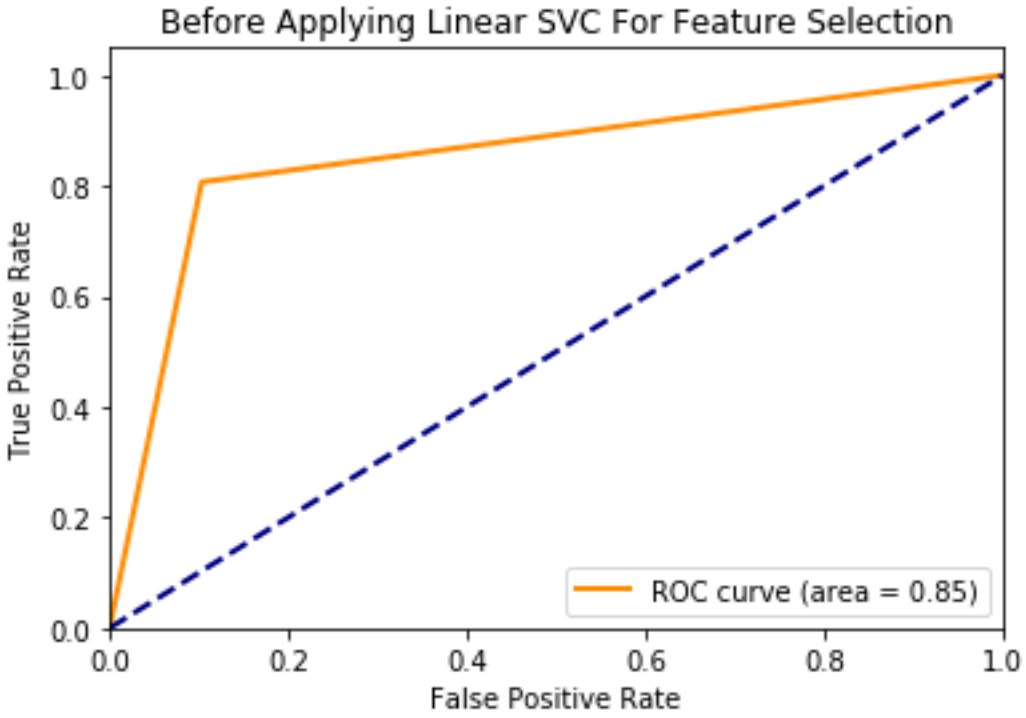

Figure 5

Observations:

We were able to observe some relationships that are prevalent in the dataset obtained. The videos belonging to category Sports, Travel and Events takes time to feature in the trending videos, whereas videos like Comedy, Education, Non-profits and activities either feature in early stages of life or never. The probability of Music and Comedy Videos appearing in trending videos is very high. The maximum number of views is also obtained by music videos. How the videos belonging to each category vary with age and views with respect to trending can be seen in the figures 6 and 7 below.

