YouTube Trend Analysis

Neelima Jagtap
dept. of Applied Data Science
San Jose State University
San Jose, US

Greeshma Venkatesh
dept. of Applied Data Science
San Jose State University
San Jose, US

Yash Jobanputra
dept. of Applied Data Science
San Jose State University
San Jose, US

Sougandhika Ayathamaraju
dept. of Applied Data Science
San Jose State University
San Jose, US

Rishikumar Ravichandran
dept. of Applied Data Science
San Jose State University
San Jose, US

***Abstract*—YouTube is one of the most popular video-sharing websites on the internet which is open source. It is the biggest platform with ample video categories in the world. The content includes a variety of music, product promotion, blogs, reviews, educational and entertainment videos, etc. It also includes various features that will help the users to explore and look into an enormous number of videos based on their choices and likings. There are several features that make the video trending and gain the most popularity among the audience. Some of them are the view count and how quickly the videos are getting views. We also consider a fact to analyze various other sites through which the same video is being viewed repeatedly. The understanding of features and deriving insights to predict the popularity of a video is very essential in the current times. Although YouTube is very famous and important, there is still an opportunity to identify and analyze significant insights. The aim of this project is to include interactive features to compare the correlation for the trend of the videos. We have called the live streaming data through API directly from the YouTube storage portal. Our objective is to work on the US region of YouTube data. The objective of the project is to perform machine learning models for three different target variables to fully grasp the features that affect video popularity. For this purpose, we are going to analyze the trending data and compare them through machine learning algorithms like Linear Regression for the Views analysis. Logistic Regression, Naive Bayes, Random Forest, Decision Tree, KNN for the Category prediction and number of days required for a video to be trending list prediction. The implemented models are compared and evaluated to see which model gives the best accuracy on given data.**

***Index Terms*—Live streaming data, Linear Regression, Naive Bayes, Random Forest, KNN, trending analysis.**

I. Introduction

YouTube was established in February 2005, and since then it is a very productive and famous user-created content sharing website in the internet world. As per the statistics of year 2021, YouTube has over 2.3 billion users in the world. The increase in the user enrolling is directly proportional to the videos uploaded on the portal. Approximately 400 hours of videos are shared on the platform per minute across the globe. The viewers watch more than a billion hours of videos on YouTube every day. Due to these billions of views are generated from those videos. Hence, it has become a place for the users to upload the content on YouTube and generate income from those videos. There are loads of factors included to manipulate the reputation of videos. YouTube has numerous features to assist users in finding interesting videos among a large number of videos, including a search engine, front page highlight, and related videos recommendation. YouTube can collect data on how viewers can interact with these features, which can help it improve its service. Active users who on a regular basis upload video content will be benefited.

YouTube is available in over 100 countries and 80 languages. More than 60% of businesses use YouTube to promote their products or services. 90% of digital market viewers in the United States use YouTube to share essential information. As a result, utilizing trending analysis data gathered from textual material, an approach may be devised to determine the YouTube user's perception of video content. The best option for interpreting the significance of each factor determining the success of a video in YouTube depends on the views of the video and how valuable the content that has been published.

This project aims to contribute to the advancement of trending analysis techniques by applying different machine learning algorithms like Logistic and Linear Regression, Naive Bayes, Random Forest, Decision Tree, and KNN with a Binary Classification approach. The usage of this method was chosen because it worked well on dataset and subjective classification. Linear Regression and Decision Tree algorithms are excellent at categorizing texts with a small number of data or document snippets, whereas Naive Bayes and KNN are excellent at classifying texts with a large number of data or a full-length document. The use of a mix of these models results in higher accuracy and performance.

II. Motivation

Over the years YouTube has expanded beyond just being able to access websites into mobile apps, network television, and the ability to link with other services. The inspiration

behind working on this project is the popularity the platform has gained among the public. Performing a trend analysis on YouTube videos and predicting the trend of videos in advance would help a large sector of people who target YouTube as one of the main sources for expanding their business and making profit.

It has received a major importance in the field of marketing which supports various business purposes. Business professionals have learnt that YouTube videos are one of the easiest means to increase their sales to make an impact of their products or brands on targeted audiences. YouTube being a content library makes it a powerful tool by keeping the trending videos easily viewable by many people. It is also a great platform for content creators to share their knowledge and ideas which would help the viewers to know more about their field of interests. Predicting the trend analysis would help the content creators to make the videos according to public demand and needs.

The purpose of this is to support creators and businesspersons and help them to understand about the public needs and preferences. They will also be able to build collaborations with various sponsors in advance if the video trends can be predicted in advance.

III. Literature survey

Several pieces of research have been performed on the YouTube platform as it is one of the largest social media platforms that share video content for people across the globe. Over the years it has grown from attracting 30,000 visitors to 100 million visitors per day. Data mining, NLP, text mining, spam detection, sentiment analysis are some of the areas of research that are performed on YouTube. Recommendations of YouTube videos have been researched and analyzed by a lot of researchers until now, but the analysis and research for YouTube trending videos still have a lot of scopes to be researched.

The literature survey is carried out in meaningful subsections based on our project goal as follows:

A. YouTube Trend Analysis:

Zhou [1] explored that the ordering of the videos in a related video playlist plays an important role in the click-through rate because they observed that there is a very high percentage of people who click on a specific video and then watch its related videos. They also inferred that there is a direct interrelationship between the view count of a video and the average view count of its top referred videos. Thus, they concluded that the impact on videos viewed is based on the YouTube video recommendation [1].

Figueiredo [2] studied the pattern of videos that emerge at the top in the YouTube playlist, videos that are deleted because of copyright issues, and the type of videos that get selected due to random searches. They inferred that popularity, and the trending of videos depend on a single type. The videos that emerge at the top list on YouTube tend to gain popularity at a faster rate when compared to other videos like the ones which have protection for copyright. In their research, they have used a multivariate linear (ML) model. Using this model, they have fitted different weights for the daily views of videos. To obtain better growth they have incorporated radial basis functions into the ML model [2].

