Received: 15 April 2024 | Revised: 12 June 2024 | Accepted: 25 June 2024 | Published online: 5 July 2024

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2024, Vol. 00(00) 1–13
DOI: 10.47852/bonviewJCCE42023120

BON VIEW PUBLISHING

Selection of Optimal YouTube View Count Prediction Model Using Data Envelopment Analysis Approach

Garima Babbar[1], Adarsh Anand[1,*] and Mohini Agarwal[2]

[1]*Department of Operational Research, University of Delhi, India*

[2]*College of Humanities and Social Sciences, Grand Canyon University, United States*

Abstract: This study investigates the dynamics of video view counts on YouTube to gain a comprehensive understanding of its influence on digital media engagement. By analyzing the impact of increasing parameters in view count models and employing various distribution functions for the viewing rate, the research employs a unified approach to assess model responses to a broader set of parameters. This methodology integrates different distribution functions, resulting in a range of models that capture diverse viewing behavior patterns. The effectiveness of these models is evaluated using Data Envelopment Analysis (DEA), a robust analytical tool within the framework of Multiple Criteria Decision Making (MCDM) techniques. The results highlight that models with two parameters exhibit superior efficiency across multiple datasets, effectively representing the complex dynamics of YouTube's view count patterns. This study advances the academic discourse on digital platforms by offering a detailed analysis of YouTube's view count dynamics. The results provide actionable insights for customizing content and engagement strategies to match observed viewing behavior patterns, thereby improving content dissemination and audience interaction on the platform.

Keywords: Data Envelopment Analysis (DEA), Multiple Criteria Decision Making (MCDM), social media, view count, YouTube

1. Introduction

YouTube, a cornerstone of the digital landscape, has evolved from a mere video-sharing platform to a global influencer, shaping our daily interactions with digital content. Having amassed a user base of over 2.7 billion active users by 2024, YouTube has significantly influenced various sectors, fundamentally reshaping how people consume media [1]. From entertainment to education, it stands as a multifaceted hub with unparalleled reach and influence [2].

The platform's exponential growth has been fuelled by technological advancements and mobile device proliferation. Users not only consume but actively participate by generating and sharing content, creating a dynamic and interactive online community. This active engagement transforms YouTube beyond a passive consumption platform into a vibrant digital ecosystem where users contribute to the platform's content diversity, fostering collaborative interactions and shaping the collective digital narrative [3].

For content creators and businesses, YouTube's Partner Program is a vital revenue source [4]. An in-depth analysis of view count dynamics proves instrumental in formulating effective monetization strategies. By understanding the factors influencing view counts, creators can optimize ad revenue [5]. This strategic approach enhances the overall financial viability of content creation on YouTube, ensuring sustainable income streams for creators and fostering a thriving ecosystem within the platform's economic framework.

On the educational front, YouTube has become an invaluable resource, offering a vast repository of tutorials, lectures, and skill-building content [6]. The interactive nature and visual aspects of videos enhance the learning experience, rendering complex concepts more readily understandable [7]. Simultaneously, YouTube has evolved into a go-to destination for entertainment. Its expansive array of content spans from amateur videos to professionally produced shows, creating a diverse tapestry of entertainment options. The platform has become synonymous with streaming and sharing music, contributing significantly to the music industry's releases. Furthermore, YouTube's impact on global pop culture is undeniable, with viral sensations, memes, and influencers shaping societal trends [8].

Amid this transformative journey, the dynamics of video view counts on YouTube have emerged as a critical focal point [9]. Examining these dynamics provides creators, marketers, and researchers with a lens into the preferences of the vast YouTube audience, enabling tailored content strategies and informed decision-making [10, 11]. The fluctuations in view counts not only reflect content popularity but also hold the key to understanding the pulse of modern digital media engagement.

The dynamics of YouTube's viewership have attracted considerable scholarly attention over the past decade, with extensive research dedicated to understanding this phenomenon. Researchers have delved into understanding the intricate factors influencing video view counts on YouTube [12–16]. They have also explored

***Corresponding author:** Adarsh Anand, Department of Operational Research, University of Delhi, India. Email: aanand@or.du.ac.in

various facets of user behaviour, content popularity, and algorithmic influences on this prominent video-sharing platform.

Previous research underscores the significance of user engagement and its impact on content popularity on YouTube. Hu et al. [17] demonstrated the correlation between user interactions, such as likes and comments, and the overall popularity of videos. The diversity of user behaviour and viewing patterns on YouTube has been a subject of investigation. Sonoda et al. [18] explored the impact of user diversity on content popularity, shedding light on the varying preferences that influence video consumption.

The interplay between content creation and viewership dynamics is central to leveraging YouTube effectively across various domains - be it in marketing, education, entertainment, or information dissemination [19–21]. Hence, it becomes essential for stakeholders in these domains to analyze the factors impacting view counts. This enables them to customize their content and strategies, ensuring effective resonance with the intended audience.

Against this backdrop, our present study takes a focused approach to understanding how view count models respond to an increase in parameters. By exploring this, we aim to provide a clearer picture of what makes certain content more popular and how variations in parameters impact the overall viewership dynamics [22]. We employ a unified approach and incorporate various distribution functions in order to unveil a set of models designed to capture the diverse patterns of viewing behaviour on YouTube. The ability to understand and predict these fluctuations in view counts is paramount, as it directly influences the prevailing trends in how individuals interact with modern digital media. The study addresses the necessity of unravelling the complexities inherent in view count models to gain insights into the patterns of user engagement and content reception on this influential platform.

The current study uses the Data Envelope Analysis (DEA) method within the realm of Multiple Criteria Decision Making (MCDM) techniques. The approach has been widely applied across different industries, such as engineering, defence, hospitality sector, and the transport sector, to assess efficiencies, as demonstrated in studies [23–28].

The proposed research's core lies in utilizing the Data Envelopment Analysis (DEA) method to quantitatively measure and evaluate the performance efficiency of these models. Through this, our findings aim to contribute nuanced insights to the ongoing academic dialogue on digital platforms, providing a sophisticated analysis of YouTube's evolving landscape and the intricate dynamics involved in modeling viewing behaviour. This research, in essence, enhances our understanding of the multifaceted nature of YouTube's dynamics, thereby contributing significantly to the broader discourse on digital media platforms.

The remainder of the article is organized as follows: In Section 2, the authors delve into the foundational elements of the proposed modeling framework, followed by the presentation of the proposed modeling framework in Section 3. Section 4 elaborates on model illustration, while Section 5 covers the DEA framework. Sections 6 and 7 are dedicated to discussion and conclusion, respectively, with references provided subsequently.

