Examine How Content-unaware Attributes Influencing the Video Views of YouTuber

Chenrui Feng*

Department of Art and Science, St George Campus, University of Toronto, Toronto, Ontario, M5S 1C6, Canada

*Corresponding author: chenrui.feng@mail.utoronto.ca

Abstract. YouTube is one of the most popular apps worldwide, and many people create channels to earn income as YouTubers. This research focuses on identifying content-unrelated attributes that impact video views and examining their influence. Initially, data cleaning is conducted to remove missing observations that are not useful for model fitting. Data exploration is performed to gain a basic understanding of the relationships between variables and the response, which may aid in variable selection and model construction. Linear regression models are then fitted to predict video views based on content-unrelated variables. In this paper, these models are compared to select the best one, followed by diagnostics using AIC, R-squared, and MSE. Additionally, the selected model is evaluated for its fit by checking linear assumptions, the presence of outliers, and the significance of predictors. Finally, potential drawbacks of the methods used in this research are discussed, along with consideration of future research directions.

Keywords: Video views; YouTubers; linear regression model; model selection; diagnostic.

1. Introduction

Social media has significantly influenced lives in recent years, leading to the invention and proliferation of numerous apps. One of the main drivers is the desire to reduce the information gap, with individuals from higher-income households and higher education levels tending to use social media more frequently [1]. Among these platforms, YouTube stands out as one of the most popular worldwide. In response to high viewer demand, many individuals create channels and use the platform to generate income [2]. Video views are a key metric for YouTubers, as they directly impact popularity and income potential [3]. However, achieving high popularity is challenging due to increasing competition among YouTubers. It is, therefore, crucial to understand the factors that enhance channel popularity, which is often reflected in video views.

While many creators focus on improving video content quality to boost views, external factors unrelated to content may also play a role but are often overlooked [4]. This research examines content-unrelated attributes that may influence YouTubers' video views. These factors are widespread but difficult for YouTubers to alter directly, as they are often not immediately apparent. Identifying them requires attention to video and account details that can impact channel popularity. Predicting YouTuber popularity based on content-unaware factors such as channel title, description, and thumbnail-elements not typically included in traditional popularity analyses-can provide new insights [5]. The interaction between YouTubers and their audience, including video pace, subscriber count, and the number of uploads, is also crucial in predicting video views [6]. Additionally, social behaviors, along with cultural and geographical factors like latitude and language, should be considered in the analysis of popularity [7].

Video length, another important factor, affects video views since viewers often prefer shorter content due to limited time [8]. Emotional attachment expressed by YouTubers in their videos can also impact viewer mood and, consequently, video views [9]. In many cases, older channels tend to receive more views than newer ones, though newer YouTubers can still gain popularity [10]. Certain channel categories are more attractive than others, resulting in higher video views for those focused on popular topics. Although many content-unrelated factors influencing video views cannot be controlled by YouTubers, understanding their effects remains valuable, as it can help creators focus on other factors that enhance their popularity. This research aims to explore the relationship between content-unaware attributes and YouTubers' video views.

In summary, after consideration and optimization, as video views of YouTubers is numerical variable, it would be suitable to use linear regression model to predict the video views of YouTubers by some content-unaware factors.

2. Methods

2.1. Data Description

The data comes from the Kaggle website, specifically the dataset titled "Global YouTube Statistics 2023," containing 996 observations. This sample size is sufficient for conducting data analysis and identifying general trends to address the research questions. The dataset includes 28 variables, with many valuable attributes unrelated to the content of videos, as they pertain to the main features of the YouTuber. These characteristics make the dataset especially useful for this study.

2.2. Variable Introduction

After uploading the dataset, it is essential to check for missing values and remove them using the `na.omit()` function to ensure that all attributes for the observations are valid. A new dataset is then created by adding meaningful variables derived from the original attributes. Video views are selected as the response variable for the project, with a focus on identifying content-unrelated attributes to predict or influence YouTuber video views.

