Prediction of Popularity of YouTube Video using Support Vector Machine & Random Forest

Hiu Ting Lau SID: 210665192

Abstract – This paper presents the prediction of popularity of YouTube videos with the video characteristics as the variables and the most relevant video characteristics for popularity prediction. The dataset is taken from Kaggle, which includes 13 video features, and pre-processed to generate the most relevant and suitable data for later processing. A new column 'popular' is added by inputting 1 for videos with top 25% of views and 0 for the remaining 75%, and thus the dataset is an imbalanced data. The finalized dataset contains 882 observations with 6 video characteristics left as variables. 2 classification machine learning models, support vector machine (SVM) and random forest (RM), are performed on the data to predict the popularity of the videos. The results show that RF performs better in the prediction of popularity than SVM with a small margin. Both models perform excellently by correctly predicting popular video with accuracy higher than 90%. Among all the video characteristics, number of likes is the most relevant features that help predict the popularity of video.

Keyword: YouTube video popularity, Support Vector Machine, Random Forest

1. Introduction

1.1 Background

Since the emergence of Web 2.0, social media which is based on user-generated contents has been promoted. Among all the social media platform, such as Facebook and Instagram, YouTube which focuses on video sharing, has gained its popularity since it launches. From 2012 to 2022, YouTube gained approximately 300% increase in active user from 700 million to 2.6 billon [1]. With 500 hours of videos are uploaded every minute to YouTube [2], not all the videos can become popular. To reach a high audience base in YouTube, there are 2 methods: paid promotion and YouTube internal recommendation. Paid promotion requires creators to pay for a duration of promotion with specified parameters. YouTube internal recommendation of videos depends on secret algorithms by YouTube. However, features considered for recommendation includes various video characteristics, including title keywords, number of likes and comments, which might lead to the 'rich-get-richer' phenomenon [3]. From a marketing perspective, the number of views in YouTube determine the audience reach.

Therefore, in order to maximize the probability of gaining high view counts and broad audience coverage for best promotion of brands, it is crucial for marketers and video creators to understand the factors that influence the popularity of their videos and to predict the popularity of the videos before they make any decision concerning video content and promotion, to stand out and promote the brand, and predict the popularity of the video based on the video characteristics and maximize the promotion and marketing ability of the videos.

In this paper, 2 models are performed to predict the popularity of YouTube video, which are support vector machine and Random Forest. Feature Importance are also extracted with Random Forest to present the extent of association between different video characteristics and video popularity.

1.2 Problem Statement and Hypothesis

The research questions are i) which model, support vector machine or random forest, gives a better prediction of the popularity of YouTube video, and ii) To what extent does each video characteristics influence the prediction of popularity of YouTube videos.

1.3 Aims

This paper aims to determine i) the best model in predicting popularity of YouTube videos and ii) the most influential video characteristics on predicting the popularity of YouTube videos.

2. Literature Review

YouTube determines popularity of a video based on the view count [4] and recommend videos to relevant audience based on a company-designed secret algorithms. Past literatures tried to predict the popularity of videos with sets of video and channel characteristics, including video-specific features such as number of subscribers, comments, likes and length etc. [3], [4], [5] and total number of views of all previous videos in the channel as an indicator of past performance of a channel. Inconsistent conclusions are made, with the number of previous views, number of comments and video age being the generally most significant factors affecting popularity.

Models for prediction on classification of popular video and online content were also conducted by past literatures. Video-specific features are included in the analysis for prediction with different approaches in the definition in the indicator of popularity. In [8], popularity of videos was determined by classes of the ratio of number of likes to number of views. 3 classes, high, medium and low,

are defined according to the calculated popularity parameter. In [7], Li et.al. defined popularity by taking the number of comments, views, likes and dislikes into the calculation of popularity score. 4 classes are defined according to the calculated popularity score. Various models are tested, including Random Forest, XGBoost, Support Vector Machine and Logistic Regression etc. [6], [7], [8] Among all models, XGBoost presented the highest accuracy, followed by Random Forest, Support Vector Machine and Logistic Regression [6]. All models presented similar accuracy in prediction on classification of popular news content. In [7], Li et. al. presented a different conclusion. While the ranking of prediction accuracy among different models remained the same, XGBoost shows a more significant strength in classification prediction of video popularity relative to Random Forest and Decision Tree. In [8], Batta et. al. added on to the previous conclusion and presented the ranking among XGBoost, Random Forest, Decision Tree and KNN Classification. The ranking remained the same with KNN classifier being the model with lowest accuracy.

