DOI: 10.23977/autml.2024.050205
ISSN 2516-5003 Vol. 5 Num. 2

Machine Learning Based Short Video Comment Count Prediction

Yimeng Ren[1], Bo Xiao[2]

[1]School of Internet of Things Engineering, Jiangnan University, Wuxi, 214122, China
[2]School of Business, Ningbo University, Ningbo, 315211, China

Keywords: Short videos, video comment count prediction, machine learning algorithms

Abstract: With the rapid development of today's society, the widespread use of the Internet and the rapid development of network information technology, people are taking the initiative to accept and use the Internet. Compared with general network goods, short videos have a higher degree of dissemination as well as better acceptance. In the development of the Internet industry, short video content has been enriched, driving the growth of user scale and viscosity, and becoming a major source of incremental mobile Internet hours and traffic. At the same time, viewers also prefer to post their personal feelings to the response comment section when watching, thus interacting and communicating with other people watching, thus generating a large amount of comment information. These comments intuitively express the user's likes and dislikes and demands, and the number of comments also reflects the popularity of the video to a certain extent. Predicting the number of comments on a short video allows short video operation platforms to understand and master the popularity of the video, increase traffic investment in potential short videos in a targeted manner, and help potential short videos to gain higher popularity and create a more appealing short video ecological environment. In this context, this design uses data analysis tools to analyse and predict the number of short video comments. In this paper, three models, XGBoost, Random Forest model and LASSO, are used to analyse and predict the number of short video comments, and by comparing the MSE, interpreting the SHAP graph, the model with the strongest prediction ability is selected, so as to predict the hotness of short video.

1. Introduction

1.1 Background and Significance

In recent years, the Internet has become more and more popular and has become an important part of people's lives. According to the 51st Statistical Report on Internet Development in China published by China's Internet Information Centre (CNNIC), as of December 2022, the number of Internet users in China had reached 1.067 billion, an increase of 35.49 million compared with December 2021, and the Internet penetration rate had reached 75.6%.

As a new public opinion leader, short video has been developing rapidly by virtue of its short duration, concentrated content, entertainment and strong expressive power. After years of rapid growth and continuous improvement, short video has entered a new stage of mature development. As an emerging audiovisual industry, short video has further strengthened and highlighted its cultural and economic attributes, becoming a new medium for public opinion propaganda, information dissemination, cultural construction and economic and social development. Nowadays, a large number of high-quality videos have emerged, the content form is increasingly diversified, the scale of the industry is expanding, and it has become the main engine of the network audiovisual industry. In the future, short video platforms will further seek new breakthroughs, such as joining live broadcasting, e-commerce and other businesses, while the current head of short video platforms have been developing online live broadcasting business, and seek to deepen the relationship with other content creators, while developing new features to deepen the interactivity between creators and users. On the technical level, the increase in 5G penetration rate and the development of artificial intelligence and big data technology will provide new support for short video platforms. Coupled with the state to strengthen the regulation of the industry, platforms on the short video content released by users to strengthen the audit, the short video industry will become more and more standardised. From a comprehensive point of view, the short video industry has great potential for development.

As a mainstream Internet application, the market competition pattern of the short video industry is relatively stable, with platform factions blossoming and constantly changing and innovating to enhance their competitiveness. Among them, Jitterbug and Shutterbug are in the top tier of the industry, and a competitive landscape consisting of Jitterbug, Shutterbug and their related product matrices has been formed. As a popular short-video social platform, Jitterbug has hundreds of millions of active users and a huge amount of video content.

On this creative and competitive platform, users interact with content by watching, liking and commenting, thus generating a large amount of user behaviour data. These data not only reflect users' interests and preferences, but also provide rich content resources and social interactions for the Jieyin platform.

Therefore, it is necessary to use data analysis tools to assess the performance and effectiveness of short videos, and effectively guide the improvement of video creation quality.

Web commenting is also known as user-generated content, or UGC for short. It has attracted a lot of attention since its inception, and is recognised for its objectivity, openness and a host of other characteristics. This, coupled with the fact that the audience of short videos is very wide and random, makes video reviews very suitable for studying the spread of videos.

1.2 Research Status

At present, in the field of short video comment count analysis, data analysis tools are mainly used in the following two aspects: one is to choose the appropriate mathematical model to analyse the number of short video comments, which is used to assess the changes in the number of video comments, and the other is to find the reasons for the changes in the number of comments on short videos, which can be used by short video operation platforms to select short videos with more social influence and increase the exposure and recommendation, and short video creators can also be directed to improve the quality of short videos to increase the number of comments on individual videos as much as possible. Short video creators can also improve the quality of short videos in a directional way to increase the number of comments on individual videos as much as possible.