First, the continuous variable Latitude of YouTubers is transformed into categorical variables to better represent the geographical features of different regions. The variables Country and Channel_type are also reclassified into English vs Non-English and Music vs Non-Music, respectively, as both the English and Music channels dominate the platform. The urban_population_rate and unemployment rate of the YouTuber's country are considered important external factors, which YouTubers cannot change, but may still influence video views.

In addition to these external factors, highest_yearly_earnings, uploads, and subscribers are selected as variables corresponding to the YouTubers themselves, to detect their effects on video views. A new meaningful variable, Age_of_channel (calculated as 2023 - created_year), is created as a special feature of the channels and is included in the project for further analysis.

2.3. Model Introduction

In this process, first split data to train and test part under the condition of random sampling, then use train data to fit the model, use test data to check the prediction error. In addition, perform the exploratory data analysis to view the relationship between attributes and predictors, which is meaningful to build model in the next stage.

The model used in this research is a linear regression model, as video views of YouTubers serve as a numerical variable. Four models are built in total. The first three models are relatively simple, differing only in the number of predictors included, while the final model is more complex as it incorporates interactions between variables. A comparison between these models is conducted by examining AIC, R-squared, and MSE. In general, a model with a higher R-squared, lower AIC, and lower MSE is considered superior.

Additionally, it is essential to diagnose the final model by verifying the assumptions of linear regression, detecting any influence from outliers or influential points, and using F-statistics and p-values to ensure the model's fit and the significance of predictors.

3. Results and Discussion

3.1. Random Sampling

According to Table 1 and Table 2, the mean and median of all numerical predictors were first calculated for both the training and testing groups and then compared. The results indicate that the mean and median for the two groups are similar for nearly all variables, confirming that random sampling is adequately achieved. This similarity minimizes bias in fitting the model using the training data.

Table 1. Mean of attributes for observations in training group and testing group

Variables	Training group	Testing group
video views	12835411290	12802536647
subscribers	23876871	24692177
uploads	12852.1	15085.28
age of channel	10.116	10.265
%unemployment rate	8.734	9.135
urban population rate	0.665	0.692
highest yearly earnings	10393617	10477249

Table 2. Median of attributes for observations in training group and testing group

Variables	Training group	Testing group
video views	8458730300	9338773418
subscribers	18450000	18750000
uploads	1177	1197.5
age of channel	10	10
%unemployment rate	5.36	8.43
urban population rate	0.825	0.825
highest yearly earnings	5400000	5400000

3.2. Exploratory data analysi

Exploratory data analysis was conducted to uncover the basic relationships between meaningful variables and the response variable, video views. Figure 1 demonstrates that YouTubers with higher video views tend to have higher annual earnings, more uploads, longer channel age, and a larger subscriber base compared to those with lower video views. Music channels also appear more attractive than other channel types. However, video views do not vary significantly with factors like the unemployment rate or urban population rate in the YouTuber’s country. In contrast, factors such as the YouTuber’s country (whether they are from an English-speaking country) and geographical position (latitude) seem to influence video views to some extent.