2. Building Block of the Proposed Modelling Framework

2.1. Assumptions

The model works under the following assumptions:

1) The attainable maximum view-count for a YouTube video remains constant.
2) Each viewer is restricted to providing only one view.
3) The rate of viewership at any given time is directly proportional to the remaining pool of potential viewers.

2.2. Notations

$V(t)$	*"The cumulative no. of video views obtained by time 't'"*
N	*The total no. of potential viewers for a video"*
$F_i(t)$	*"CDF," $i = 1, 2, 3$*
$f_i(t)$	*"PDF," $i = 1, 2, 3$*
b	*Rate of viewing*
β	*Learning parameter*

3. Proposed Modelling Framework

Aggrawal et al. [29] laid the groundwork by developing a framework that delineated marketing science theory in relation to YouTube's video view counts. Expanding upon this foundation, Irshad et al. [30] introduced a model capturing the dynamics of YouTube video popularity. Their model integrated two pivotal factors: the dissemination of information among netizens, akin to word of mouth, and the number of subscribers to a specific video channel. Notably, Irshad et al. [30] proposed an alternative formulation to Aggrawal et al.'s [29] framework. This alternative considered a time-dependent rate of viewing, a critical element in determining a video's accumulated views, as expressed in Equation (1):

$$\frac{dV(t)}{dt} = b(t)[N - V(t)] \tag{1}$$

In the above equation, $V(t)$ represents the cumulative number of views up to time 't', N denotes the total number of potential viewers of a video, and $b(t)$signifies the time-dependent rate of viewing.

Utilizing the hazard rate approach, a widely acknowledged method in marketing science literature [31], Equation (1) can be expressed as:

$$\frac{dV(t)}{dt} = \frac{f(t)}{1 - F(t)}(N - V(t)) \tag{2}$$

"Where; $F(t)$ is the cumulative distribution function and $f(t)$ is the probability density function."

The above modelling framework can be solved to derive a closed form solution using the initial condition $V(t = 0) = 0$, resulting in the following equation:

$$V(t) = NF(t) \tag{3}$$

Adopting the standard approach outlined in Equation (3) above, the current research has varied the number of parameters for $F(t)$. This variation in parameters has led to the formulation of distinct models, each illustrating different viewing behaviour. Thus, within this modeling framework, three cases have been examined, distinguishing the models based on the number of parameters incorporated into the function $F(t)$.

3.1. Case I: Distribution characterized by a singular parameter

M_1: When the view count adheres to an exponential behavior with viewing rate b, it signifies a rapid and continuous growth

pattern, indicative of escalating popularity and widespread audience engagement. Exponential behavior in view counts suggests a compounding effect, where the rate of increase is proportional to the existing count, often reflecting viral content or sustained audience interest over time. This pattern is instrumental in identifying content that experiences accelerated visibility and resonates with a broad audience.

$$V(t) = N\left[1 - e^{-bt}\right] \tag{4}$$

3.2. Case II: Distribution characterized by two parameters

M_2: When the viewing pattern follows a logistic distribution, represented by rate parameter b and a learning parameter β, it implies a growth pattern that eventually reaches a saturation point. In a logistic distribution, the initial phase witnesses rapid growth, representing heightened interest. However, as the content saturates its potential audience, the rate of increase gradually diminishes. This distribution model is particularly relevant for understanding content that experiences both initial popularity surges and eventual stabilization as it reaches its audience limit.

$$V(t) = N\left[\frac{1 - e^{-bt}}{1 + \beta e^{-bt}}\right] \tag{5}$$

M_3: When the view-count conforms to the Weibull distribution, distinguished by rate parameter b and a shape parameter β, it suggests a versatile growth pattern with adjustable shapes. The Weibull distribution accommodates a range of behaviours, from initially slow to rapid accelerations or even decelerations, depending on the value of the shape parameter. This flexibility allows for the modeling of diverse scenarios, making it a valuable tool for capturing the nuanced dynamics of YouTube video view counts across various content types and audience interactions.

$$V(t) = N\left[1 - e^{-b(t)^{\beta}}\right] \tag{6}$$

M_4: When viewership aligns with a normal distribution, characterized by a mean μ and standard deviation σ, it reflects a typical or bell-shaped pattern. In this scenario, the majority of view counts cluster around the mean, with fewer instances of extreme values as determined by the standard deviation. This normal distribution model is instrumental in understanding the central tendency and variability of view counts, providing insights into the expected range of viewership for a given content type on YouTube.

$$V(t) = N.Normal(t, \mu, \sigma) \tag{7}$$

3.3. Case III: Distribution characterized by three parameters

M_5: When viewership pattern adheres to a Para-logistic distribution, denoted by shape parameter c, scale parameter α, and location parameter θ, it signifies a distribution with distinctive characteristics. This distribution allows for a nuanced representation of YouTube viewership dynamics, accommodating scenarios with unique patterns and trends.

$$V(t) = N\left\{1 - \left[1 + \left(\frac{t-\theta}{c}\right)^{\alpha}\right]^{-\alpha}\right\} \tag{8}$$

Using the above proposed models, we examine different setups of parameters for $F(t)$ to find the most effective one. The goal of this paper is to figure out the number of parameters for $F(t)$ that work best in representing the complexities of YouTube's view count patterns. By comparing various configurations, the study aims to identify the most suitable set of parameters, shedding light on what makes a particular setup optimal for modeling in the context of viewership on YouTube.

4. Numerical Illustration

The models presented in this study underwent validation using four distinct datasets. The dataset, described in Table 1, comprises view counts of YouTube videos collected at 4-hour intervals. This particular time frame was chosen to consolidate the dynamic and rapidly evolving landscape of video consumption on the platform. Encompassing videos from a wide array of genres and categories, the dataset ensures a diverse and representative sample, contributing to the robustness of the analysis.

4.1. Model parameters

This research utilized the nonlinear least square method to estimate the unknown parameters to assess the effectiveness of the proposed models. Tables 2–5 present the estimated parameter values obtained for each of the four datasets.