This paper will be based on the above two aspects to analyse and predict the number of short video comments. It is necessary and important to use data analysis tools to evaluate, this design will use three models, eXtreme Gradient Boosting(XGBoost), Random Forest model(RF) and The Least Absolute Shrinkage and Selection Operator(Lasoo), to evaluate the performance and effectiveness of short videos, to analyse and predict the number of comments on short videos, by comparing the Mean

Square Error(MSE) of the three models, interpreting the SHapley Additive exPlanations(SHAP) plots of the models, and selecting the one that has the strongest fitting ability to predict the hotness of each short video.

2. Literature Review

Nowadays, short videos have formed a novel social mode. How to use machine learning to empower the development of short videos is a topic of great concern for short video platforms nowadays.

For example, for the problem of missing user labels due to the rapid growth of the user volume of short video platforms, some users' authentication requests cannot be passed in a timely manner, Xuesong Lin [1] uses the XGBoost decision tree model to train the authentication labels of short videos to supplement the lack of authentication labels.

Furthermore, Xiaowei Sun [2] adopts a multimodal fusion approach, which is used to identify emotional states such as happiness, sadness, anger, and disgust in the sample with high accuracy.

One of the hottest topics in the field of using machine learning to fuel the development of short videos is how to allow users to actively or passively see their favourite content and increase user stickiness. With the explosive growth in the number of short videos, information overload has begun to appear, and users need to spend a lot of time to find their own favourite content and high-quality content in the massive amount of content. Short video platforms want to better meet user needs and increase user stickiness, they need to do accurate resource analysis to understand users. Therefore, how to effectively enable short video platforms to collect and store a large amount of user data, mine high-quality content through analysis, and convert data into insights so that users can quickly find their favourite content and high-quality content to retain users is the most concerned and the hottest as well as the most challenging topic for short video platforms nowadays.

To address this area of great potential application, Ying Feng et al [3] proposed a cross-media collaborative filtering neural network model for short video click rate estimation. Using RNN structure to process users' historical behaviours and data augmentation to process the training data, a multimodal feature and interest evolution based short video click rate prediction model (MMIE) is proposed.

Based on data mining methods and feature engineering techniques, Ren Ronghuan [4] conducted a prediction study on users' click-playing behaviour based on video clicking behaviour data of industrial real users, through in-depth analysis of users' behavioural data, mining of users' attribute features and video features, and so on. Experiments show that the DeepFM model performs best in prediction accuracy metrics, with an AUC value of 4.38%, 3.44%, and 3.93% higher than the other three decision tree models, and a better prediction performance, second only to the fastest LightGBM model, resulting in the best overall performance.

N.H.T.M. De Siva and R.A.H.M. Rupasingha [5] use seven classification algorithms Random Forest, Support Vector Machine (SVM), Logistic Regression, Multilayer Perception (MLP), Decision Tree, Naive Bayes, and ensemble learning algorithm that combined the five individual algorithms, in order to help users identify the integrity and quality of short videos and take their decisions to view those videos.The accuracy of each algorithm, Precision, F-Measure, and Recall are considered and ensemble learning with 97.21% testing accuracy, random Forest is identified as the prime individual algorithm with 96.89% testing accuracy.

Hariharan R and R. Sophia Janit [6] analyze large amounts of user behavior data such as: search, like, watch, comment, subscribe. By analyzing this data, they use Knn algorithm, predict user preferences and recommend videos which users are likely to like.

Laiche F et al.[7] use quality of experience(QoE) to describes the level of satisfaction or annoyance of users when using a multimedia service or application. They examine socio-environmental factors and user engagement characteristics and investigate their correlation with QoE. And they build a metric for estimating end-to-end QoE for a specific aspect of user behaviour. Then, by simulating mathematical metrics, they use machine learning models to predict QoE.

3. Method

3.1 eXtreme Gradient Boosting Model

EXtreme Gradient Boosting (XGBoost) was developed by Tianqi Chen in 2014 as an improvement and extension of Gradient Boosting Decision Tree(GBDT) [8]. XGBoost offers significant advantages in terms of accuracy, flexibility and efficiency, is good at capturing dependencies between complex data, and able to obtain effective models from large-scale datasets. XGBoost is widely used in various machine learning tasks, including classification, regression, sorting, recommender systems, and so on, and it has been widely used and achieved good results in Kaggle and industry.

As the name suggests, XGBoost is a machine learning algorithm based on Gradient Boosting Tree, which combines the features of the Gradient Boosting framework and Decision Tree Models to progressively improve the predictive performance of the model by iteratively training a series of decision trees. Its goal is to optimise the loss function so that the error between the predicted and actual values is minimised.