3. Data Processing

Open-source data are taken from from Kaggle [9].

3.1 Description of Data

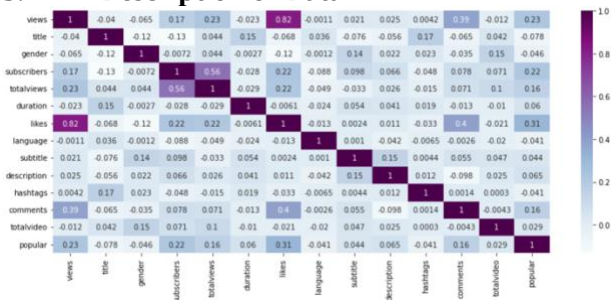

Figure 1. Correlation between Variables

Variables with high correlation do not bring additional information to the analysis. Since there are 12 variables in the dataset, correlations between variables are studied to avoid the analytical complexity and dimensional problems. From figure 1, totalviews and subscribers have the highest with moderate correlation (0.56), but it is still acceptable.

Figure 2 present s the distribution of different classes among the variables ‘gender’, ‘subtitle’, ‘language’ and ‘description’. Gender is dominated by the class 1 ‘Male (54.8%) and presents a significantly low proportion of female creators (8.8%). Subtitles and description are available for most of the videos with both percentage of class 1 being higher than 50%. The dominant language used in videos is English, followed by Hindi which fall off with a large margin.

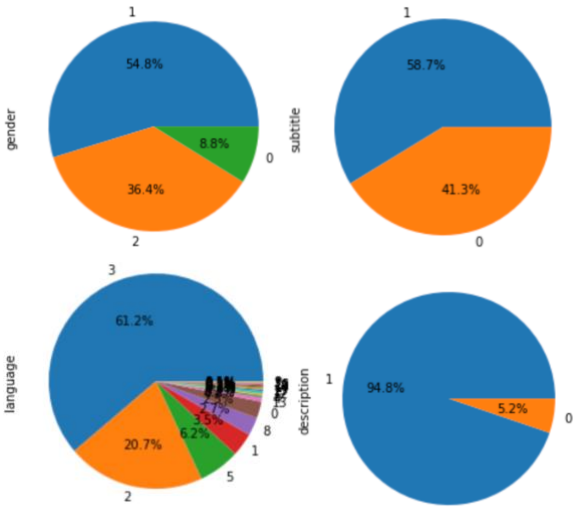

Figure 2. Pie Chart for distribution in gender, subtitle, language and description

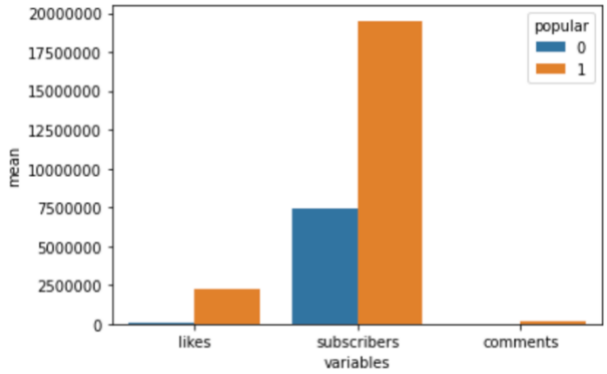

Figure 3. Means in ‘Popular’ and ‘Unpopular’

Means of likes, subscribers and comments are calculated for group of popular and unpopular video. Figure 3 shows that the mean of all variables in group popular (1) are higher than group unpopular (0). Subscribers shows the highest difference (nearly 70%) in mean with mean subscribers = 19,511,136 in group 1 and 7,415,069 in group 0.