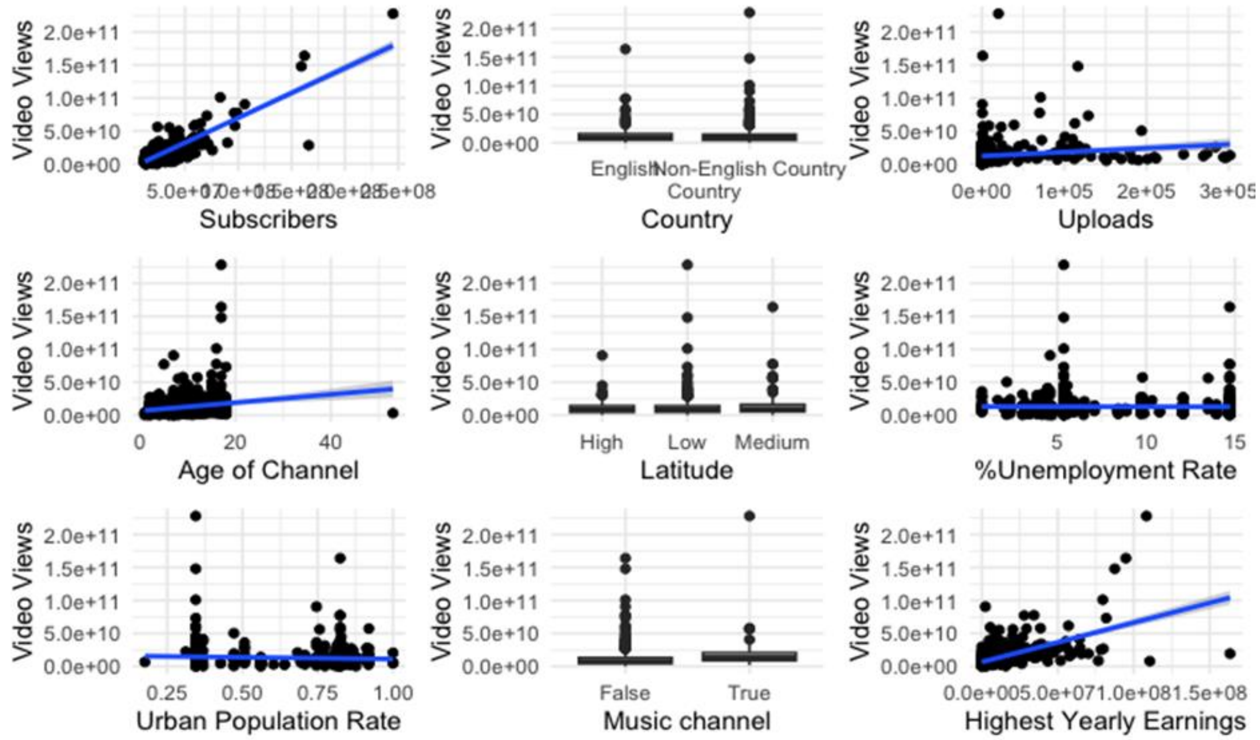

Fig. 1 Relationship between some meaningful variables and video views of YouTuber

3.3. Model Construction and Selection

This paper constructs the following models:

Model 1: $video\ views = \beta 0 + \beta 1 \times subscribers + \beta 2 \times uploads + \beta 3 \times Age\_of\_channel + \beta 4 \times Country + \beta 5 \times Latitude + \beta 6 \times Unemployment.rate + \beta 7 \times Urban\_population\_rate + \beta 8 \times Music\_channel + \beta 9 \times highest\_yearly\_earnings + \varepsilon.$

Model 2: $video\ views = \beta 0 + \beta 1 \times subscribers + \beta 2 \times uploads + \beta 3 \times Age\_of\_channel + \beta 4 \times Country + \beta 5 \times Latitude + \beta 6 \times Music\_channel + \beta 7 \times highest\_yearly\_earnings + \varepsilon.$

Model 3: $video\ views = \beta 0 + \beta 1 \times subscribers + \beta 2 \times uploads + \beta 3 \times highest\_yearly\_earnings + \varepsilon.$

Model 4: $video\ views = \beta 0 + \beta 1 \times subscribers + \beta 2 \times uploads + \beta 3 \times Age\_of\_channel + \beta 4 \times Country + \beta 5 \times Latitude + \beta 6 \times Unemployment.rate + \beta 7 \times Urban\_population\_rate + \beta 8 \times Music\_channel + \beta 9 \times (highest\_yearly\_earnings \times Age\_of\_channel) + \beta 10 \times (subscribers \times Age\_of\_channel) + \varepsilon.$