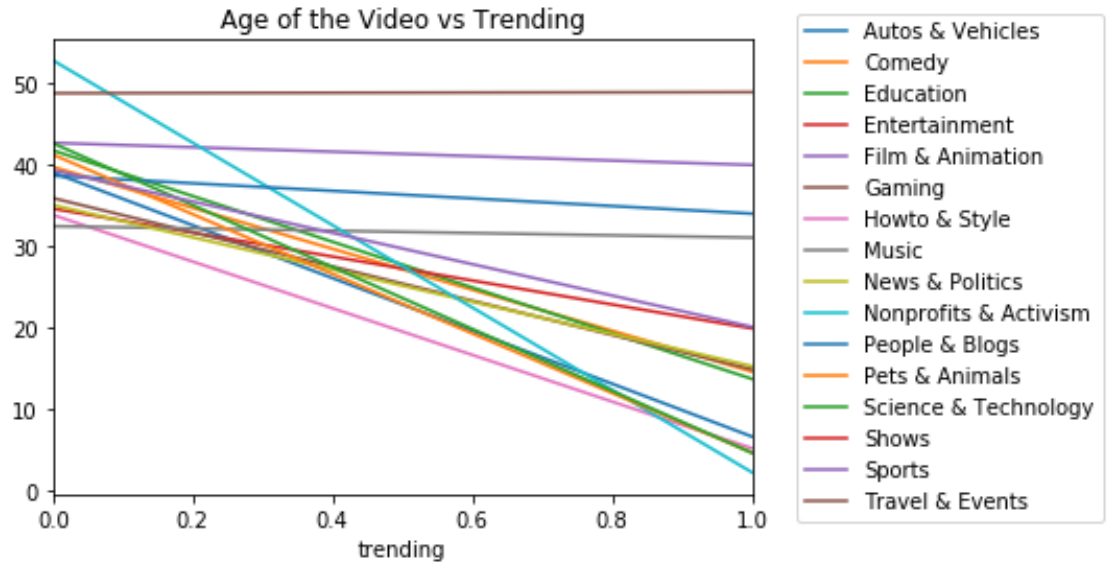

Figure 6

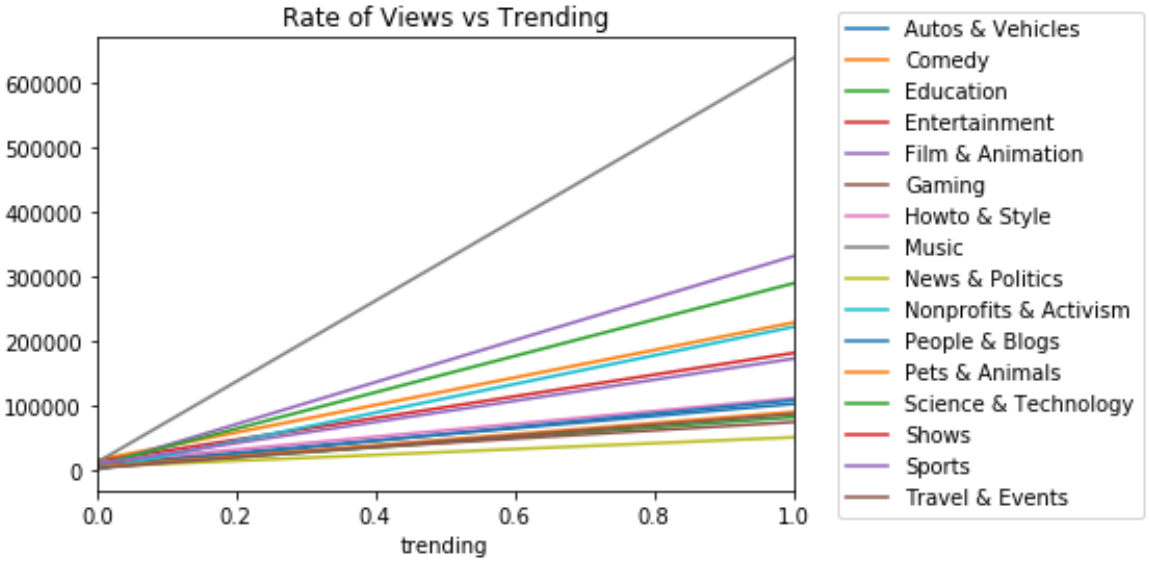

Figure 7

Conclusion and Future Research work:

Developing a model for prediction of popularity is not easy task as different criteria like the types of parameters considered, the type of dataset used, type of data in the datasets differs from one model to another. Our model provides a new method to predict the popularity of YouTube videos using Logistic regression analysis based on several parameters like the video ID, trending date, title, channel title, channel ID, published time, tags, views, likes, dislikes, comments and description of the videos popular in the regions of Germany, USA, France, Great Britain and Canada. We have also used K-NN classifier to compare the accuracy with the Logistic regression. The model provides about 85% accuracy for Logistic regression model.

Future work of our model aims at improving the accuracy of Logistic regression. The channel title in our model is calculated by calculating the frequency of its appearance in the dataset and ranking accordingly. We can also handle channel title by ranking the channel title based on the number of subscribers that the particular channel holds to improve the accuracy. Tags could also be handled by collecting the data maintained by Google per day about the trends and then ranking the tags. We have considered dataset of only five countries and it can be extended for a dataset maintained across the globe. The prediction of popularity using our model can also be extended in prediction of future popularity of a video using early view patterns.

References:

[1]. **Using early view patterns to predict the popularity of YouTube videos** by Henrique Pinto, Jussara M. Almeida, Marcos A. Gonçalves, presented in WSDM '13 Proceedings of the sixth ACM international conference on Web search and data mining.

[2]. **Popularity Prediction of Videos in YouTube as Case Study: A regression Analysis study** by Soufiana Mekouar, Nabila Zrira, El-Houssine Bouyakhf, presented in BDCA'17 Proceedings of the 2nd international Conference on Big Data, Cloud and Applications.

[3]. http://youtube-trends.blogspot.com/p/about-youtube-trends.html

[4]. https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[5]. https://en.wikipedia.org/wiki/Confusion_matrix

[6]. https://en.wikipedia.org/wiki/Precision_and_recall

[7]. Git Hub repository of Ayush Singh (ayushkumarsingh97@gmail.com) for scarping of YouTube Videos.