3.2 Pre-Processing Steps

3.2.1 Renaming

Data are renamed for easy referencing and misspelling correction as shown in Table 3 – Assigned Name.

3.2.2 Conversion of Categorical Variables to Numeric Variables

Conversion is performed on 4 variables: gender, description, subtitle and language. For gender, Female = 0, Male = 1, Company = 2 are assigned. For description and subtitle, No = 0, Yes = 1. For language, 'Kannada' = 0, 'Telugu' = 1, 'Hindi' = 2, 'English' = 3, 'Japanese’ = 4, 'Tamil' = 5, 'Urdu' = 6, 'Punjabi' = 7, 'Malayalam' = 8, 'Mongolian' = 9,

'Russian' = 10, 'Italian' = 11, 'Hindi/English' = 12, 'Sanskrit' = 13, 'English + Hindi' = 14, 'Konkani' = 15, 'Arabic' =16, 'Urdu + Arabic' = 17, 'Nawayathi' = 18 are assigned.

3.2.3 Transformation of Type and Values

Since title is a string which cannot be analyzed with the algorithms, it is transformed into 'number of words in the title'. Commas in views, durations, totalviews and comments are also deleted.

3.2.4 Null Value Handling

Null values in gender, likes, language and subtitle are handled. There are 314 null values for gender, 21 for language, 1 for likes and 1 for subtitle. For gender, 'company' is assigned in replace of null values by manual observations of channel name. For likes, language and subtitle, instances with null values are dropped due to their small number of null values compared to the sample size (1 for subtitle and likes, 21 for language).
Lastly, all variables are converted to type int64 for easy reading of table and data analysis.

3.2.5 Adding of an Indicator of Popularity

Column 'popular' are added with popular = 1 for videos with top 25% of view count, and popular = 0 for the remaining videos. Column 'views' is dropped.

3.3 Feature Selection

Both test are conducted from scipy.stats. Through the process of feature selection from the result of One-way ANOVA and chi-square test, duration, hashtags, totalvideo, gender, subtitle and description are dropped from the dataset.

3.3.1 One-way ANOVA on Numeric Variables

Variables	p-value	Variables	p-value
title	0.020	likes	1.599e-21
subscribers	7.627e-11	hashtags	0.224
totalviews	2.951e-6	comments	2.035e-6
durations	0.076	totalvideo	0.392

Table 1. One-Way ANOVA Results

Numeric variables include title, subscribers, totalviews, duration, likes, hashtags, comments and totalvideo. One-way ANOVA is used to test the presence of statistical difference in population means among popular and unpopular videos [10]. The null hypothesis is thus population means of video characteristics of popular and non-popular videos are the same. The alternative hypothesis is there are differences in the population mean among popular and unpopular videos. Null hypothesis will be rejected if p-value is lower than 0.05. From table 1, p-values of hashtags, duration and totalvideo are below 0.05 and show that there is no significance difference in the popular mean of these variables between popular and unpopular videos, and thus the 3 variables are dropped from the dataset.

3.3.2 Chi-Square on Categorical Variables

Variables	p-value	Variables	p-value
gender	0.386	subtitle	0.224
language	5.275e-5	description	0.079

Table 2. Chi-Square Test Results

Categorical variables include gender, language, subtitle and description. Chi-square test is used to examine the independence between 2 categorical variables [11], which are popular and variables in this case. The null hypothesis is that there is no relationship between the video characteristics and 'popular', and the alternative hypothesis is there is relationship between the video characteristics and 'popular' in the population. Null hypothesis will be rejected if p-value is lower than 0.05. From table 2, p-values of gender, subtitle and description are below 0.05 and show that there is no relationship between these video characteristics and 'popular', and thus the 3 variables are dropped from the dataset.

3.3.3 Finalized Dataset

Variables	Assigned Name	Description
Video Title	title	Title of the video
Total Channel Subscribers	subscribers	Number of channel subscribers
Total Chanel Views	totalviews	Number of views of all videos uploaded on the channel
No of Likes	likes	Number of likes of the video
Language of the Video	language	Language used in the video
No of Comments	comments	Number of comments of the video
	popular	0 = unpopular, 1 = popular

Table 3. Description of Video Features for Finalized Dataset

The finalized dataset includes 6 video characteristics: title, likes, subscribers, totalviews, comments and language with 882 observations. The dataset is standardized with StandardScaler from sklearn.