Table 3. Model comparison tool by AIC, R-squared and M

model	AIC	R-squared	MSE
model1	13492.15	0.8083178	9.951×10^7
model2	13489.23	0.807613	9.952×10^7
model3	13482.03	0.8057713	9.889×10^7
model4	13416.14	0.8539861	8.554×10^7

After constructing the four models and calculating the AIC, R-squared, and MSE, as shown in Table 3, Model 4 performs the best compared to the other three models. This is evident as it has the lowest MSE, indicating that it minimizes the estimated error. Model 4 also exhibits the highest R-squared value, signifying a good fit to the data and that a larger proportion of the variation in the response variable is explained by the predictor variables. Although AIC penalizes model complexity, Model 4 still has the smallest AIC value, demonstrating that the improvement in fit accuracy outweighs the cost of a slightly more complex model structure.

3.4. Model Construction and Selection

Then check the assumptions of the linear regression model constructed. As Figure 2 below, Residual plots show that the points are presented without any patterns. Thus linearity assumption, uncorrelated errors assumption, and constant variance assumption are all satisfied. The QQ-Residuals plot also shows an upward slope and it is nearly straight so the normality assumption of this model is also satisfied.

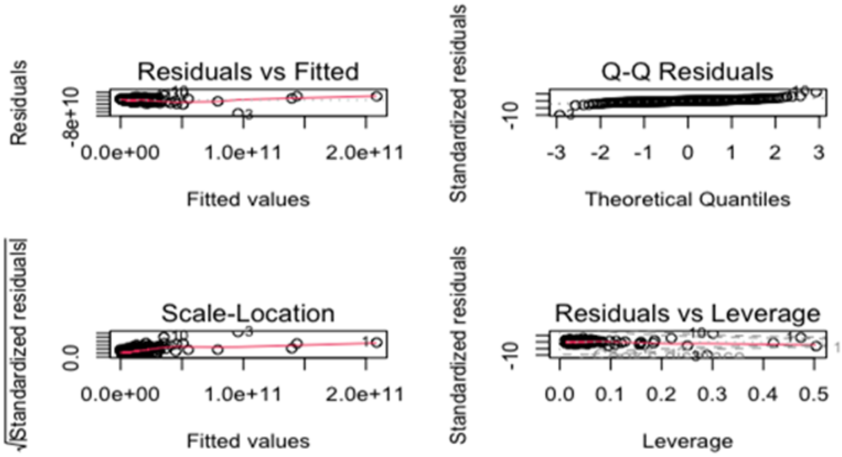

Fig. 2 Residual plots and QQ plots

According to Figure 3, the DFFITS plot and Cook’s Distance plot were examined to identify any influential points that could affect model performance. In the DFFITS plot, nearly all observations are close to the middle horizontal line, indicating that the influence of individual data points on the predicted values is generally low. Similarly, the Cook’s Distance plot shows that removing data points would have a minimal impact on the model fit. Therefore, no adjustments to the observations are necessary to improve the accuracy of the model fit.