The exact process of the XGBoost algorithm is as follows:

1) Input training samples x and set relevant parameters. The number of iterations is K, f is the function space consisting of all trees, fk is a single decision tree model with initial value f_0 = 0. The XGBoost model can be represented as Equation 1.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

2) Define the objective function of the model, see Equation 2.

$$\mathrm{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

3) The additive structure of XGBoost is shown in Equation 3, which can be obtained by substituting it into the objective function and performing a Taylor expansion to obtain Equation 4.

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{3}$$

$$\begin{aligned} \mathrm{Obj}^t &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{i=1}^{T} \left[G_j \omega_i + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right] + \gamma T \end{aligned} \tag{4}$$

$G_j = \sum g_i$,$H_j = \sum h_i$, g_i and h_i are the first and second order derivatives of the loss function, respectively. Let the first order derivative of Obj^t be 0, the corresponding value of the leaf node can be found.

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{5}$$

The value of the objective function at this point is of the form of Equation 6.

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\left[\frac{G_j^2}{H_j+\lambda}\right]+\gamma T \tag{6}$$

4) The greedy strategy is used to generate a new decision tree to minimise the value of the objective function, and to find the corresponding predicted values of the leaf nodes ω_j^* , the newly generated decision tree $f_i(x_i)$ is added to the model to obtain Equation 7.

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{7}$$

5) Keep iterating until the end of N iterations and output the XGBoost model consisting of N decision trees.

The main benefits of XGBoost are as follows:

1) Taylor expansion of the loss function of the second order, so that the optimal solution is solved more efficiently.

2) The objective function is added to the regular term, the complexity of each regression tree will be penalised, which is conducive to reducing overfitting.

3) Parallel computing: XGBoost supports parallel computing at the feature granularity, which can effectively use multiple CPUs to speed up the training process.

4) Flexibility: XGBoost supports user-defined any second-order derivable function as the objective function.

5) Missing value processing: For samples with missing values, XGBoost can automatically learn the split direction of the samples without pre-processing the missing values.

6) Support for multiple languages: XGBoost provides a practical interface including Java, Python, R and other languages.

7) Built-in cross-validation: XGBoost algorithm cross-validation can be used in each round of Boosting iterations, which will be convenient to obtain the optimal number of Boosting iterations.

The main disadvantages of XGBoost are as follows:

1) The algorithm has too many parameters and complex tuning parameters for beginners.

2) The XGBoost algorithm has good processing speed and accuracy for low and medium dimensional data, but compared to deep learning models, XGBoost is unable to model spatio-temporal location, and does not capture high dimensional data such as images, speech, and text well.

3.2 Random Forest Model

Random Forest model (RF) belongs to the integrated learning algorithm, which is a kind of Bagging method, proposed by Breiman in 2001 [9]. RF is a combination of Bootstrap and CART tree, which takes the CART tree as the base classifier, and uses the Bootstrap method to extract a number of different training samples, and when training a single tree it is not not use all the features for attribute segmentation, but will randomly select features.The principle of RF is shown in Figure 1. As shown in the figure, RF consists of several decision trees, each of which is obtained by training on different samples and feature subsets to prevent overfitting. The final prediction result is obtained by integrating the prediction results of each decision tree through voting or averaging, etc. That is, the classification results of several weak classifiers are voted and selected to form a strong classifier.