4. Learning Methods

2 supervised learning methods are used in this paper, support vector machine (SVM) and Random Forest (RM). Supervised learning is used as data are labelled and the algorithms allows a more accurate outcome. In this paper, the column 'popular' is predicted with the remaining columns as variables. Classification models with 2 classes are used to predict whether a video will be popular (= 1) or unpopular (= 0). Classification report, confusion matrix and ROC curve will be generated from both models. Classification report states the precision, recall, f1 scores and accuracy of the model in predicting the classes. Precision states the how many positive predictions are correct. Recall states the how many of the positive observation has the model correctly predicted, and f1-score is an average of the precision and recall score. The formulas of these are given below. Confusion matrix helps explain the scenario in classification report. ROC curve shows the ability of the models in classifying popular and unpopular videos with AUC (Area Under Curve) score relative to the default curve of random estimation (50%). Additionally, feature importance is generated from the RF model to show which variables are of most influence on the prediction of popularity. SVM and RF are selected based on the following algorithm-based features and the evidence of accurate prediction shown from past literature.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positve}$$

$$Recall\ = \frac{True\ Positive}{True\ Positive + False\ Negatives}$$

$$f1\ score = \frac{Precision * Recall}{Precision + Recall} * 2$$

4.1 Support Vector Machine (SVM)

SVM is an algorithm that can be used for binary classification of 2 classes, in this case, popular and unpopular by determining the optimal hyperplane. With the closest data point to the hyperplane as the supporting vectors, the optimal hyperplane is determined by maximizing the range from supporting vectors. Since the data are non-linear, kernel function with the best hyperparameters is used to determine algorithm for the optimization of the process of n-dimensional classification. SVM is selected since it is popular to use with complex and small dataset with its high predicting power.

4.2 Random Forest (RF)

RF is an algorithm that can be used for classification. Hyperparameters are required to be set before running RF, including the node size (max_depth), number of trees (n_estimators), the function to measure the quality of a split (criterion) and the number of features sampled (max_features). After that, RF works on the bagging principle which begins by selecting randomly a subset of samples from the training data and construct decision tree for each subset. The data are separated into different tree based on the criterion in the nodes of the trees. The final output will be based on the majority voting according to the output from each decision tree. RF is selected since the prediction is less influenced by outliers and feature importance can be generated since the model is always robust, and thus gives a more comprehensive and accurate prediction.

5. Analysis, Testing and Results

5.1 Cross Validation & Hyperparameter Selection

Before running SVM and RF on the dataset, cross validation is conducted to determine the predictive accuracy of SVM and RF model. Hyperparameters are also selected through cross validation with GrindSearchCV from sklearn. GrindSearchCV selects the best hyperparameters for both models from the given set, and thus does not require manual decision on hyperparameters selections. The dataset is split into 10% testing data and 90% training data for a 10-fold cross validation.

Average accuracy of cross validation is 0.909 for SVM and 0.928 for RF as shown in table 4. This implies that the predictive accuracy of the 2 models is good with RF outperforms SVM with a small margin. The average cross validation accuracy will later be compared to the accuracy after running the model.

Best hyperparameter for SVM are regularization parameter C = 10, gamma = 10 and kernel = 'poly'. For RF are n_estimators = 100, criterion = 'gini', max_depth = 10 and max_features = 'sqrt'.