Table 1
Data description

Dataset	Category	Upload Date	Video Title
DS-I	Animation and Film	11-07-2023	Painkiller | Official Trailer | Netflix
DS-II	Kids	13-07-2023	Hindi Kids Rhymes
DS-III	Gaming	11-07-2023	Palia - Official Beta Release Trailer
DS-IV	News and Politics	13-07-2023	Delhi Flood Alert

Table 2
Model parameter values (estimated) for DS-I

	N	b	β	μ	σ	θ	c	α
M_1	13466.2	0.08	-	-	-	-	-	-
M_2	12953.77	0.13	1	-	-	-	-	-
M_3	13295.67	0.07	1.05	-	-	-	-	-
M_4	12674.73	-	-	9.13	9.2	-	-	-
M_5	15194.83	-	-	-	-	0.89	1.12	10.47

Table 3
Model parameter values (estimated) for DS-II

	N	b	β	μ	σ	θ	c	α
M_1	72058.8	0.1	-	-	-	-	-	-
M_2	70554.81	0.16	1	-	-	-	-	-
M_3	81201.28	0.2	0.63	-	-	-	-	-
M_4	72742.99	-	-	6.49	14.94	-	-	-
M_5	83779.75	-	-	-	-	0.12	7.54	1

Table 4
Model parameter values (estimated) for DS-III

	N	b	β	μ	σ	θ	c	α
M_1	13248.11	0.11	-	-	-	-	-	-
M_2	12888.35	0.18	1	-	-	-	-	-
M_3	15416.56	0.2	0.63	-	-	-	-	-
M_4	13517.52	-	-	6.09	13.94	-	-	-
M_5	15554.57		-	-	-	0.18	6.85	1

4.2. Model validation

The evaluation of the models has been carried out by using various comparison metrics. The outcomes of the goodness-of-fit assessment are displayed in Tables 6–9.

The criteria for evaluating goodness of fit provide valuable quantitative insights into how well the proposed models align with observed data. Nevertheless, it is important to highlight that among the six proposed models, no single model emerges as the unequivocal best fit.

Table 5
Model parameter values (estimated) for DS-IV

	N	b	β	μ	σ	θ	c	α
M_1	34366.97	0.1	-	-	-	-	-	-
M_2	34850.08	0.18	2.73	-	-	-	-	-
M_3	34879.59	0.03	1.42	-	-	-	-	-
M_4	33919.51	-	-	8.84	6.44	-	-	-
M_5	37078.37	-	-	-	-	0.75	11.6	1.562

Table 6
Values of comparison parameters for DS-I

	R2	Bias	Variance	MAE	RMSE	RMSPE
M_1	0.996	30.874	227.893	109.203	230.018	227.825
M_2	0.992	9.623	308.372	256.08	308.525	308.388
M_3	0.996	19.225	217.897	129.001	218.761	217.853
M_4	0.970	27.945	596.153	434.618	596.821	596.129
M_5	0.998	0.313	139.736	101.196	139.737	139.735

Table 7
Values of comparison parameters for DS-II

	R2	Bias	Variance	MAE	RMSE	RMSPE
M_1	0.950	408.97	3333.91	2768.95	3359.32	3333.97
M_2	0.903	619.3	4602.29	3879.99	4644.47	4602.36
M_3	0.993	-37.883	1214.04	718.353	1214.64	1214.03
M_4	0.949	-59.698	3367.43	2240.33	3367.97	3367.42
M_5	0.988	97.71	1610.57	1242.84	1613.58	1610.6

4.3. Graphical analysis

Figures 1–4 represent the accuracy of the proposed models with respect to the original data.

The figures presented above illustrate that the models show a satisfactory alignment with the original data, indicating a favorable correspondence between the proposed models and the obtained values.

5. Data Envelopment Analysis (DEA)

Data Envelopment Analysis (DEA) is a methodology designed to measure the relative efficiency of organizational units. Its primary

Table 8
Values of comparison parameters for DS-III

	R2	Bias	Variance	MAE	RMSE	RMSPE
M_1	0.908	56.005	894.077	681.814	895.866	894.108
M_2	0.861	88.963	1097.59	860.117	1101.26	1097.63
M_3	0.949	18.796	669.512	443.793	669.781	669.498
M_4	0.883	15.283	1009.77	654.463	1009.89	1009.76
M_5	0.951	0.001	657.25	464.925	657.25	657.25

Table 9
Values of comparison parameters for DS-IV

	R2	Bias	Variance	MAE	RMSE	RMSPE
M_1	0.954	223.32	2251.18	1810.4	2262.52	2251.23
M_2	0.992	103.67	967.016	659.607	972.707	966.962
M_3	0.994	-65.701	846.216	588.796	848.832	846.178
M_4	0.979	110.84	1529.42	1176.28	1533.54	1529.38
M_5	0.995	-1.189	742.929	569.379	742.93	742.928

advantage lies in its ability to comprehensively evaluate the dynamic interaction between different inputs and outputs, a task conventional ratio analysis often struggles to accomplish.

Data Envelopment Analysis (DEA), originally proposed by Charnes et al. [32] 1978 and subsequently expanded by Banker et al. in 1984, is a powerful technique categorized under Multi-Criteria Decision Making (MCDM) methods. CCR (Charnes, Cooper, and Rhodes) models, according to Charnes et al. [32] adhere to a constant returns to scale approach, whereas BCC (Banker, Charnes, and Cooper) models, following Banker et al. [33], employ a variable returns to scale approach.

DEA serves as a non-parametric method utilizing linear programming problems (LPP) to assess the performance efficiency of decision-making units (DMUs), where alternatives are termed DMUs. It excels in determining relative efficiencies among DMUs with similar considerations for various parameters. Noteworthy characteristics include its proficiency in handling DMUs with nearly identical parameters, with the approach tailored to minimize inputs and maximize outputs through the application of linear programming models. This methodology proves instrumental in evaluating and ranking DMUs based on their relative efficiency.

Hence, utilizing DEA facilitates acquiring performance insights for diverse decision-making units, along with the ability to propose specific input or output targets for those DMUs demonstrating lower performance.

Charnes et al. [32] introduced the concept that the efficiency of a decision-making unit (DMU) can be computed by evaluating the ratio of the weighted sum of outputs to the weighted sum of inputs. Mathematically, this computation is represented as follows:

$$Efficiency = \frac{Weighted\ sum\ of\ the\ output}{Weighted\ sum\ of\ the\ input} \tag{9}$$