Gabor [3] researched on two different social media platforms, YouTube and Digg. They observed that there is a linear relationship between the videos that are viewed early and the videos which gain popularity in the future. They used log transformation on the data to obtain this pattern of results. Through this study, they were able to predict the videos that trend for a longer time when linear regression is used on the initial data [3].

B. Feature Extraction:

There is a significant difference between Feature Selection and Feature Extraction. Feature selection deals with the ranking of the most important features that can serve to describe the whole dataset while Feature extraction is more based on dimensionality reduction which means reducing the number of features from the existing ones. Sayed et al. developed a new algorithm named chaotic crow search algorithm (CCSA) for feature selection. This algorithm is an improvised version of the Crow Search Algorithm (CSA) that boosts the low convergence rate by adopting a chaotic search method for optimizing features and maximizing classification accuracy. In this paper, the algorithm is tested on 20 different datasets and six different criteria for evaluation are adopted which are mean fitness value, best fitness, worst fitness, average feature selection size, p-value from Wilcoxon rank test, and standard deviation. The experiment concludes that the algorithm performance is better than other popular feature selection algorithms with regards to fitness value [4].

Bio-inspired algorithms have tremendously increased in recent years for optimization problems. These algorithms are developed inspired by biological species and their instinct. One such algorithm is the grasshopper optimization algorithm (GOA). Lewis et al. proposed GOA inspired by grasshoppers by mathematically modeling the species behavior. When the GOA is tested for performance and compared with other optimization algorithms, GOA is better in unimodal test functions as only one global optimum is present in these cases. The limitation of GOA is the convergence rate [5].

Mafarja [6] proposed two different approaches to the binary grasshopper optimization algorithm (BGOA). The first approach is developed by implementing transfer functions and the second approach considers the optimal position mechanism. The convergence rate is improved significantly as binary

bits are frequently changed. The algorithm is tested on 25 different datasets and compared with other algorithms. It can be concluded that BGOA is a superior algorithm that can be implemented in real-time problems [6].

C. ML Regression Algorithms:

As per the survey, in 2018 YouTube touched 1.5 billion users worldwide, and this number is predicted to extend to 1.86 billion by 2021. YouTube is the biggest platform providing a variety of video content in the world. The increasing number of videos uploaded on this platform is directly proportional to the increasing number of users registered on YouTube. The gist of working on this data is to classify the positive and negative content that is being uploaded daily as it is very crucial for the users to evaluate how useful the content is being published on the platform.

The combination of Naïve Bayes and Support Vector Machine algorithms to perform sentiment analysis is a very good approach. If we talk about Naïve Bayes, it is excellent in grouping texts with tiny numbers of data and document particles. Moreover, SVM is very good at categorizing texts with comparatively full-length documents. The conjunction of Naïve Bayes and SVM gives more accuracy levels and robust performance with the use of 70% training data and 30% testing data. As per the study, these algorithms will produce excellent performance test values, clarity of 91%, recall of 83%, and F1 score of 87%. In this project, authors contribute to the development of a sentiment analysis approach using the Naïve Bayes and SVM together that uses a binary classification method. The reason behind selecting the combination of these algorithms is the categorizing approach would be working better on the document extract and prolonged doc to break down sentiment, intuitive classification, and repeatedly better than the result that was published in the past [7].

D. Streaming Data:

In the recent decade, the machine learning techniques on the steaming data were mostly done using supervised learning like classification which aims to administer the data. The center of attention was only on developing the model to focus on precise decisions made from the steaming data. However, still, now there is no consensus of addressing the way of learning from the streaming data. To use the machine learning techniques on data streaming, there are several processes to be involved. The methods are data initialization, learning the dataset, and usage of steam data in the machine learning algorithms [8].

As we are aware, the initial step for machine learning is introducing the data to the model and cleaning the raw data. But, in this case, the cleaning cannot be done so easily because, for the steaming data, continuity is important. Any single job or instance of the steaming data will be related to the next data on the flow. So, it is important to combine the data for getting valid information. The inappropriate data in the pipeline may result in high noise and end up in making a failure model. The next stage is the learning phase, where the model should identify the important aspects of the given data. For example, what time is considered as high traffic for the website. The usage of this data from the model will help the researchers to identify the requirements of the users. There are many types of research carried out using other techniques using regression, unsupervised learning, and pattern mining. In further consideration, these techniques are enhanced with artificial intelligence to bring more efficiency out of the model [8].

E. Machine Learning: Algorithms, Real-World Applications and Research Directions:

According to the paper author mentioned, the digital world has a plethora of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, and so on, in this age of the Fourth Industrial Revolution (4IR or Industry 4.0). Knowledge of artificial intelligence (AI), particularly machine learning (ML), is required to intelligently analyze these data and construct the associated smart and automated applications. There are several types of machine learning algorithms in this field, including supervised, unsupervised, semi-supervised, and reinforcement learning. Deep learning, which is part of a larger family of machine learning technologies, can also effectively examine enormous amounts of data. In this paper, we provide a detailed overview of machine learning techniques that can be used to improve an application's intelligence and capabilities. As a result, the main contribution of this research is to explain the concepts of numerous machine learning approaches and their applicability in diverse real-world scenarios [9].

Types of real-world data: 1) Structured data: well-defined, structured data, easily accessible, highly standardized format, and data is stored in the table example name, id, address, phone number, other parameters and maintain its relational database property. It is easy to process. 2) unstructured data: Word processing documents, multimedia, blog entries, sensor data, emails, video, images, business data. It is difficult for analysis 3) semi-structured: HTML, XML, non-relational data but it is not stored like structured data set format. 4) Metadata: Data about data describes more relevant information about the data to the users.

The challenges and research direction: 1) Poor quality, useless features small amount of training data these factors lead towards the failure of the machine learning model.2) To select the particular perfect model is also a big task for the developers if they select the wrong model then results are desirable according to the client requirement and then there is a loss of effort.3) If they select the hybrid model and modify the rules of the existing model then it leads towards a challenging area. The success ratio of any machine learning algorithm model depends on its real data and the performance of the algorithm. To predict or classify the data into a successful

decision matrix mainly depends on the past, current data, and knowledge related to the application development [9].