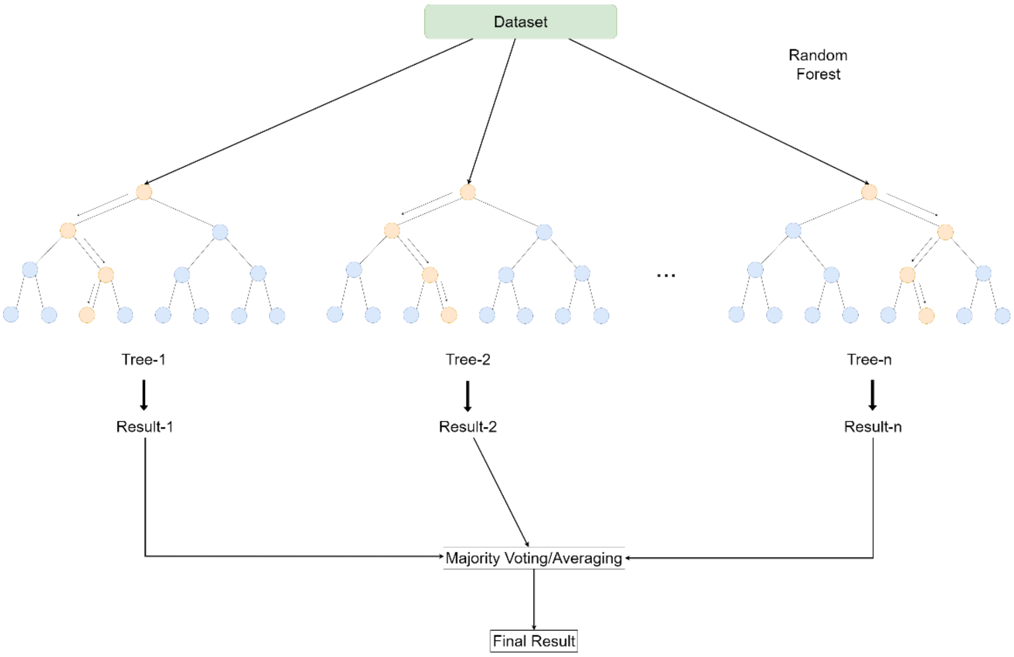

Figure 1: Maximum Pooling Operation Process

The exact process can be divided into three steps:

Step 1: Extract different training sets using Bootstrap method. For the original training set with N samples, N samples are extracted repeatedly with put-back, and the newly generated samples can be used to train a single CART tree next.

Step 2: Randomly select some of the features to do attribute segmentation on a single CART tree. From M total features m are randomly selected (where m<<M), applied to the growth of a single CART tree. During the growth of the random forest, m is kept constant and each decision tree is allowed to grow freely and no pruning operation is performed.

Step 3: Keep repeating step 1 and step 2 until p decision trees are formed, forming a random forest.

Step 4: Predict the test samples based on the random forest created in the above steps. Use each trained decision tree to classify the test sample for prediction, the results obtained for each decision tree are counted and the prediction with the highest number of votes is taken as the final classification result.

The main advantages of RF are as follows:

1) Ability to handle high-dimensional features: the RF is able to handle input samples with high-dimensional features without dimensionality reduction.

2) Strong generalisation ability: sample perturbation and attribute perturbation enhance the differences between the base learners and increase the randomness, resulting in a small variance of the final integrated model and a further increase in generalisation ability.

3) High robustness: RF is insensitive to missing values, outliers and noise, and has good robustness, which can effectively deal with more complex real-world problems.

4) Interpretability: by ranking the importance of features, the Random Forest model can provide results of feature selection and variable importance analysis to help us understand the nature of the problem.

5) The call parameters are simple, and the only parameter to be concerned with is n_estimator.

The main disadvantages of RF are as follows:

1) Not applicable to highly correlated features: on certain noisy sample sets, the Random Forest model is prone to overfitting and requires feature selection or dimensionality reduction.

2) High computational complexity: since the random forest model consists of multiple decision

trees, it has a high computational complexity for training and prediction.

3) High memory consumption: due to the need to store multiple decision trees, the random forest model takes up a large memory space, which may be limited in resource-constrained situations.

3.3 LASSO Regression Model

The Least Absolute Shrinkage and Selection Operator (Lasso) was proposed by Robert Tibshiran in 1996.The Lasso regression restricts the model coefficients by adding L1 regularisation, which has the effect of restricting the model coefficients to as many 0s as possible, which means that feature selection can be performed through this (unselected feature coefficients are reduced to 0.) The objective function of Lasso is shown in Equation 8.

$$J(\beta) = \sum^{'} (y - X\beta)^2 + \sum^{'} \lambda|\beta| \tag{8}$$

where λ is the regularisation factor, transforming the above problem into an equivalent convex optimisation problem, see Eqs. 9. In Eqs. 9, the sum of the absolute values of all regression coefficients is bounded to be no more than a constant t when the sum of squared errors is minimised.

$$\begin{cases} argmin \sum^{'} (y - X\beta)^2 \\ \quad s.t. \sum|\beta| \leq t \end{cases} \tag{9}$$

The advantage of Lasso is to change the penalty term from L2 to L1 paradigm, which can reduce some unimportant regression coefficients to 0, and achieve the purpose of eliminating variables and feature selection.

The disadvantage of Lasso is that the selection of the regularisation factor is more difficult; if the regularisation factor is too large, it will result in an underfitted model, and if it is too small, it will not have a limiting effect on the model.