	SVM	RF
Average CV Accuracy	0.909	0.928

Table 4. Average CV Accuracy for SVM and RF

5.2 Class-specific Results

	precision	recall	f1-score	accuracy
0	0.97	0.94	0.96	0.93
1	0.81	0.89	0.85	

Table 5. Classification Report with SVM

	precision	recall	f1-score	accuracy
0	0.97	0.96	0.96	0.94
1	0.85	0.89	0.87	

Table 6. Classification Report with RF

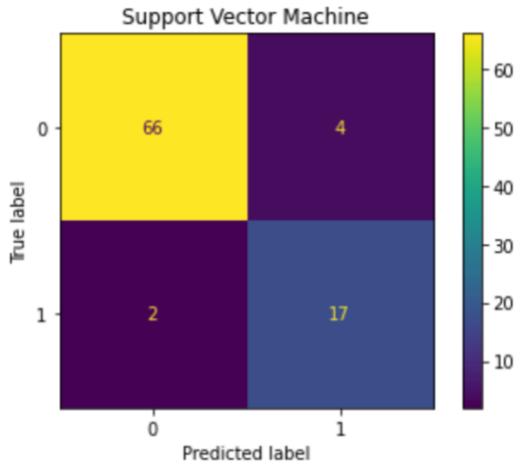

Figure 4. Confusion Matrix with SVM

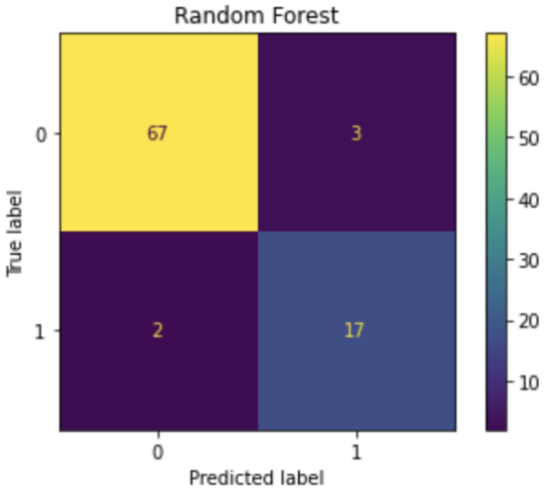

Figure 5. Confusion Matrix with RF

From the result of classification report in table 5 and table 6, it shows that both SVM and RF model make a good prediction on the training data with high accuracy (93% and 94% respectively). RF model makes a slightly more accurate prediction than SVM model. The results are consistent with the average CV accuracy, which show that both models predict constantly well with different data set.

Generally, both models perform excellently in predicting popular videos with fair precision and recall score higher than 80%. F1-score are also high for both models, indicating that both SVM and RF make accurate prediction on popular video. It should also be noted than precision score of RF model (85%) is higher than the precision score of SVM model (81%). This means that RF model is highly accurate in correctly predicting a popular video and outperformed SVM. For the ability in identifying popular video, SVM and RF model perform at the same excellent level with 89% recall score.

By investigating deeper on the classification report, the higher precision and recall score in row 0 than in row 1 shows that both models perform worse in predicting and identifying popular videos than unpopular video. This shows that both models are better in capturing the characteristics of unpopular video than that of popular video.

The above discussion are also shows in figure 4 and figure 5 with the confusion matrix with results from SVM and RF model on testing data.

5.3 AUC-ROC Curve

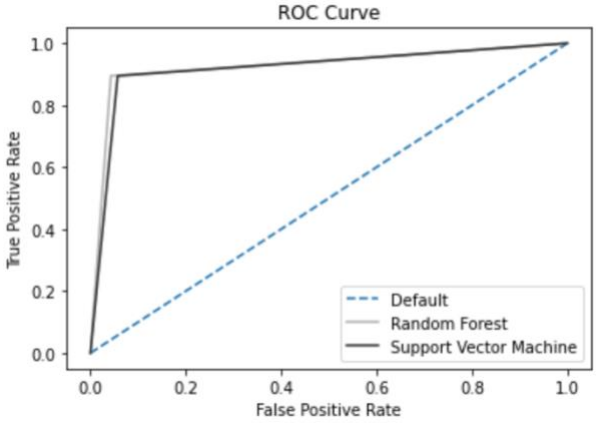

Figure 6. ROC Curve with SVM and RF

AUC scores are 0.919 for SVM model and 0.926 for RF model. Since both scores are higher than 90%, both models perform excellently in classifying popular and unpopular songs, with RF model again outperform SVM model with a small margin. This is also evident from figure 6 that area under the RF's ROC curve is slightly larger than that of SVM's RC curve.