IV. Methodology

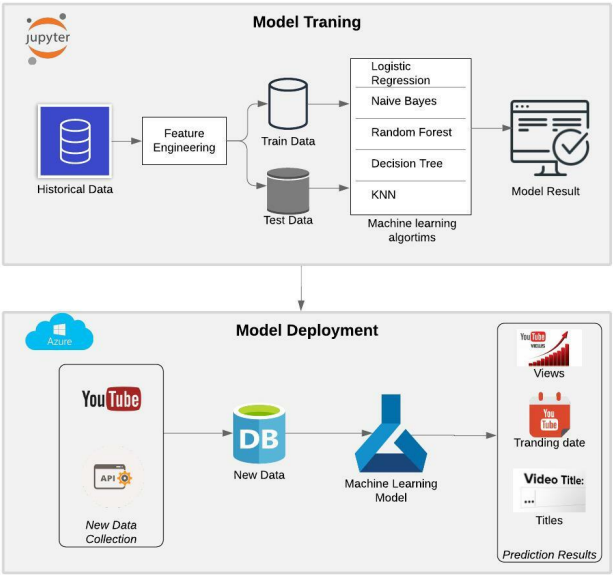

Fig. 1. Methodology

In this project we collected data from the Kaggle website, on the website data contains different region files in which we are using the USA and Mexico country data for further analysis.

The first CSV file contains USA region data and the other file contains Mexico region data. For this project we merged the CSV files. It contains 409409 rows and 16 columns.

video_id, category_id, trendind_date, publish_time, count of like, count of dislike, count of comments,tags,comments disables, description, video_removed counts,dislike_percentage, description.

The dataset contains a title JSON file in which it provides the information about Catergory_id, and category_name. In this project we convert JSON files to CSV. In the country CSV file it contains the category_id and in the title the csv file contains category_name. So we performed the normalization on the both CSV files on common category_id. The final file contains attributes such as video_id, category_id, trendind_date, publish_time, count of like, dislike_percentage, count of dislike, count of comments,tags,comments disables, description, video_removed counts, description and category_name.

On the above file we performed Exploratory analysis. By using correlation matrix it shows that all features are not important for implementing the algorithms. So we selected count of like,count of dislike, count of comments, category_id. The final data set size becomes (40949* 4).The next step is dividing data in the training data set and testing data set (80:20). After that we applied the logistic regression, KNN, decision tree,Random forest and Naive Bayes. Find the modal evaluation metric such as accuracy, F1 score, recall value. This part is called model training and for the model validation we used microsoft Azure services. YouTube provides the APIs to download the specific streaming metadata content. According to the project needs we created the Google API key and used it in YouTube open source APIs to download the new metadata of the trending videos.The streaming CSV file we applied to linear regression model which was implemented in Azure services provided in the demo section in the report.

A. Exploratory Data Analysis

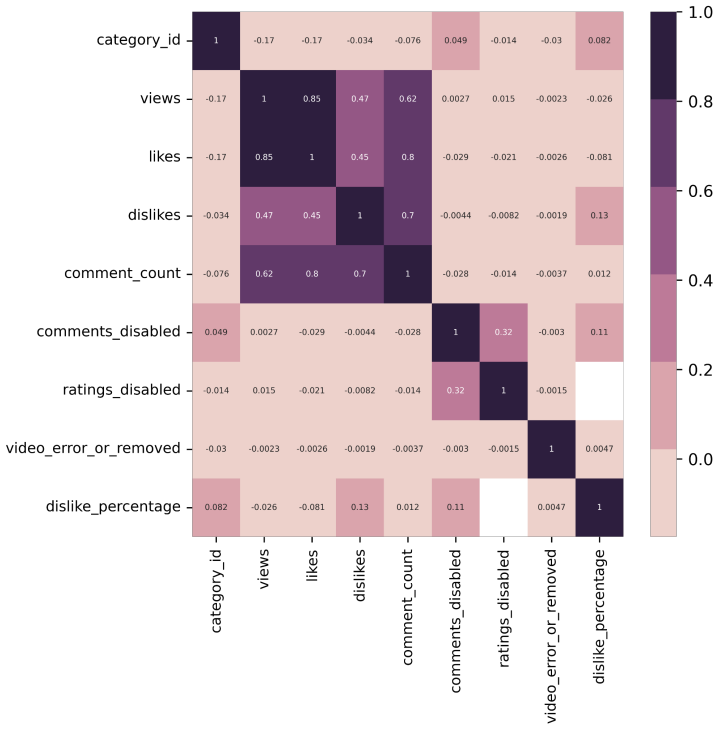

Fig. 2. Heatmap

A correlation matrix shows the characteristics and values between the attributes of the dataset. Every variable represents the connection with every other attribute. It is helpful while summarizing the data. With the help of a heatmap, we can show advanced insights. In the above figure (correlation matrix heat map), a positive correlation is shown by the dark shade of the color, and a negative correlation is shown by the lighter shade of the color. Here, Likes and Views have the highest correlation i.e 0.85. Also, Likes and Comment_Count have a correlation of 0.85. On the other hand, Dislikes and Rating Disabled, Likes and Video_error_removed, Ratings_disabled and Dislikes features have negative correlations. Hence, we

can conclude that Views, Likes, Dislikes and Comment_count attributes will be more useful in showing analysis and understanding the information provided.