3.4 SHAP Visualisation

This design uses Python language for model interpretation to complete the SHapley Additive exPlanations (SHAP) visualisation. Breaking down its name, SHapley indicates that for each feature in a single sample, its SHapley Value is computed.SHapley value is derived from the theory of co-operative games, which is a contribution-based allocation; Additive indicates that the Shapley Values of features are additive for a single sample; exPlanation indicates that the deviation of the predicted value from the "mean of predicted values" can be explained by summing up the shapley values of all features. The sample predicted values are calculated in Equation 10.

$$f(x) = g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{10}$$

In Equation 14, $f(x)$ is the sample predicted value, ϕ_0 is the predicted mean, z_i' is the vector of indications for the sample, ϕ is the SHapley Value of each feature. For a given sample, if the feature is not in its decision path, then the SHAP value of the corresponding feature is 0, $\phi = 0$, which indicates that the feature does not impute to the sample and does not contribute to the final predicted value. ϕ_0 is the prediction mean, which is a constant. z is the indicator vector of the sample, which indicates whether the corresponding feature exists (1 or 0). M is the number of input features.

The SHAP values are visualised as shown in Figure 2 to demonstrate the relative importance of each feature. The vertical axis sorts the features by the sum of the SHAP values of all samples, and the horizontal axis is the SHAP value. Each point represents a sample and the gradient colour represents the original value of the feature (red corresponds to high values and blue to low values). The graph averages the absolute values of the SHAP values of each feature, and then ranks the features in descending order from top to bottom based on that average to obtain a feature importance distribution. The higher the ranking, the greater the contribution of the feature factors to the model's accuracy prediction; features with positive SHAP values contribute positively to the final prediction results; conversely, features with negative SHAP values contribute negatively to the final prediction results. As an example, the first line shows that high downloads are positively influencing the number of reviews and low downloads are negatively influencing the number of reviews. The level of importance of a feature indicates the magnitude of the feature's role in the construction of the boosting tree. A feature is more important if it is used as a dividing attribute more times in all trees.

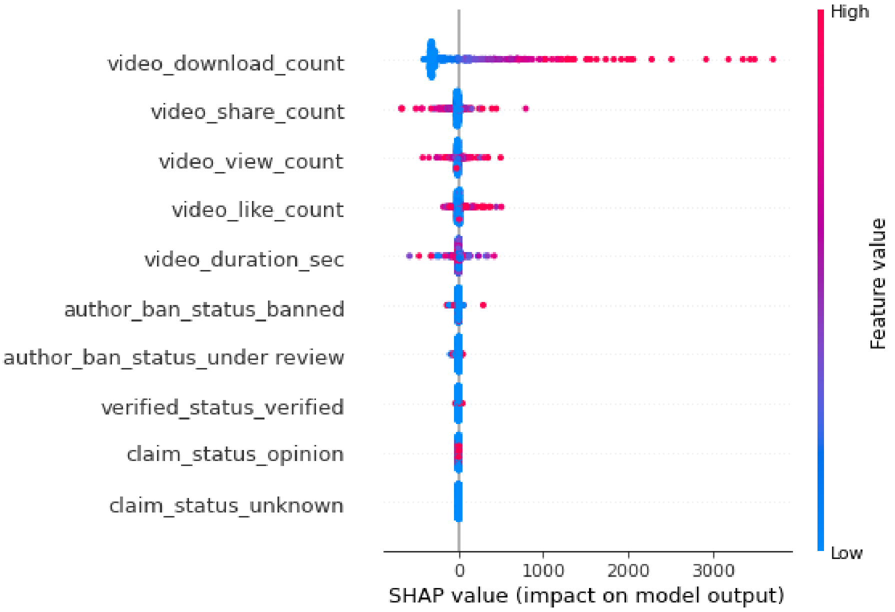

Figure 2: SHAP plot of the number of video comments based on the XGBoost

4. Empirical Analysis

4.1 Data

A dataset is a collection of samples. From a data perspective, reality is uneven and scattered. From the computer's point of view, the normalised input facilitates the computer to learn the laws and leads to the development and improvement of algorithms. From a human perspective, datasets facilitate the validation of the performance of machine learning algorithms within a certain range.

The dataset of this design has a total of 11 features, see Table 1. The distribution of each feature value in the data is shown in Figure 3. The importance of each feature is obtained by the method of gradient boosting, based on the tree after boosting.