Figure 1
Graphical representation for DS-I

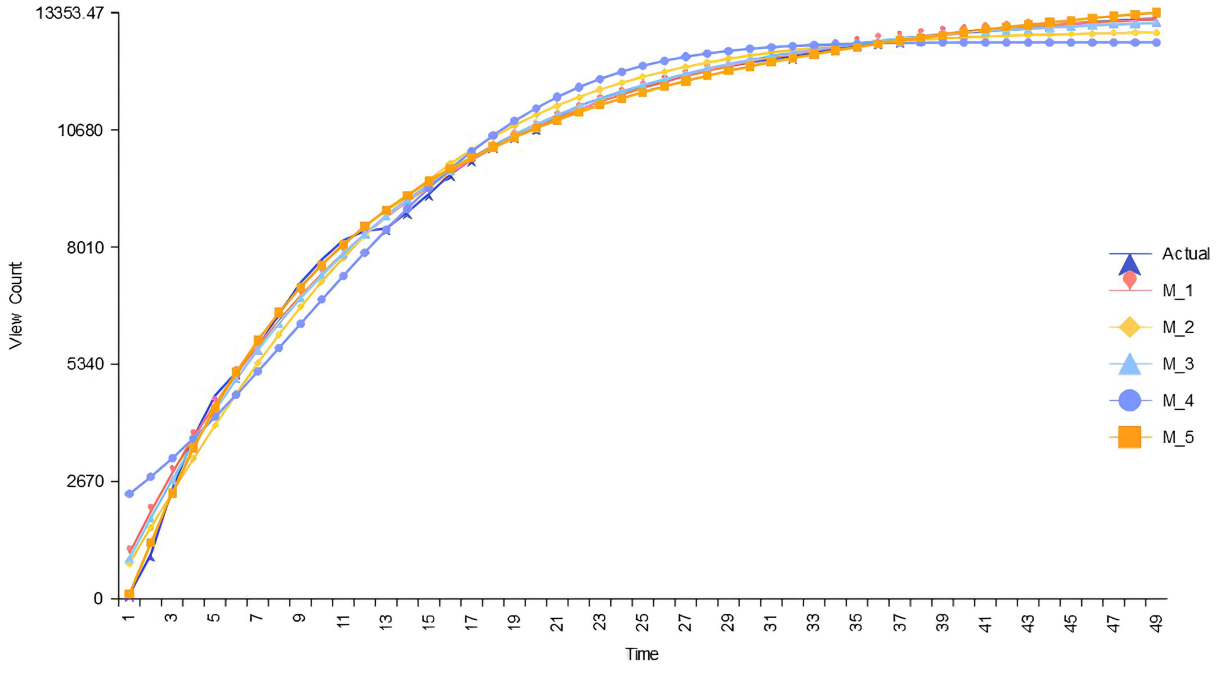


Figure 2
Graphical representation for DS-II

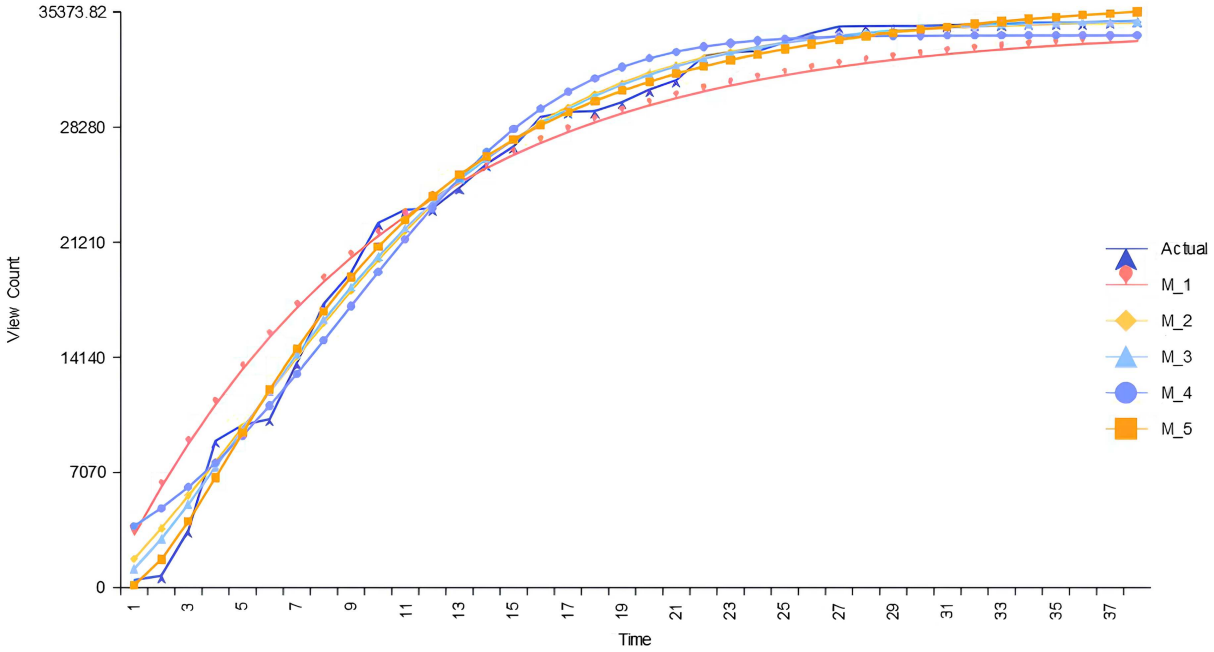
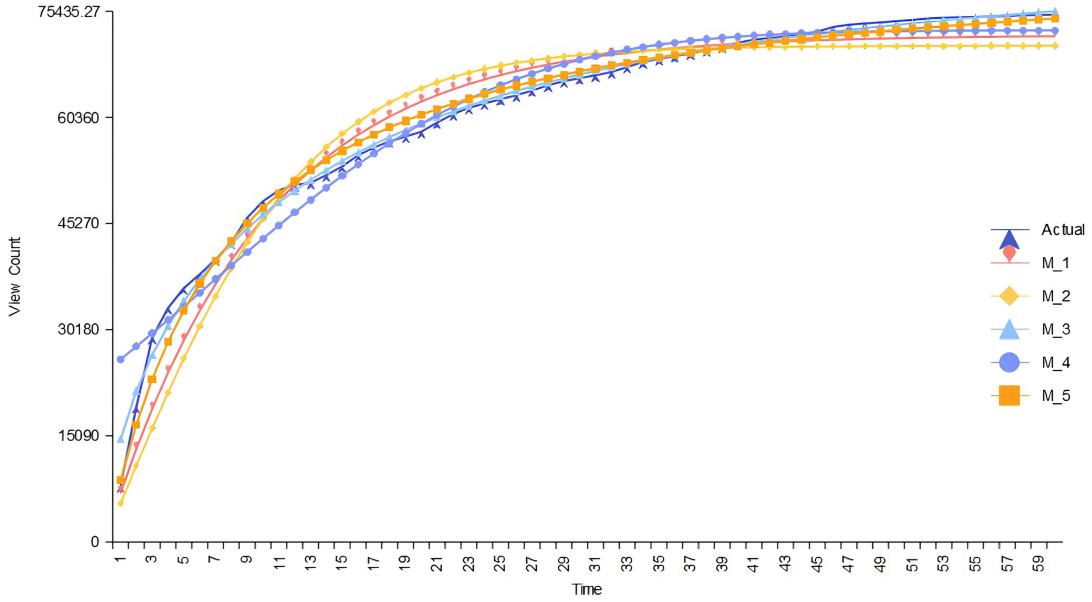


Figure 3
Graphical representation for DS-III



Equation (9) can be rewritten as:

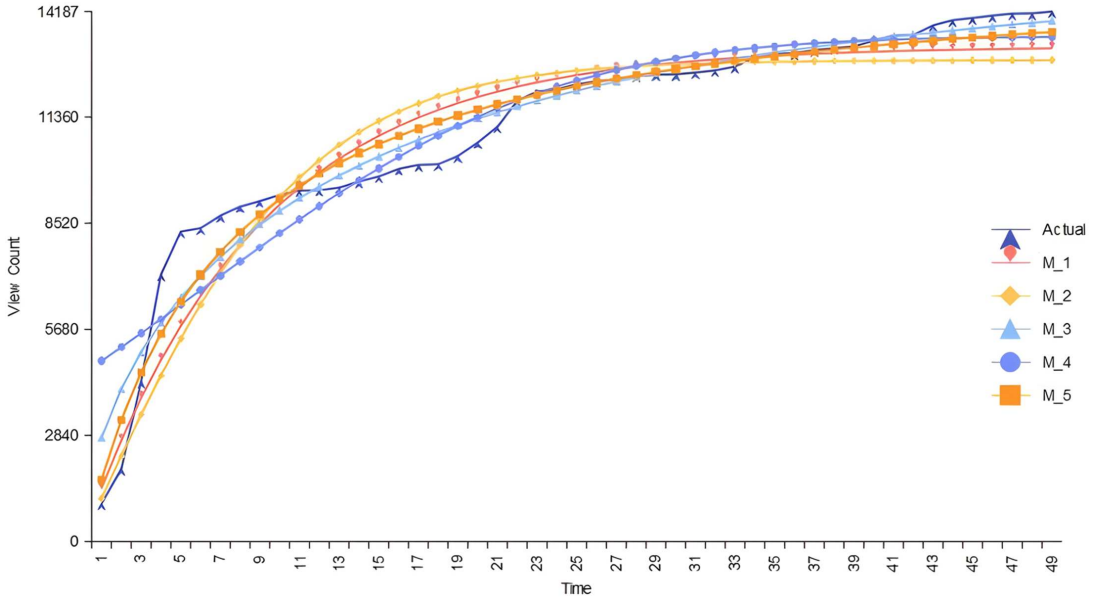
$$Max\, E_j = \frac{\sum_{r=1}^{s} w_r Output_{r,j}}{\sum_{i=1}^{m} z_i Input_{i,j}} \tag{10}$$

$$0 \leq \frac{\sum_{r=1}^{s} w_r Output_{r,j}}{\sum_{i=1}^{m} z_i Input_{i,j}} \leq 1;\ n = 1, 2, 3....., j, ...N \tag{11}$$
$$w_r, z_i \geq 0;\ \textit{for all } i,\ r$$

In the above equations, E_j denotes the efficiency of the j^{th} DMU, with $r = 1\ to\ s$, i $= 1\ to\ m\ and\ n = 1\ to\ N$. $Output_{r,j}$ represents the r^{th} output of the j^{th} DMU, where w_r signifies the weight of $Output_{r,j}$. Similarly, $Input_{i,j}$ stands for the i^{th} input of j^{th} DMU, and z_i denotes the weight of $Input_{i,j}$.

In 1978, Charnes pioneered a linear programming model known as CCR. This model, designed for Data Envelopment Analysis (DEA), offers flexibility by allowing computation through either maximizing output or minimizing input criteria. This versatility showcases the practical application of DEA. The above fractional form is thus linearised and solved using linear programming techniques.

Figure 4
Graphical representation for DS-IV



The CCR model in its **input-oriented** form can be precisely formulated as:

$$\begin{aligned} g_k &= \min\left(\sum_{i=1}^{m} v_i x_{ik}\right) \\ &\textit{subject to}: \\ -\sum_{r=1}^{s} u_r y_{rk} + \sum_{i=1}^{m} v_i x_{ik} &\geq 0 \; \textit{for}\; k = 1, ..., n \\ \sum_{r=1}^{s} u_r y_{rk} &= 1 \\ u_r &\geq 0,\; r = 1, ..., s \\ v_i &\geq 0,\; i = 1, ..., m \end{aligned} \tag{12}$$

The CCR model in its **output-oriented** form can be formulated as:

$$\begin{aligned} g_k &= \max\left(\sum_{r=1}^{s} u_r y_{rk}\right) \\ &\textit{subject to}: \\ \sum_{r=1}^{s} u_r y_{rk} - \sum_{i=1}^{m} v_i x_{ik} &\geq 0 \; \textit{for}\; \mathrm{j} = 1, ..., n \\ \sum_{r=1}^{s} v_i x_{ik} &= 1 \\ u_r &\geq 0,\; r = 1, ..., s \\ v_i &\geq 0,\; i = 1, ..., m \end{aligned} \tag{13}$$

where n is the number of alternatives/DMUs, m is the number of input criteria, s is the number of output criteria, x_{ik} *and* y_{rk} are the values of i^{th} input criteria and r^{th} output criteria for k^{th} alternative, u_r *and* v_i are the non-negative variable weights that can be evaluated by the optimal solution of minimisation problem.

Assume that there exists n^{th} DMUs and the j^{th} DMU produces $s$outputs $(y_{ij}, ..., y_{sj})$ by using $m$inputs $(x_{1j}, ..., x_{mj})$. λ_j denotes the weight of the inputs/outputs of the j^{th} DMU. The efficiency score of the observed DMU is given as the optimal value to the following linear programming problem:

$$\begin{aligned} &Max\theta_0^* = \theta; \\ &\textit{Subjected to}: \\ &\sum_j \lambda_j x_{ij} \leq x_{i0},\; i = 1, 2, ..., m \\ &\sum_j \lambda_j y_{rj} \geq \theta y_{r0},\; \mathrm{r} = 1, 2, ..., s \\ &\lambda_j \geq 0,\; j = 1, 2, ..., n \end{aligned} \tag{14}$$

The aforementioned model represents the general output-oriented CCR model, adhering to the constant return to scale principle. This model aims to maximize output while maintaining inputs constant, thereby improving the relative efficiency of DMUs.

$$\begin{aligned} &Min\,\theta_0^* = \theta; \\ &\textit{Subjected to}: \\ &\sum_j \lambda_j x_{ij} \leq \theta x_{i0},\; i = 1, 2, ..., m \\ &\sum_j \lambda_j y_{rj} \geq y_{r0},\; \mathrm{r} = 1, 2, ..., s \\ &\lambda_j \geq 0,\; j = 1, 2, ..., n \end{aligned} \tag{15}$$

Likewise, the aforementioned model corresponds to the general input-oriented DEA model, also adhering to the constant return to scale principle. This model seeks to minimize inputs while holding outputs constant, thereby enhancing the relative efficiency of DMUs.