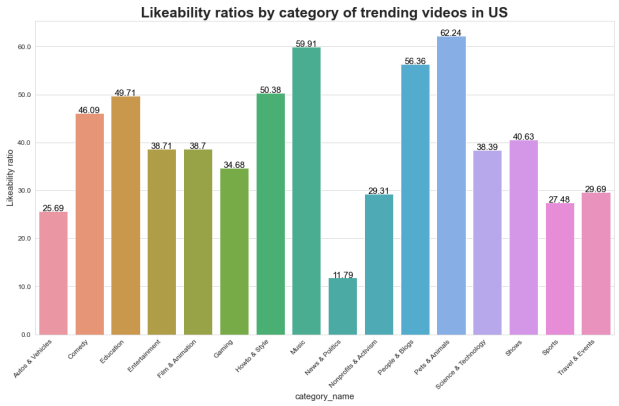

Fig. 3. Likeability ratio by category

As we can clearly see the most like video is based on pets and animals and followed by the music category. The lowest category video is news and politics. For this representation here we used the group by clause for the category section. According to this analysis, we can suggest upcoming Youtubers create new videos based on the high likability ratio. Where in the News and politics section people have different opinions regarding this particular topic that's why it has the lowest ratio.

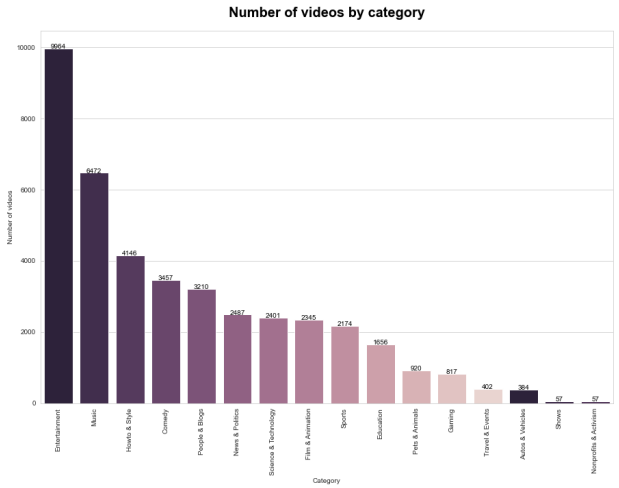

Fig. 4. Videos per category

The videos are segregated based on the categories which are directly proportional to their content. A bar graph is represented for showing the number of videos on various categories, on which it is observed that the highest count of the videos is uploaded under the category called entertainment. The videos related to the contents holding Story, Drama, Dance mainly fall under the entertainment category. Music stands as the second-highest count on the category list. This category includes the videos published by corporate music channels, independent music channels, and also, solo music channels. On the other extreme, it is observed that Nonprofits and Activism stakes a smaller number of videos. As the name indicates, this category videos are uploaded for non-profit activities or to create social awareness to the people. Mostly, the videos included in this category are those that have the content on campaigning, even promotions or appreciation to an activity. So, the category is one of the key aspects to be addressed while posting a video on YouTube since it tremendously contributes to the reach of the videos by providing suggestions to the appropriate audience who search for the videos in the same categories.

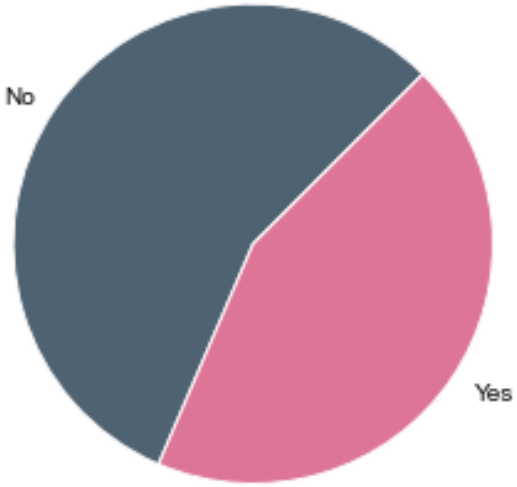

Fig. 5. Title with Capitalized words

A Pie chart has been used to represent the proportion of titles that contain capitalized words and the proportion of titles that do not contain capitalized words. Pie chart is used as it is easy to divide the categories into slices according to its numerical proportion. The arc length of each slice represents the proportion of both the respective parameters. The Pie chart is divided into a series of two segments: 'No' and 'Yes'. In the chart, No represents the proportion of titles that does not contain capital words and Yes represents the proportion of titles that contains capital words. As per our EDA analysis we inferred that about 56% of the total titles does not contain capitalized words and about 44% of the total titles contain capitalized words.

The video title length and number of trending videos distribution is plotted using python. It can be seen from the above figure that among the trending videos most of the videos have titles with medium length. Here the title length is obtained from the title name and it represents the number of characters used in the title of the video. Most of the Videos that are trending have used around 50 characters in its title. The length of title is also a viable factor and it should be in the mid-range which means lengthy titles should not be adopted while very short titles also do not completely convey the essence of videos.

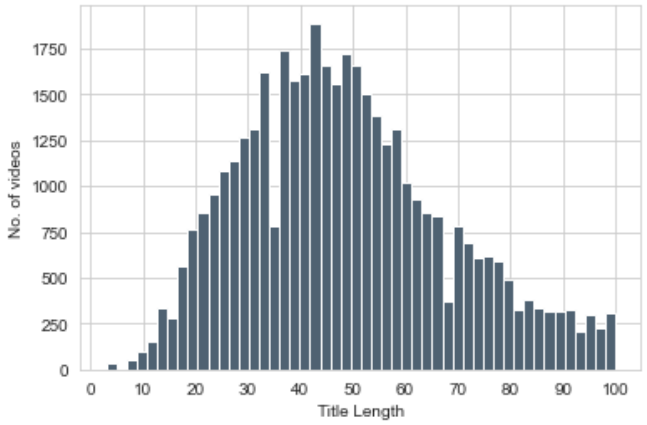

Fig. 6. Title length

V. Modeling

In this project we will be using 5 models for evaluation: Linear Regression, Logistic Regression, KNN, Decision Tree, and Random Forest.