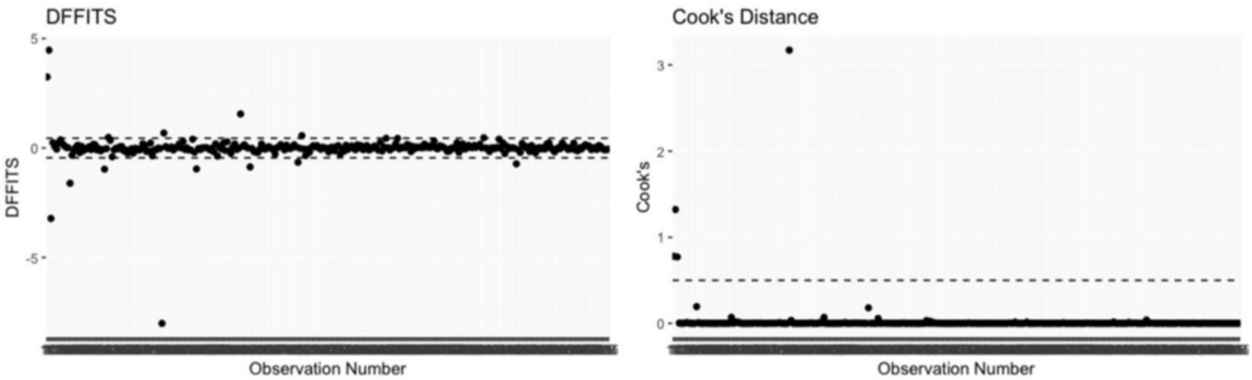

Fig. 3 DFFITS and Cook’s Distance plots

Finally, p-values for all coefficients were calculated, revealing that most are low, with four of them below 0.05. This indicates that the predictors in Model 4 effectively influence YouTube video views. Additionally, the F-statistic for Model 4 is as high as 129, with a corresponding p-value close to 0. This result suggests that the model explains a significant proportion of the variance in the response variable, providing strong evidence to reject the null hypothesis and confirm that at least one of the predictors is statistically significant.

4. Conclusion

As shown by the results, Model 4 is optimal compared to the other three models. It uses variables such as uploads, Age_of_Channel, Country, Latitude, Unemployment.rate, Urban_population_rate, Music_channel, the interaction between highest_yearly_earnings and Age_of_Channel, and the interaction between subscribers and Age_of_Channel to predict YouTuber video views. This allows for a comprehensive answer to the research question: "How do content-unrelated factors affect YouTubers' video views". The model not only identifies attributes influencing video views but also examines their effects.

However, some limitations exist in this optimal model. First, while interaction terms improve prediction accuracy, they make the interpretation of predictors less straightforward than a simple linear equation. Additionally, potential collinearity between subscribers and highest_yearly_earnings could negatively impact the model's fit. Although Model 4 performs best among the four models tested, the R-squared value is still below 1, indicating room for improvement with potentially better models. Moreover, Model 4's complexity increases the variance of predictors, although it reduces bias.

Another limitation lies in the dataset itself; some variables included in the model do not significantly impact video views, even though the most relevant predictors likely affecting video views were selected. Future studies could explore additional content-unrelated attributes that are independent and more significantly impactful on video views. This approach would be valuable for YouTube channel owners aiming to enhance their popularity and revenue.

References

[1] Hruska J, Maresova P. Use of social media platforms among adults in the United States—Behavior on social media. Societies (Basel, Switzerland), 2020, 10(1): 27.

[2] Han B. How do YouTubers make money? A lesson learned from the most subscribed YouTuber channels. International Journal of Business Information Systems, 2020, 33(1): 132-143.

[3] Sukhee Han. Anatomizing Popular YouTube Channels of English-speaking Countries. International Journal of Internet, Broadcasting and Communication, 2020, 12(4), 42-47.

[4] Yildirim G, Kocaelli H A. Assessment of the content and quality of YouTube videos related zygomatic implants: A content–quality analysis. Clinical Implant Dentistry and Related Research, 2023, 25(3), 605-612.

[5] Al-Mamouri H, Baiee W R. Maximizing video popularity prediction: A holistic approach utilizing metadata and thumbnail analysis. AIP Conference Proceedings, 2024, 3097(1).

[6] Welbourne D J, Grant W J. Science communication on YouTube: Factors that affect channel and video popularity. Public Understanding of Science (Bristol, England), 2016, 25(6): 706-718.

[7] Halim Z, Hussain S, Hashim Ali R. Identifying content unaware features influencing popularity of videos on YouTube: A study based on seven regions. Expert Systems with Applications, 2022, 206: 117836.

[8] Nisa M U N, et al. Optimizing prediction of youtube video popularity using xgboost. Electronics (Basel), 2021, 10(23): 2962.

[9] Ladhari R, Massa E, Skandrani H. YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise. Journal of Retailing and Consumer Services, 2020, 54: 102027.

[10] Baertl M. YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence (London, England), 2018, 24(1): 16-32.