Table 1: Characteristics of the dataset

diagnostic property	hidden meaning
claim_status	Declare the state, divided into OPINION and CLAIM
video_id	Video ID
video_duration_sec	Video length may affect the number of comments

video_transcription_text	Video to text
verified_status	Verification status, divided into verified and not verified
author_ban_status	The author banning status, which is divided into three states: active, banned and under review
video_view_count	The number of views may have a positive correlation with the number of comments
video_like_count	The number of likes may have a positive correlation with the number of comments
video_share_count	The number of shares may have a positive correlation with the number of comments
video_download_count	The number of downloads may have a positive correlation with the number of comments

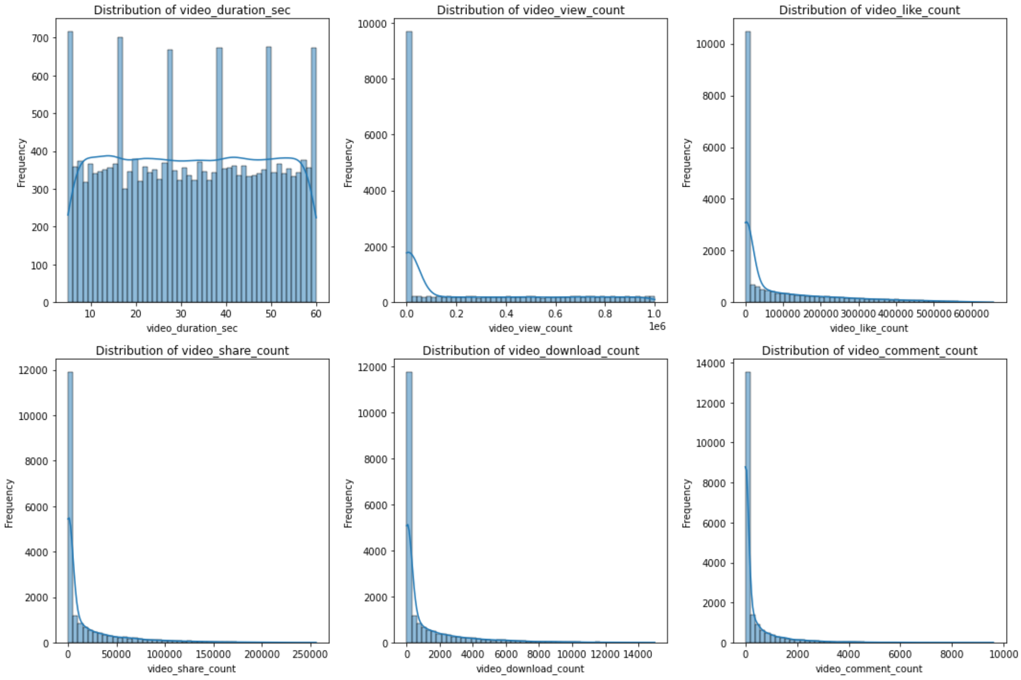

Figure 3: Distribution of individual eigenvalues in the data

4.2 Indicators for Evaluation of Experimental Results

(1) Indicators for evaluating the results of the experiment

1) This design uses Mean Square Error (MSE) as an evaluation index, and the closer the MSE value is to 0, the better the model fit.

MSE is used to measure the deviation of an observation from its true value, and is often used as a measure of how well an observation matches the true value in machine learning models.

During the training of the model, the formula for the training set is as follows:

$$Train = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), \dots, (x_N, y_N) \tag{11}$$

N is the total number of training samples,n $= 1,2,\dots,$ N.

The formula for the test set is as follows:

$$Test = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), \dots, (x_N, y_N) \tag{12}$$

N is the total number of training samples,n = 1,2, . . . , N.

The training model is f(x).The formula for the predicted value is as follows:

$$\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m, \dots, \hat{y}_M\} \tag{13}$$

The formula for the MSE indicator is shown in Equation 14, m is the number of observations, and $h(x_i)$ is the observed value of the i th observation, and y_i is the true value. It is the ratio of the square of the deviation between the observed value and the true value and the number of observations. MSE is a convenient way to measure the "average error", MSE can evaluate the degree of variation of the data, and the smaller the value of MSE, the better accuracy the prediction model has in describing the experimental data.

$$\mathrm{MSE}(X, h) = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2 \tag{14}$$

4.3 MSE

The MSEs of the three models were compared to obtain Table 2.

Table 2: Comparison of model indicators

modelling	MSE
XGBoost	486.21
Random Forest	441.71
LASSO	414.92

The larger the Mse value, the worse the prediction effect; the closer the MSE value is to 0, the better the model is fitted. The data in the table shows that the LASSO model has the smallest RMSE, the best fitting ability, and a significant advantage. The RMSE of the RFl and the XGBoost becomes larger in order, and the span of the difference becomes larger, and the performance is poorer.