Linear Regression is a machine learning algorithm which is used to understand the relationship between several input variables (X1, X2.... Xk) and a single output variable (Yt). The equation of linear regression is as given below where b0 is the bias and b1,b2...bk are coefficients of the input variable.

$$Yt = b0 + b1X1 + b2X2 + ... + bkXk$$

The input variables X are the independent variables, and the output variable Y is the dependent variable. In higher dimensions the classification line is called plane or hyperplane. To apply linear regression the relationship between the data needs to be linear. Linear regression by default assumes that the input and output are not noisy, to avoid outliers, data cleaning needs to be performed. When a data set has correlated input variables it always overfit the data to avoid that the high correlated values should be dropped out. To obtain accurate and reliable prediction it is required to perform normalization on the dataset by rescaling the input variables.[11]

K-Nearest Neighbor (KNN) is a supervised machine learning algorithm. To perform classification KNN follows a method of classifying the nearest neighbors. First an unknown new data point is picked and the closest data points around the new data point is picked using distance-based approaches.

There are 3 distance functions used to calculate the distance of the nearest neighbours

Euclidean Distance: It is the distance between two points in Euclidean space, it is the length of a line segment between the two points. The equation of the Euclidean distance is as shown below.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Hamming Distance: It is the distance between two strings which is of equal length, it is the number of positions at which the corresponding symbols are different.

$$d(a,b) = a \oplus b$$

Manhattan Distance: Manhattan distance is calculated as the sum of the absolute differences between the two vectors.

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Logistic Regression: Logistic model is a statistical model and is used to model probability of various classes. The representation used for Logistic regression:

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).

Below is an example logistic regression equation:

$$y = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. [14]

Decision trees: Decision trees are a tree-like model of decisions along with consequences, their utility, and their chance outcomes. There are two types of decision trees, namely, classification trees where the target variable (output variable) is discrete value and regression trees where target variable continuous value (real numbers). In the field of data mining, a decision tree describes data (but the resulting tree can be an input for the process of decision making) [12].

Random forest: Random forest is a learning method for regression and classification by constructing multiple decision

trees.

There are various Random forest algorithms, some of them are described below:

1) Bagging: This is a type of random forest training algorithm:

Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

Sample, with replacement, n training examples from X, Y; call these Xb, Yb. Train a classification or regression tree fb on Xb, Yb. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

or by taking the majority vote in the case of classification trees (3).

2) Random subspace method:: Random forests also include another type of scheme: they use a modified version of tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called ”feature bagging” [13].

VI. Model Implementation & Results

A. Prediction of views:

In this prediction model, the label data is “views” and input variables, or features are likes, dislikes, comments counts, category id, publishing hour, rating disabled. In the raw data set file 17 features to select the features here, we are using the correlation matrix. In the correlation matrix, it shows that views, likes, dislikes, comment count have positive values where all other features values are in the negative state.

Min Max normalization was performed on the data for the linear transformation. For the minimum value features, it transformed to 0 value and for the maximum value, it transformed to 1. The dependent variable means label and independent variable are features for the modeling here we are using a multiple linear regression model. The value of the R2 score is 0.78.

```
print("Coefficient of determination: %.2f" % r2_score(Y1_test, y_pred))
```

```
Coefficient of determination: 0.78
```

Fig. 7. Prediction of views

B. Prediction of video Category:

For training the model we need Video details (CSV) and video category (JSON) data. Here we applied the Naive Bayes algorithm for the training data set and the accuracy of the model is approximately 90%.

```
acc_nb = mn.score(X_test, y_test)

print('Accuracy', acc_nb)
```

```
Accuracy 0.8996336996336997
```

Fig. 8. Video category prediction

For prediction of video type here we used the count vector function in which the video title for the history data in converts into number.

```
titles = ["Barking cat plays with toy",
          "Best fashion looks for Spring 2018",
          "Olympics opening ceremony highlights",
          "Warriors basketball game versus the cavs",
          "CNN world news on donald trump",
          "Police Chase in Hollywood",
          "Ed Sheeran - Perfect (Official Music Video)",
          "how to do eyeshadow"
           ]
```

```
titles_counts = vector.transform(titles)

predict = mn.predict(titles_counts)

predict
```

```
array([15, 24, 17, 17, 25, 26, 10, 26])
```

Fig. 9. Video type prediction

In the above example here, we want to predict the video type for the incoming new data such as “Ed Sheeran” belongs to the musical category.

	Predicted Video Type	Video Title
0	Entertainment	Hilarious cat plays with toy
1	People & Blogs	Best fashion looks for Spring 2018
2	Sports	Olympics opening ceremony highlights
3	Sports	Warriors basketball game versus the cavs
4	News & Politics	CNN world news on donald trump
5	News & Politics	Police Chase in Hollywood
6	Music	Ed Sheeran - Perfect (Official Music Video)
7	Howto & Style	how to do eyeshadow

Fig. 10. New incoming data prediction

C. Prediction of Trending Days of a Video in YouTube:

In this project, the data collected from YouTube consists of the date and time the video is trending and the timestamp of the video when it is published on YouTube. These features are used to find the numbers of the days taken for a video to get on a trending list. Initially, these columns are transformed into the required date format. The published date is in a timestamp column which is divided into two columns separately representing date and time. When the two columns are in the same format the date difference is obtained and stored in new column which is named as Days_to_Trend.

```
date_diff = Trend_Video_US_CAT
date_diff['date_diff'] = date_diff['trending_date']-date_diff['publish_date']
Trend_Video_US_CAT['Days_to_Trend'] = date_diff['date_diff']/np.timedelta64(1,'D')
Trend_Video_US_CAT.drop(['date_diff'], axis= 1)
```

d	trending_date	title	channel_title	category_id	publish_time	views	likes	dislikes	comment_count	category	publish_date	Days_to_Trend
E	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	17:13:01	748374	57527	2966	15954	People & Blogs	2017-11-13	1.0
Y	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	07:30:00	2418783	97185	6146	12703	Entertainment	2017-11-13	1.0
4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	19:05:24	3191434	146033	5339	8181	Comedy	2017-11-12	2.0
Y	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	11:00:04	343168	10172	666	2146	Entertainment	2017-11-13	1.0
A	2017-11-14	I Dare You: GOING BALD!?	nigahiga	24	18:01:41	2095731	132235	1989	17518	Entertainment	2017-11-12	2.0
..	...	...	...	...	...	...	...	...	...	...	...	...
w	2018-06-14	The Cat Who Caught the Laser	AaronsAnimals	15	13:00:04	1685609	38160	1385	2657	Pets & Animals	2018-05-18	27.0