5.3 Feature Importance

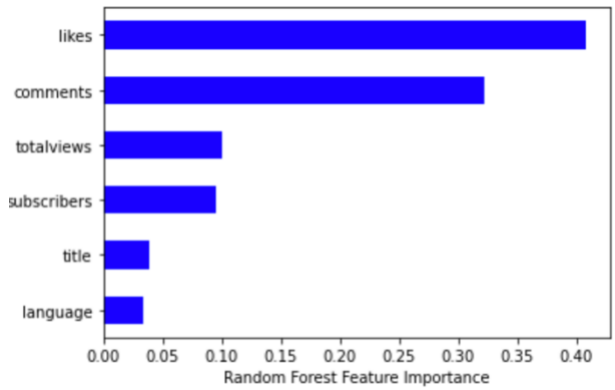

Figure 7. Feature Importance with RF

Feature importance is generated with RF model, indicating the level of relevance of each video characteristics to the popularity of video. From figure 7, it shows that number of likes and comments are the most relevant variables associated to the popularity of video, each with importance of 0.409 and 0.322 respectively. Their relevancies outweigh

with a large margin comparing to the rest of the video characteristics. Meanwhile, with the top 4 being the number of likes, number of comments, total views count in the channel, and number of subscribers, it shows that video popularity is heavily predicted by the audience interaction with the video and the channel. It should be noted that video creators-defined variables, i.e., title and language, are of the lowest importance with much lower level of importance compared to the top 4 video characteristics. This implies that video creators cannot easily predict the popularity of videos by deciding length of title or language used.

6. Conclusion

6.1 Main Findings

In conclusion, the research questions and aims of this paper have been successfully answered and achieved. The first and major aim of this paper is to determine the best model for the prediction of popularity of YouTube videos by comparing the accuracy of and relevant scores generated from support vector machine and random forest model. The average cross validation accuracy gives a first glance of the predictive power of the model on unseen data. Both cross validation accuracy from SVM and RF are higher than 90% which indicate excellent prediction, with RF outperforms SVM with a small margin. The excellent predicting power of SVM and RF are then proven with the high accuracy and scores from the classification report. The accuracy for both models are higher than 90% with precision scores higher than 80%. The results are consistent with the average cross validation accuracy. Meanwhile, the accuracy and precision score with RF model are all slightly higher than that with SVM model. This indicates that RF model presents a slightly more accurate prediction on the popularity of video given set of video characteristics data. The output of ROC curve also agrees with the results from the classification report. AUC scores of both models are higher than 90% and indicate that both models perform excellently on classifying popular and unpopular videos. Again, RF has a slightly higher AUC score than SVM and shows that RF is a better model in both predicting and classifying the popularity of YouTube videos. The results are consistent with the past literature [6], presenting SVM is a better model in predicting the popularity of a YouTube video than SVM.

The second minor aim in this paper is to determine the most relevant video characteristics associated with the popularity of video. This is primarily done by the display of feature importance from the RF model. The results shows that number of likes and number of comments are the 2 most important features in predicting the popularity of the video. This is partially consistent with the result of number of comments being most influential feature from past literature [4]. However, number of total views of the channel is not as importance in this paper when compared to the result from past literature [3]. The total views ranked third with a much lower importance than number of comments in predicting the popularity of videos. Meanwhile, all 4 audience-determined video characteristics, i.e., number of likes, comments, total views and subscribers, rank higher than the creator-determined video characteristics, i.e., length of title and language. This shows that the prediction of popularity of YouTube videos is highly dependent on the interaction and reaction by the audience, and thus video creators do not have much available adjustments that can significantly influence the prediction of the popularity of the videos. This is also consistent with the results from past literature [4] as mentioned above.

6.2 Limitations

6.2.1 Comprehensiveness of the Dataset

The dataset is primarily focusing on the video and channel characteristics. This means that social network of the video creators is neglected from the analysis. In general, people who are more famous or have a larger social network can relatively gain a higher popularity in YouTube due to their celebrity effect. Since the samples are taken randomly, the size of social networks of the video creators are unknown, and this might affect the number of views they gained, and in turns affect the prediction of popularity.