4.4 Visualisation of SHAP Values

Figure 2 shows the SHAP values of each influential factor when using the XGBoost model to predict the number of comments. From Figure 2, it can be seen that the biggest determinant of the number of video comments when using the XGBoost model is the number of video downloads, which is strongly positively correlated with the number of video comments, and has a strong positive impact on the number of short video comments. A high number of video downloads means that the video touches the user more and the user is more likely to have the willingness to express opinions and emotions through comments. The main factors affecting the number of video downloads also include the number of video shares, the number of video views, the number of video favourites and the video duration. Among them, when the raw value of the number of video favourites is high, it mainly positively affects the number of short video comments, and in lesser cases, it negatively affects the number of comments. As a positive expression, a high number of video favourites means more likes or approvals from users, when they are more likely to comment. However, when a user likes a video but does not want to spend too much time commenting on it and does not have a strong desire to express his or her opinions or emotions, the user may just click the favourite button without commenting.

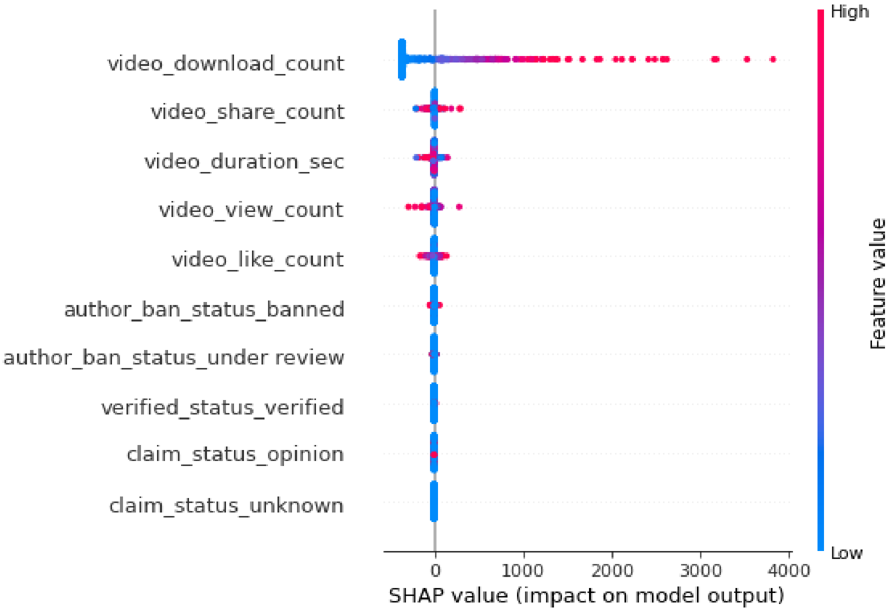

Figure 4: SHAP plot of the number of video comments based on the RF model

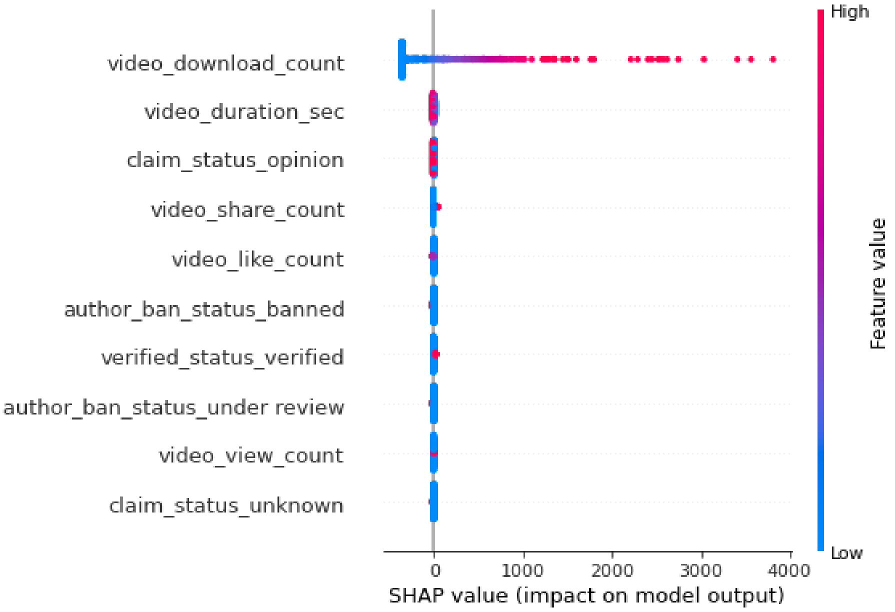

Figure 5: SHAP plot of the number of video comments based on the LASSO model

Figure 4 shows the SHAP values of each influential factor when predicting the number of comments using the RF model. Compared to the XGBoost model, the number of video downloads has a greater influence on the number of video comments in the RF model, and videos are more likely to receive more comments when they have a higher number of downloads. The number of video shares, video duration, number of video views and number of video favourites are also important determinants of the number of video comments, with only a slightly different degree of importance compared to the XGBoost model. The number of video favourites has more positive SHAP values, which means that the number of video favourites tends to increase the number of comments on a video, but the number of video favourites is not the only factor that determines how many comments there are. When the raw values of video duration, number of views and number of favourites are large, the SHAP values for the number of video comments are both positive and negative and are comparable in absolute value.