Fig. 11. New column creation

The raw dataset does not contain any labels and to transform this data from unsupervised to supervised, labels are defined. A new column named 'class' is added for class labels which defines the classes. Class 1 is defined for the videos that trend on the same day while class2 is defined for videos that took more than one day to trend after they are published. Later, the labels column is hot encoded.

```
#output variable is ready when video takes more than 1 day its tier 2
Trend_Video_US_CAT['class'] = np.where((Trend_Video_US_CAT['Days_to_Trend']>1), 'class2', 'class1')
Trend_Video_US_CAT['class']= Trend_Video_US_CAT['class'].map({'class1' : 0 , 'class2' : 1} )
Trend_Video_US_CAT.head()
```

e	title	channel_title	category_id	publish_time	views	likes	dislikes	comment_count	category	publish_date	date_diff	Days_to_Trend	class
4	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	17:13:01	748374	57527	2966	15954	People & Blogs	2017-11-13	1 days	1.0	0
4	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	07:30:00	2418783	97185	6146	12703	Entertainment	2017-11-13	1 days	1.0	0
4	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	19:05:24	3191434	146033	5339	8181	Comedy	2017-11-12	2 days	2.0	1
4	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	11:00:04	343168	10172	666	2146	Entertainment	2017-11-13	1 days	1.0	0
4	I Dare You: GOING BALD!?	nigahiga	24	18:01:41	2095731	132235	1989	17518	Entertainment	2017-11-12	2 days	2.0	1

Fig. 12. Adding a new label

The class label is used as a target to fit the model. This means that we can predict the number of days a video will take to trend after it airs on YouTube. To understand the implementation of different models, four models are implemented for this prediction. The dataset is divided into training and testing based on random splitting. The data is divided in such a way that 30% of data is allotted to testing while 70% data is trained. The models implemented are Logistic Regression, KNN, Random Forest and Decision Tree. Each of the models are fitted to the training and testing data. The predicted target value is determined.

D. Decision Tree Confusion Matrix, AUC Score and ROC Curve:

	Predicted Negative	Predicted Positive
Actual Negative	161	727
Actual Positive	671	10726

```
roc_auc_score(Y_test, y_pred3)
```

```
0.5612155818624627
```

Fig. 13. Decision Tree AUC score

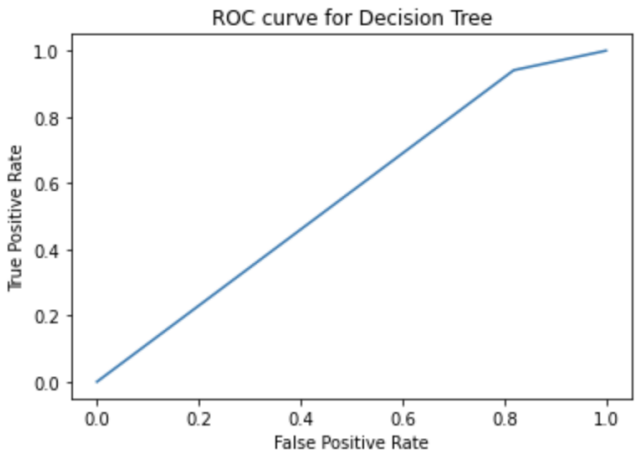

Fig. 14. Decision Tree ROC curve

E. Random Forest Confusion Matrix, AUC Score and ROC Curve:

The below fig. 15 and fig. 16 shows the AUC score and ROC curve of Random forest confusion matrix.

	Predicted Negative	Predicted Positive
Actual Negative	49	839
Actual Positive	48	11349

```
roc_auc_score(Y_test, y_pred2)
```

```
0.5254842727697426
```

Fig. 15. Random forest AUC score

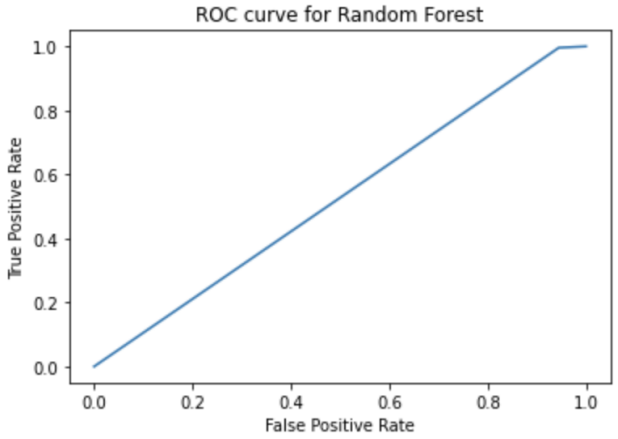

Fig. 16. Random forest ROC curve

F. KNN Model Confusion Matrix, AUC Score and ROC Curve:

	Predicted Negative	Predicted Positive
Actual Negative	25	863
Actual Positive	127	11270

```
roc_auc_score(Y_test, y_pred1)
```

```
0.5085049349164905
```