A DMU is said to be efficient if $\theta_0^* = 1$, otherwise it is said to be CCR inefficient.

In our context, we designate non-beneficial criteria or parameters as inputs, while beneficial criteria are categorized as outputs. Tables 10–13 represent the decision matrix for DS-1-IV respectively.

The efficiency of the models will be now evaluated using the aforesaid approaches:

1) Based on CCR general input and output-oriented

To understand the current modeling framework under both general input and general output-oriented CCR Model, presented below are a series of equations pertaining to M_1 for DS-I. Similarly, corresponding sets of equations for other datasets and alternatives can be formulated.

Table 10
Decision matrix for DS-I

Criteria →	N	b	β	μ	σ	θ	c	α	Output
Alternative ↓	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Beneficial
M_1	13466.201	0.08	-	-	-	-	-	-	13195.94
M_2	12953.774	0.126	1	-	-	-	-	-	12899.72
M_3	13295.668	0.072	1.051	-	-	-	-	-	13114.89
M_4	12674.728	-	-	9.133	9.19	-	-	-	12674.64
M_5	15194.826	-	-	-	-	0.886	1.123	10.466	13353.47

Table 11
Decision matrix for DS-II

Criteria →	N	b	β	μ	σ	θ	c	α	Output
Alternative ↓	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Beneficial
M_1	72058.799	0.102	-	-	-	-	-	-	71896.17
M_2	70554.811	0.155	1	-	-	-	-	-	70542.14
M_3	81201.28	0.198	0.633	-	-	-	-	-	75435.27
M_4	72742.993	-	-	6.486	14.938	-	-	-	72730.6
M_5	83779.745	-	-	-	-	0.12	7.541	1	74409.43

Table 12
Decision matrix for DS-III

Criteria →	N	b	β	μ	σ	θ	c	α	Output
Alternative ↓	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Beneficial
M_1	13248.114	0.114	-	-	-	-	-	-	13198.95
M_2	12888.354	0.18	1	-	-	-	-	-	12884.48
M_3	15416.557	0.199	0.633	-	-	-	-	-	13928.74
M_4	13517.523	-	-	6.086	13.935	-	-	-	13503.51
M_5	15554.571		-	-	-	0.184	6.853	1	13639.68

Table 13
Decision matrix for DS-IV

Criteria →	Output	α	c	θ	σ	μ	β	b	N
Alternative ↓	Beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial	Non-beneficial
M_1	33567.21	-	-	-	-	-	-	0.099	34366.97
M_2	34715.92	-	-	-	-	-	2.732	0.181	34850.084
M_3	34773.51	-	-	-	-	-	1.417	0.033	34879.59
M_4	33919.41	-	-	-	6.436	8.844	-	-	33919.51
M_5	35373.82	1.562	11.598	0.748	-	-	-	-	37078.369

2) For the first alternative M_1 (input-oriented) for DS-I

$$
\begin{aligned}
&Min\theta_0^* = \theta;\\
&13466.2 * \lambda 1 + 12953.77 * \lambda 2 + 13295.7 * \lambda 3 + 12674.7 * \lambda 4\\
&\quad +15194.82 * \lambda 5 <= 13466.2 * \theta;\\
&0.08 * \lambda 1 + 0.126 * \lambda 2 + 0.072 * \lambda 3 <= 0.08 * \theta;\\
&1 * \lambda 2 + 1.051 * \lambda 3 <= 0;\\
&9.133 * \lambda 4 <= 0;\\
&9.19 * \lambda 4 <= 0;\\
&0.886 * \lambda 5 <= 0;\\
&1.123 * \lambda 5 <= 0;\\
&10.466 * \lambda 5 <= 0;\\
&13195.9 * \lambda 1 + 12899.7 * \lambda 2 + 13114.9 * \lambda 3 + 12674.6 * \lambda 4\\
&\quad +13353.5 * \lambda 5 >= 13195.9;\\
&\lambda_j >= 0,\ j = 1, ...5
\end{aligned}
\tag{16}
$$

3) For the first alternative M_1 (output-oriented) for DS-I

$$
\begin{aligned}
&Max\theta_0^* = \theta;\\
&13466.2 * \lambda 1 + 12953.77 * \lambda 2 + 13295.7 * \lambda 3 + 12674.7 * \lambda 4\\
&\quad +15194.82 * \lambda 5 <= 13466.2;\\
&0.08 * \lambda 1 + 0.126 * \lambda 2 + 0.072 * \lambda 3 <= 0.08;\\
&1 * \lambda 2 + 1.051 * \lambda 3 <= 0;\\
&9.133 * \lambda 4 <= 0;\\
&9.19 * \lambda 4 <= 0;\\
&0.886 * \lambda 5 <= 0;\\
&1.123 * \lambda 5 <= 0;\\
&10.466 * \lambda 5 <= 0;\\
&13195.9 * \lambda 1 + 12899.7 * \lambda 2 + 13114.9 * \lambda 3 + 12674.6 * \lambda 4\\
&\quad +13353.5 * \lambda 5 >= 13195.9 * \theta;\\
&\lambda_j >= 0,\ j = 1, ...5
\end{aligned}
\tag{17}
$$

Similarly, the model can be written for other alternatives.

The linear programming problem was resolved utilizing Lingo software. Tables 14–17 display the scores related to both the input and output-oriented approaches for the CCR model. The results indicate multiple efficient models for each dataset.

Table 14
Solution of DMUs for CCR for DS-I

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*1*	*1*
M_2	*1*	*1.000002*
M_3	*1*	*1*
M_4	*1*	*1*
M_5	*1*	*1*

Table 15
Solution of DMUs for CCR for DS-II

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*1*	*1*
M_2	*1*	*1*
M_3	*0.9300275*	*1.075237*
M_4	*1*	*1*
M_5	*1*	*1*

Table 16
Solution of DMUs for CCR for DS-III

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*1*	*1*
M_2	*1*	*1*
M_3	*0.9*	*1.104707*
M_4	*1*	*1*
M_5	*1*	*1*

Table 17
Solution of DMUs for CCR for DS-IV

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*1*	*1*
M_2	*0.9991891*	*1.000812*
M_3	*1*	*1*
M_4	*1*	*1*
M_5	*1*	*1*

5.1. Super efficiency

Super efficiency also known as the super efficiency model or super efficiency DEA, is an extension of DEA that allows for the identification of the most efficient DMUs within the group of already efficient DMUs. In Data Envelopment Analysis (DEA), efficient Decision-Making Units (DMUs) cannot be ranked solely based on their efficiency scores, as they all attain the maximum score of unity. Nonetheless, it is acknowledged that efficient DMUs may exhibit variations in their actual performance in practical scenarios. To overcome this limitation, several approaches have been devised to enable a comprehensive ranking of efficient DMUs. Andersen and Petersen [34] proposed super-efficiency as a ranking methodology, aiming to distinguish the performance levels of exceptionally efficient DMUs.