Figure 5 shows the SHAP values for each influencing factor when predicting the number of comments using the LASSO model. In the LASSO model, the number of video downloads is almost the only factor that influences the number of video comments. From XGBoost model to RF model to LASSO model, the number of video downloads has more and more influence on the number of video comments.

Combined with MSE, it can be seen that among the three models, XGBoost, RF and LASSO, the LASSO algorithm is more accurate and can find the most influential factors more precisely; while the XGBoost model has poorer performance although it is more detailed in analysing the number of video comments.

5. Conclusions and Outlook

5.1 Conclusion

Therefore, the research in this paper mines and analyses the comment information of short videos, providing a direction for short video creators to optimize short video content and improve traffic, and also providing a method for short video operation platforms to screen high-quality videos, continuously improve their quality and popularity, promote the construction of a more high-quality network ecological environment, and further release the potential of digital life.

This design uses data analysis tools to evaluate the effectiveness and performance of short videos, three models, XGBoost, RF and LASSO, will be used to analyse and predict the number of comments on short videos, by comparing the model's MSE, interpreting the model's SHAP chart and selecting the model with the strongest predictive ability to predict the hotness of each short video.

According to MSE, it can be seen that among the three models, XGBoost, RF and LASSO, the LASSO algorithm predicts more accurately and can find the most influential factors more precisely; while the XGBoost model can meticulously find out multiple influencing factors when analysing the number of video comments, but the prediction performance is poor. According to the LASSO model, the heavyweight feature that affects the number of short video comments - the number of video downloads - can be found to capture the essence of the matter; according to the XGBoost model and the RF model, other influencing features such as the number of video shares, the video duration, the number of video views, and the number of video favourites can be found.

5.2 Shortcomings and Future Prospects

In addition to focusing on the amount of short video traffic, short video platforms and video creators should also pay more attention to the rich content and emotional information contained in the network comments, and deeply excavate and analyse the laws of users' viewing behaviour and emotional tendencies. Generally speaking, users mainly comment on the video they watch to express their feelings, so we can analyse the user's viewing habits, emotional expression and other information from these data to extract valuable information.

It is also possible to select useful indicators based on the text content, number of likes and other relevant information of short video network comments to further conduct more detailed correlation analysis, as well as to produce word clouds, sentiment analysis charts and other data visualisations, to find their intrinsic correlations, and to mine effective information.

References

[1] Lin Xuesong. Research on short video user data mining and preference strategy [D].Beijing University of Posts and

Telecommunications, 2024. DOI:10.26969/d.cnki. gbydu.2022.000783.

[2] X. Sun, "Research and Implementation of Short Video Product Recommendation Algorithm Based on Multimodal Features," 2022 2nd International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2022, pp. 392-395, doi: 10.1109/NetCIT57419.2022.00099.

[3] Feng Y,Zhao G, Implementation of Short Video Click-Through Rate Estimation Model Based on Cross-Media Collaborative Filtering Neural Network [J]. Computational Intelligence and Neuroscience. 2022, 2022.

[4] Ren Ronghuan. Research on short video playback prediction based on data mining method [D]. University of International Business and Economics, 2024. DOI:10. 27015/d. cnki. gdwju. 2022. 001259.

[5] N. H. T. M. De Siva and R. A. H. M. Rupasingha, "Classifying YouTube Videos Based on Their Quality: A Comparative Study of Seven Machine Learning Algorithms," 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, 2023, pp. 251-256, doi: 10. 1109/ICIIS58898. 2023.10253580.

[6] H. R and R. S. Janit, "Social Media Content Recommendation System using Knn Algorithm," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1465-1468, doi: 10.1109/ICACITE57410.2023.10182968.

[7] Laiche F, Ben Letaifa A, Elloumi I, et al. When machine learning algorithms meet user engagement parameters to predict video QoE[J]. Wireless Personal Communications, 2021, 116(3): 2723-2741.

[8] Chen T, Guestrin C. XGiBoost: A Scalable Tree Boosting SystemiC]. Prceedings ofthe 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. acm, 2016.

[9] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.