Fig. 17. KNN model AUC score

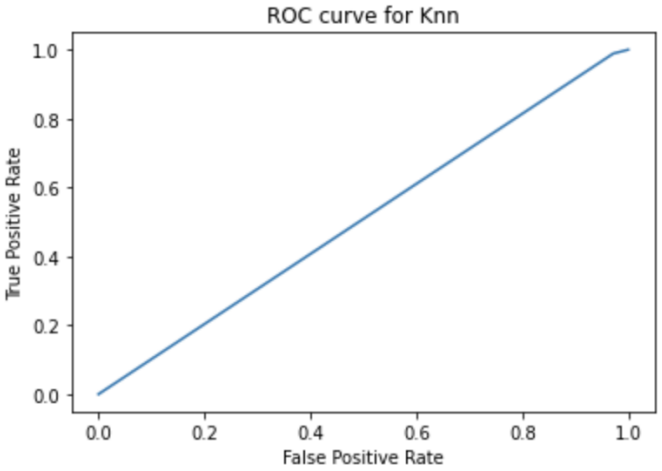

Fig. 18. KNN model ROC curve

It can be seen from the above confusion matrix that for each of the implemented models the true positive and predicted are high but when compared among the models, it can be seen that random forest has predicted correctly the majority of the data points while decision trees have lowest true positive cases. This provides an overview comparison between the models but for more detailed evaluation other metrics are evaluated.

```
Logistic Regression
-------------------------------------------
Accuracy Score: 0.927960927960928
Precision Score: 0.9278375733855186
F1 Score: 0.9617952835770435
-------------------------------------------
Knn Model
-------------------------------------------
Accuracy Score: 0.9224257224257224
Precision Score: 0.9277039225340554
F1 Score: 0.9587008141112618
-------------------------------------------
Random Forest Model
-------------------------------------------
Accuracy Score: 0.9219373219373219
Precision Score: 0.9277226911415226
F1 Score: 0.9593661285538748
-------------------------------------------
Decision Tree Model
-------------------------------------------
Accuracy Score: 0.8696784696784696
Precision Score: 0.9286776931828127
F1 Score: 0.9298576122672507
-------------------------------------------
```

Fig. 19. Overall model comparison

To perform the model evaluation, different metrics are evaluated like f1 score, accuracy, and precision. It can be seen from the model evaluation metrics that the accuracy of all the models fitted is around the same range of 92% except for the decision tree algorithm which has 86%. The f1 score and precision are also good indicating the model fitting is appropriate.

G. ML Model Demo in Microsoft Azure Studio

We are using Microsoft Azure for machine learning modeling demonstration. We have extracted the live streaming data from the YouTube platform. After extracting the file, we have cleaned the data and removed unnecessary columns. Now, the data is ready to be put in the Machine Learning Studio in Azure. A student account was created in Azure to perform the modeling on our prepared data. To perform the modeling section in Azure was very easy and we got accurate results. Microsoft Azure had mostly all the predefined algorithms stored in their dataset that we used to train and evaluate our data. We started with uploading a .csv file into the Azure studio. Following process was done by the studio tools only, we just needed to arrange the correct ordering of the tools, and the task was performed automatically by the tools. We uploaded the file, cleaned missing values (It was already done in Python. However, we added the tool to check the functionality. After that select the required features and then split the data into 80:20 ratio, 80% is for training and 20% is for the testing. After splitting the data, we applied the

ML model (Linear Regression). Once modeling is done, we score the model with testing data and trained data. At the end we evaluated the model with 73% accuracy. The set up looked like the figure below.

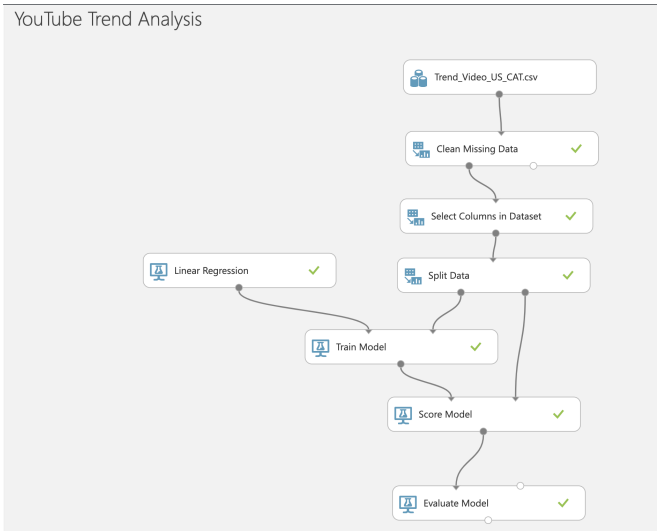

Fig. 20. Modeling in Azure

Mean Absolute Error	1213902.512663
Root Mean Squared Error	3651252.705845
Relative Absolute Error	0.460374
Relative Squared Error	0.264594
Coefficient of Determination	0.735406

Error Histogram

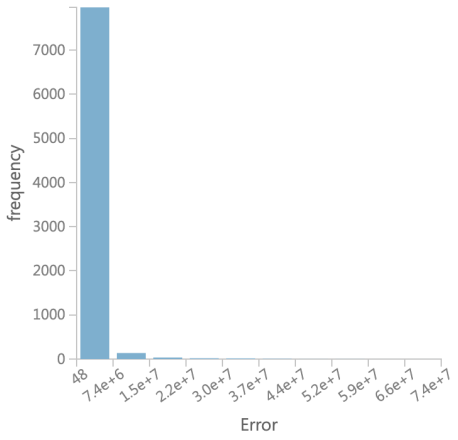

Fig. 21. Azure model evaluation and result

VII. Conclusion

YouTube has been a platform for a wide variety of content and sudden popularity of a video is often regarded as a random act while the exact analysis is still ambiguous but by applying machine learning models, the features can be determined, and the popularity index can be disintegrated into statistics. In this project, YouTube trending data is analyzed to comprehend the factors that define trending statistics of a video in YouTube.

The real-time streaming data from YouTube is extracted using YouTube API and python script. This real time data is used to conduct data exploration to find the correlations of the features extracted. Feature selection is performed based on the correlation and heatmaps. The prediction is performed for three different categories to fully determine the defining factors for a video to be on the trending list. The days a video requires to be on the trending list once it is published is predicted by feature extraction of date difference and trained the dataset for different models.