The super efficiency model works by imposing an additional constraint on the already efficient DMUs, requiring them to maintain their current level of efficiency while minimising their resource utilisation. This constraint aims to identify the most efficient DMU that achieves the highest level of efficiency while using the least amount of resources.

By applying the super-efficiency model, it becomes possible to identify the best-performing DMUs among the group of efficient units. These DMUs serve as benchmarks for other efficient units, showcasing best practices and providing insights into further performance improvement.

Tables 18–21 show the super-efficiency scores for all datasets.

The analysis of super efficiency results reveals that M_3 emerges as the most efficient model for DS-I and DS-II, while M_2 demonstrates superior efficiency for DS-III and DS-IV.

6. Discussion

The proposed models with different number of parameters were evaluated and compared using Tables 2–5. Tables 6–9 showcase the performance of the proposed models across the four considered datasets. For better evaluation of the performance of the proposed models, we employed Data Envelopment Analysis (DEA) to assess their relative efficiency and

Table 18
Super-efficiency scores for DS-I

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*NFS*	*0*
M_2	*1.009707*	*0.9903863*
M_3	*1.104291*	*0.9055586*
M_4	*NFS*	*0*
M_5	*NFS*	*0*

Table 19
Super-efficiency scores for DS-II

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*NFS*	*0*
M_2	*1.002082*	*0.9979223*
M_3	*-*	*1.074007*
M_4	*NFS*	*0*
M_5	*NFS*	*0*

Table 20
Super-efficiency scores for DS-III

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*NFS*	*0*
M_2	*1.003*	*0.9965885*
M_3	*-*	*1.104707*
M_4	*NFS*	*0*
M_5	*NFS*	*0*

Table 21
Super-efficiency scores for DS-IV

Alternatives	*Input-Oriented*	*Output-Oriented*
M_1	*NFS*	*0*
M_2	*-*	*1.000812*
M_3	*3.107811*	*0.3217699*
M_4	*NFS*	*0*
M_5	*NFS*	*0*

to quantify their performance to identify the optimal models. The results showcase the efficiency variations among models, with M_3 excelling in DS-I and DS-II, while M_2 outperforming in DS-III and DS-IV. This dataset-specific efficiency highlights the need for a tailored approach in model selection. The diverse characteristics of datasets contribute to distinct model performances, emphasizing the importance of context-specific considerations.

6.1. Research contributions

This study makes significant contributions to the understanding of YouTube's view count dynamics. By employing a unified approach and incorporating diverse distribution functions, it introduces a set of parametric models capturing varied viewing behaviours. The research extends the literature by evaluating these models using the Data Envelopment Analysis method, providing a quantitative measure of their performance efficiency. This not only advances knowledge in predictive modeling on digital platforms but also enhances comprehension of the nuanced relationships between parameterization and model efficacy.

6.2. Implications for practice

Practically, this research offers valuable insights for content creators, platform developers, and marketers. Understanding the factors influencing video popularity on YouTube is crucial for optimizing content strategies. The introduced models provide a practical framework for predicting view counts, aiding creators in content planning. For platform developers, insights into user behaviour can inform algorithmic enhancements, improving content recommendations. Marketers can leverage these findings to tailor promotional strategies and maximize audience engagement.

7. Conclusion

In conclusion, this study delves into YouTube's view count dynamics, presenting a unified approach to model evaluation. The diverse set of parametric models contributes to a nuanced understanding of viewing behaviour. The application of Data Envelopment Analysis ensures a robust assessment of model performance. Through this research, we gain valuable insights into the multifaceted nature of YouTube's dynamics, bridging gaps in predictive modeling on digital platforms.

7.1. Limitations

Despite its contributions, this study is not without limitations. The focus on YouTube may limit generalizability to other platforms. The models, while diverse, might not capture all potential viewing behaviours. Additionally, the dynamic nature of online platforms presents challenges in modeling evolving user preferences.

7.2. Future research directions

Future research should extend the investigation to encompass a broader range of digital platforms, considering variations in user behaviour and content dynamics. Refinement of the proposed models based on user feedback and real-time adjustments could enhance their predictive accuracy. Exploring the influence of external factors, such as trending topics or global events, on view counts could provide a more comprehensive understanding.

Funding Support

The work done in this paper has been supported by grants sanctioned to the first author from UGC NET via Ref. No. 220510230085 and grants received by the second author from Institute of Eminence, DU, India as a part of Faculty Research Program via Ref no./IOE/2021/12/FRP.

Ethical Statement

This study does not involve any research with human or animal subjects conducted by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The DS-I data that support the findings of this study are openly available at https://youtu.be/24-YonhNS0Y. The DS-II data that support the findings of this study are openly available at https://youtu.be/S7eJes8AirA/hetrec-2011/. The DS-III data that support the findings of this study are openly available at https://youtu.be/MmlJb0Pi2-0. The DS-IV data that support the findings of this study are openly available at https://youtu.be/97AE_mAlhhc.

Author Contribution Statement

Garima Babbar: Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Funding acquisition. **Adarsh Anand:** Conceptualization, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Mohini Agarwal:** Methodology, Software.

References

[1] Singh, S. (2024). *How many people use YouTube (2024 statistics)*. Retrieved from: https://www.demandsage.com/youtube-stats/

[2] Dehghani, M., Niaki, M. K., Ramezani, I., & Sali, R. (2016). Evaluating the influence of YouTube advertising for attraction of young customers. *Computers in Human Behavior, 59*, 165–172. https://doi.org/10.1016/j.chb.2016.01.037

[3] Wegener, C., Prommer, E., & Linke, C. (2020). Gender representations on YouTube: The exclusion of female diversity. *M/C Journal, 23*(6). https://doi.org/10.5204/mcj.2728

[4] Maddox, J. (2021). What do creators and viewers owe each other? Microcelebrity, reciprocity, and transactional tingles in the ASMR YouTube community. *First Monday, 26*(1), 1–16. https://doi.org/10.5210/fm.v26i1.10804

[5] Elango, D. (2019). social media video creators monetization and business on YouTube. *SSRN*. http://dx.doi.org/10.2139/ssrn.3320253

[6] Xia, D., Zhang, Y., Qiu, Y., Zhang, S., Tian, Y., & Zhao, X. (2024). Research on the dynamic evolution law of online knowledge sharing under trust. *International Journal of Changes in Education, 1*(1), 32–40. https://doi.org/10.47852/bonviewIJCE32021834

[7] Burgess, J., & Green, J. (2018). *YouTube: Online video and participatory culture*. UK: Polity Press.