Moreover, the dataset ignores the wording used in title, description and hashtags. In the dataset, title is determined by the length of title, description is equal to 0 or 1 according to its availability, and hashtags is determined by the number of hashtags used. The lack of analysis of wordings in these 3 variables might affect the prediction of popular videos. YouTube recommends videos to users based on the relevance of videos that users generally watched. This is mainly based on referencing similar title, or similar content as indicated in the description and hashtags. Therefore, wordings used in these 3 variables might influence the popularity of videos.

Lastly, categories of the videos are not included in the dataset. YouTube categorize different videos into different categories according to their genre. The lack of category might lead to the same abovementioned problem due to the YouTube recommendation system.

6.2.2 *Classification of Popularity*

In this paper, the popularity of video is solely determined by the number of views. Videos with the top 25% views count among the dataset are classified as popular video. This approach neglects the sentiment of the audience have towards the videos, whether it is a positive or a negative popularity. For marketers and video creators, negative popularity should be avoided which might being adverse effect on brand image and business.

6.2.3 *Machine Learning Models Used*

Only 2 machine learning models are evaluated in this paper. According to the past literature [6], XGBoost is the best model in predicting popularity. Thus, this paper fails to include more classification models for prediction power comparison. Meanwhile, the processes for SVM and RF model are time consuming for both cross validation and model running. More efficient machine learning algorithms, such as logistic regression, might be a better option considering the time-accuracy tradeoff.

6.3 Possible Future Improvements

As mentioned in the limitations, approaches can be taken for future improvement.

First, social network of video creator can be considered with the number of followers in other social media platform as the indicator.

Second, natural language process (NLP) can be utilized to analyze the wording used in title, description and hashtags to investigate the association between the wordings to the popularity of videos. By identifying the common wordings used in popular videos, it will be convenient for video creators to generate title, description and hashtags, and eventually promote without cost through YouTube recommendation system. This helps reduce operation cost and broaden the audience base.

Categories of videos should also be included to identify the size of the audience base of video in different genre. Also, the sentiment of audience should be considered for the process of defining popular videos. The number of likes and dislikes are good simple indicator for the sentiment of audience towards the video. Past literature tried to include the sentiment of video in the definition of popularity parameter by including number of likes and dislikes in the calculation [7], [8]. Apart from considering the sentiment of audiences, the sentiment of video content can also be considered in the prediction model. Sentiment analysis can be conducted and define each video with class of sentiment. This helps in deciding what sentiment of video content has the highest association with the popularity of video and if the sentiment of video content influence the prediction of popularity.

Lastly, more machine learning model can be conducted with the dataset to give a comprehensive comparison of accuracy and runtime between different models. For example, XGBoost, Decision Tree, Logistic Regression and KNN Classification.

7. References

[1] Curry, D. (2023). Social App Report 2023: Revenue, Usage and Demographics for Major Social Platforms.

[2] Statista (2023). Hours of Video Uploaded to YouTube Every Minute as of February 2020.

[3] Borghol, Y., Ardon, S., Carlsson, N., Eager, D., & Mahanti, A. (2012). The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity.

[4] Chatzopoulou, G., Sheng C. & Faloutsos, M. (2010). A First Step towards Understanding Popularity in YouTube.

[5] Welbourne. D. J. (2015). Science Communication on YouTube: Factors that Affect Channel and Video Popularity.

[6] Khan, A., Worah. G., Kothari, M. Jadhav Y. H. & Nimkar A. V. (2018). News Popularity Prediction with Ensemble Methods of Classification.

[7] Li, Y., Eng, K., & Zhang, L. (2019). YouTube Videos Prediction: Will this Video be Popular?

[8] Batta, H., Murthy, A. V., & Savitri S. (2022). Predicting Popularity of YouTube Videos Using Viewer Engagement Features.

[9] Jayachandiran, K. (2022). YouTube Influencer Data. Kaggle. Available at: https://www.kaggle.com/datasets/kathir1k/youtube-influencers-data

[10] Kim, T. K. (2017). Understanding One-way ANOVA using Conceptual Figures. Korean Journal of Anesthesiology.

[11] Rana, R. & Singhal, R. (2015). Chi-square Test and Its Application in Hypothesis Testing.