The models trained are Logistic Regression, K-NN, Random Forest and Decision Tree models have the accuracy in the same range of 93% except for decision tree models with 87%. This might be due to the fact that the decision tree requires a large dataset to fully fit the model. Moreover, prediction of views is evaluated to analyze the views of the videos that get on to the trending list in YouTube. The view prediction is performed by implementing linear regression, the dataset is transformed accordingly and the r2 score is evaluated. Finally, to predict the category of the video from the title of the video, a naïve bayes model is trained and a prediction accuracy of 90% is obtained. The Model evaluation depicts that the models have been appropriately incorporated. As the raw dataset is unsupervised, the Future work of the project work can adopt deep learning techniques for further fine tuning.

VIII. Lessons Learned

1) Using API and extracting real-time data for YouTube trend analysis.
2) Learned multiple EDA techniques to improve data understanding of variables by extracting different data characteristics.
3) Understood the detailed mathematical concepts of ML models and its implementation in obtaining the required results for projects in Microsoft Azure.
4) Comparison of different machine learning algorithms and finalizing the best model for the deployment.
5) Relationship between machine learning algorithms and machine learning models, based on the dataset applied to machine learning algorithms different models can be derived.
6) Learned about different evaluation metrics to understand the best model based on its performance.

REFERENCES

[1] R. Zhou, S. Khemmarat, and L. Gao, "The impact of YouTube recommendation system on video views," *in Proceedings of the ACM SIGCOMM Internet Measurement Conference - IMC,* 2010, pp. 404–410, doi: 10.1145/1879141.1879193.

[2] I. King, W. Nejdl, H. Li, and Association for Computing Machinery, "WSDM 2011 Hong Kong," *proceedings of the 4th ACM International Conference on Web Search and Data Mining,* Feb 2011.

[3] G. Szabo and B. A. Huberman, "Predicting the Popularity of Online Content," *SSRN Electronic Journal,* Dec. 2011, doi: 10.2139/ssrn.1295610.

[4] G. I. Sayed, A. E. Hassanien, and A. T. Azar, "Feature selection via a novel chaotic crow search algorithm," *Neural Computing and Applications,* vol. 31, no. 1, pp. 171–188, Jan. 2019, doi: 10.1007/s00521-017-2988-6.

[5] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper Optimisation Algorithm: Theory and application," *Advances in Engineering Software,* vol. 105, pp. 30–47, Mar. 2017, doi: 10.1016/j.advengsoft.2017.01.004.

[6] M. Mafarja, I. Aljarah, H. Faris, A. I. Hammouri, A. M. Al-Zoubi, and S. Mirjalili, "Binary grasshopper optimisation algorithm approaches for feature selection problems," *Expert Systems with Applications,* vol. 117, pp. 267–286, Mar. 2019, doi: 10.1016/j.eswa.2018.09.015.

[7] Slamin, Universitas Negeri Jember, Institute of Electrical and Electronics Engineers. Indonesia Section, "Computer Society Chapter, and Institute of Electrical and Electronics Engineers," *Proceedings, 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE 2019),* Oct 2019.

[8] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal, and J. Gama, "Machine learning for streaming data," *ACM SIGKDD Explorations Newsletter,* vol. 21, no. 2, pp. 6–22, Nov. 2019, doi: 10.1145/3373464.3373470.

[9] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science,* vol. 2, no. 3, May 2021, doi: 10.1007/s42979-021-00592-x.

[10] M. Mohsin, "10 Youtube Stats Every Marketer Should Know in 2019 [Infographic]," *Oberlo,* Jun. 27, 2019. https://www.oberlo.com/blog/youtube-statistics.

[11] S. Polamuri, "How the logistic regression model works," *Dataaspirant,* Mar. 02, 2017. https://dataaspirant.com/how-logistic-regression-model-works.

[12] Wikimedia Foundation, "Decision tree," *Wikipedia,* Sep. 28, 2019. https://en.wikipedia.org/wiki/Decision_tree.

[13] Wikimedia Foundation, "Random Forest," *Wikipedia,* Apr. 09, 2019. https://en.wikipedia.org/wiki/Random_forest.

[14] J. Brownlee, "Logistic Regression for Machine Learning," *Machine Learning Mastery,* Mar. 31, 2016. https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

APPENDIX

Github:https://github.com/Greeshma-Venkatesh/Data-245-ML-Project_-YouTube-Trend-Analysis

Criteria	Validation
Visualization Includes exploratory analysis (heat maps and other visuals)	Yes (Performed and the results are in the EDA section).
Presentation Skills Includes time management	Practiced and pending with the Professor.
Significance to the real world	Yes, YouTube is a global platform.
Saving the model for quick demo	Yes, Model Demo demonstrated during the project presentation.
Code Walkthrough	Yes, code walkthrough was done during the project presentation.
Report Format, completeness, language, plagiarism	Have performed all the mentioned checked and validation is pending on Professor
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	GitHub link is provided in the report https://github.com/Greeshma-Venkatesh/Data-245-ML-Project_-YouTube-Trend-Analysis
Discussion / Q&A	Time allotted and Pending on presentation day.
Lessons learned - included in the report and presentation?	Yes, it's included in the report.
Prospects of winning competition / publication	Pending with Professor.
Velocity Using streamed data in real time?	Have used Live Streaming Data.
Innovation	Adapted three different three different target problems.
Evaluation of performance	Model evaluation performed.
Teamwork	All tasks were divided equally and performed as discussed by all.
Technical difficulty	Live data extraction and model demo using Azure.
Practiced pair programming?	Yes, and included in Github.
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, other artifacts	Yes, provided the evidence.
Used Grammarly / other tools for language? Grammarly free version is sufficient; can use other tools as well. Submit report screenshot on Canvas.	Yes, used Grammarly and uploaded the proof in Canvas.
Slides	Yes, submitted.
Demo	Completed.
Latex	Yes, the entire report is prepared in latex.
Used creative presentation techniques animation, effects, newer features such as those offered by prezi, etc	Yes, used Prezi and presentations performed with newer features.
Literature Survey 1. Did not miss out on any important existing work that is relevant to the project. 2. Literature survey is organized into meaningful subsections 3. All references are cited and follow standard notation used in the template	Yes, all three points covered in the literature survey.