[8] Munnukka, J., Maity, D., Reinikainen, H., & Luoma-aho, V. (2019). "Thanks for watching." The effectiveness of YouTube vlogendorsements. *Computers in Human Behavior, 93*, 226–234. https://doi.org/10.1016/j.chb.2018.12.014

[9] Lin, N., Li, D., Ding, Y., He, B., Qin, Z., Tang, J., . . . , & Dong, T. (2012). The dynamic features of Delicious, Flickr, and YouTube. *Journal of the American Society for Information Science and Technology, 63*(1), 139–162. https://doi.org/10.1002/asi.21628

[10] Weinmayer, K., Garaus, M., & Wagner, U. (2024). The impact of corporate sustainability performance on advertising efficiency. *OR Spectrum, 46*, 175–209. https://doi.org/10.1007/s00291-023-00717-z

[11] Guhl, D., Paetz, F., Wagner, U., & Wedel, M. (2024). Predicting and optimizing marketing performance in dynamic markets. *OR Spectrum, 46*, 1–27. https://doi.org/10.1007/s00291-024-00755-1

[12] Kamiyama, N., & Murata, M. (2019). Reproducing popularity distribution of YouTube videos. *IEEE Transactions on Network and Service Management, 16*(3), 1100–1112. https://doi.org/10.1109/TNSM.2019.2914222

[13] Rui, L. T., Afif, Z. A., Saedudin, R. R., Mustapha, A., & Razali, N. (2019). A regression approach for prediction of Youtube views. *Bulletin of Electrical Engineering and Informatics, 8*(4), 1502–1506. https://doi.org/10.11591/eei.v8i4.1630

[14] Wongsuparatkul, E., & Sinthupinyo, S. (2020). View count of online videos prediction using clustering view count patterns with multivariate linear model. In *Proceedings of the 8th International Conference on Computer and Communications Management,* 123–129. https://doi.org/10.1145/3411174.3411186

[15] Park, J. (2022). *The impact of YouTube's thumbnail images and view counts on users' selection of video clip, memory recall, and sharing intentions of thumbnail images*. Master's Thesis, The Florida State University.

[16] Manikandan, P., Manimuthu, A., Rajam, J. S., & Sharma , K. S. N. (2022). Prediction of YouTube view count using supervised and ensemble machine learning techniques. In *International Conference on Automation, Computing and Renewable Systems,*1038–1042. https://doi.org/10.1109/ICACRS55517.2022.10029277

[17] Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram: A first analysis of Instagram photo content and user types. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 595–598. https://doi.org/10.1609/icwsm.v8i1.14578

[18] Sonoda, A., Seki, Y., & Toriumi, F. (2022). Analyzing user engagement in news application considering popularity diversity and content diversity. *Journal of Computational Social Science, 5*(2), 1595–1614. https://doi.org/10.1007/s42001-022-00179-3

[19] Febriyantoro, M. T. (2020). Exploring YouTube marketing communication: Brand awareness, brand image and purchase intention in the millennial generation. *Cogent Business & Management, 7*(1), 1787733. https://doi.org/10.1080/23311975.2020.1787733

[20] Rahmatika, R., Yusuf, M., & Agung, L. (2021). The effectiveness of YouTube as an online learning media. *Journal of Education Technology, 5*(1), 152–158.

[21] Sharma, T., & Sharma, S. (2021). A study of YouTube as an effective educational tool. *The Journal of Contemporary Issues in Business and Government, 27*(1), 2686–2690.

[22] Kumar, A., Singh, O., Anand, A., & Kapur, P. K. (2024). Studying innovation adoption using different distribution functions. *International Journal of System Assurance Engineering and Management, 15*, 1900–1907. https://doi.org/10.1007/s13198-023-02204-8

[23] Stojčić, M., Zavadskas, E. K., Pamučar, D., Stević, Ž., & Mardani, A. (2019). Application of MCDM methods in sustainability engineering: A literature review 2008–2018. *Symmetry, 11*(3), 350. https://doi.org/10.3390/sym11030350

[24] Nguyen, N. T., & Nguyen, L. X. T. (2019). Applying DEA model to measure the efficiency of hospitality sector: The case of Vietnam. *International Journal of Analysis and Applications, 17*(6), 994–1018.

[25] Solana Ibáñez, J., Caravaca Garratón, M., & Soto Meca, A. (2020). A literature review of DEA efficiency methodology in defence sector. *Academia Revista Latinoamericana*

de Administración, 33(3/4), 381–403. https://doi.org/10.1108/ARLA-11-2019-0228

[26] Song, M., Jia, G., & Zhang, P. (2020). An evaluation of air transport sector operational efficiency in China based on a three-stage DEA analysis. *Sustainability, 12*(10), 4220. https://doi.org/10.3390/su12104220

[27] Khan, D., & Gupta, R. K. (2024). Production optimization with the maintenance of environmental sustainability based on multi-criteria decision analysis. *Environment, Development and Sustainability, 26*, 19425–19442. https://doi.org/10.1007/s10668-023-03316-8

[28] He, J., Lau, W. T., & Liu, Y. (2024). Innovative production efficiency in Chinese high-tech industries during the 13th Five-Year Plan considering environmental factors: Evidence from a three-stage DEA model. *Green and Low-Carbon Economy, 2*(1), 37–48. https://doi.org/10.47852/bonviewGLCE3202910

[29] Aggrawal, N., Arora, A., & Anand, A. (2018). Modeling and characterizing viewers of You Tube videos. *International Journal of System Assurance Engineering and Management, 9*, 539–546. https://doi.org/10.1007/s13198-018-0700-6

[30] Irshad, M. S., Anand, A., & Bisht, M. (2019). Modelling popularity dynamics based on YouTube viewers and subscribers. *International Journal of Mathematical, Engineering and Management Sciences, 4*(6), 1508–1521. https://dx.doi.org/10.33889/IJMEMS.2019.4.6-119

[31] Anand, A., Agarwal, M., Aggrawal, D., & Singh, O. (2016). Unified approach for modeling innovation adoption and optimal model selection for the diffusion process. *Journal of Advances in Management Research, 13*(2), 154–178. https://doi.org/10.1108/JAMR-03-2015-0021

[32] Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

[33] Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078

[34] Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science, 39*(10), 1261–1264. https://doi.org/10.1287/mnsc.39.10.1261

How to Cite: Babbar, G., Anand, A., & Agarwal, M. (2024). Selection of Optimal YouTube View Count Prediction Model Using Data Envelopment Analysis Approach. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE42023120