Prediction On Tiktok Like Behavior Based on Random Forest Model

Ying Nie *, Yundong Xu

School of Management, Jiangsu University, Zhenjiang, China, 212013

* Corresponding Author Email: 3210816008@stmail.ujs.edu.cn

Abstract. In recent years, the TikTok short video platform has rapidly ascended, attracting a plethora of users to participate in content creation and interaction. Predicting 'like' behavior delves deeply into user preferences, offering reference value for the platform to enhance traffic. Based on this, the present paper focuses on TikTok 'like' behavior as the research subject and employs a Random Forest model for its prediction. The model's fit was enhanced by optimizing the number of estimators (n_e) and the maximum number of features considered for splitting a node (max_f), aiming to provide a beneficial reference for TikTok and other social media platforms atop optimizing existing research. The results demonstrate that the fitted model boasts a commendable predictive performance, with an accuracy of 99.07%. The application of the model will aid the TikTok short video platform and other platforms in making informed video recommendations to users, thus improving the user experience.

Keywords: TikTok, Like Behavior, Prediction, Random Forest Model.

1. Introduction

TikTok, currently the most popular short video platform, has seen a continuous rise in the number of users and their engagement levels. The 'like' behavior on TikTok, as an expression of users' fondness for content, not only reflects the quality and popularity of the content but also holds significant commercial value for content creators and the platform. From the perspective of user behavior analysis, studying the prediction of 'like' behavior on TikTok aids in better understanding user preferences and needs. Concurrently, this provides an optimization direction for the platform's recommendation algorithms, enhancing the precision of recommended content and improving user experience. In the realm of commercial marketing, businesses can tailor their content marketing strategies based on user preferences to boost the efficacy of advertising campaigns. From the standpoint of public opinion monitoring, the predictive study of 'like' behavior on TikTok is also of paramount importance. Government agencies, enterprises, and other relevant institutions can leverage 'like' data to timely grasp social hotspots and public opinion trends, effectively prevent and control negative public sentiment, and maintain social stability. An in-depth analysis and prediction of 'like' behavior can provide valuable decision-making support for content creators, platforms, enterprises, and governments, promoting the healthy development of the short video industry. Moreover, it opens new perspectives and research directions for related fields, rendering the study of 'like' behavior prediction on TikTok of profound theoretical and practical significance.

However, existing research rarely touches upon the prediction of 'like' behavior, with a focus mainly on exploring influencing factors and analyzing the motives behind liking. Therefore, this study investigates the application of the Random Forest model in predicting 'like' behavior on TikTok based on user data, aiming to optimize existing research and offer a new perspective for platform operators and researchers.

2. Literature Review

With the pervasiveness of the internet, social media platforms have become an integral part of daily life. On these platforms, users express their appreciation, support, or agreement with the content posted by others through likes. For the platforms, understanding user 'like' behavior is crucial as it

helps in enhancing content recommendations, fostering user interaction, and increasing user satisfaction. This paper will explore methods and strategies for predicting user 'like' behavior in the hope of providing beneficial references for platforms and content creators. Prior to this, researchers have conducted relevant explorations into motivations and predictive methods.

2.1. Research on Like Behavior

In the era of rapid internet development, 'liking' has become a novel form of social interaction and an important means for people to express emotions, opinions, and attitudes on online social platforms. From Weibo and WeChat to TikTok and Bilibili, the like feature has become standard across major social platforms. However, research on like behavior has been concentrated on the analysis of liking motives and the study of influencing factors. Wenbao Lü[1] explored the motives behind users' likes by analyzing video characteristics, but the conclusions were specific to high-like short video reports on COVID-19 prevention and lacked generality and universality. Yuzhang and Haobo Zhao[2] studied like behavior from multiple perspectives, including communication psychology, interpersonal communication, and media economics, but their work lacked empirical data and was merely experiential analysis. Xian Hu, Jiang Wu, Kaiyu Liu, and others[3] conducted an empirical analysis of the factors influencing like behavior in WeChat Moments based on social cognition theory, using questionnaires and the Smart PLS2.0 tool. Zeyu Lu, Xianjin Zha, Yalan Yan[4] applied grounded theory to semi-structured interview data and performed a systematic three-level grounded analysis, sorting out the structure and relationships between concepts and categories, constructing a model of the mechanism of like behavior for intelligent recommendation content in social media environments.

2.2. Predictive Studies

Predictive research plays an essential role in science and technology, economics, society, and humanities. Xinmiao Yan, Taolan Sun, Yuhang Lu, and others[5] employed multifactorial Logistic regression to analyze the risk factors for cavities in 12-year-old children, incorporating statistically significant variables from the analysis into a machine learning algorithm to build a model. Fangfeng Zhang and Ran Ni[6] used the SMOTE algorithm to balance the original data, which improved the prediction accuracy of artificial intelligence algorithms. Minhui Zhong, Runa Zhang, Chan Yu, and others[7] constructed a postpartum depression risk prediction model using three supervised learning algorithms: Logistic regression, Support Vector Machine, and Random Forest. They utilized sequential forward selection to filter features and grid search to adjust model parameters. Yanfang Hu, Wen Xiong, Wei Gao[8] built static and dynamic features from user information and behavior log data based on relevance, proposing an ensemble learning-based method for predicting game user churn.

2.3. Research on Random Forest Model

Random Forest is an ensemble learning method that improves prediction accuracy and stability by constructing multiple decision trees and combining their results. The Random Forest model exhibits superior performance in handling unbalanced datasets and feature selection. Jinglei Shen, Huiqun Yu, Guisheng Fan, and others[9] constructed and optimized a prediction model for whether users would interact after browsing posts in an online community by analyzing upvote and downvote data. Nonetheless, issues of redundancy among decision trees and class imbalance warrant further study. Lihua Wu, Zhaojun Wang, Qinghua Chi, and others[10] incorporated user mobility and online behavior characteristics, achieving higher accuracy and recall rates than traditional prediction methods with their complex network model, enhancing network service performance and efficiency. However, the model only considered a rather singular similarity index based on local structural information, which resulted in lower accuracy for unsupervised algorithm comparison methods. Jinyong Gui, Shengjun Li, Jianhu Gao, and others[11] adopted a data-driven approach, introducing the Synthetic Minority Over-sampling Technique for boundary synthesis, and proposed a seismic prediction method for gas saturation based on Random Forest machine learning algorithm training.

Xiaozhi Wang, Fengyun Mou, Yongchuan Zhang, and others[12] proposed a signal decomposition-based prediction model, LE-RL, to fully mine traffic flow series feature rules and reduce the impact of nonlinear and non-stationary series, which showed good predictive performance and generalization ability in short-term traffic flow.

3. Data Analysis

3.1. Data Source

The data for this study were sourced from a batch of TikTok platform user behavior data released by the community, including more than 1.7 million browsing behaviors of 19-year TikTok users. The content of the data set is real and reliable, the data amount is huge, the data is new, therefore, it has universal research significance. The dataset includes basic user information (such as id, city), characteristics of short video content (such as length, music id, channel), and other features (such as whether liked, whether watched to completion, etc.). Table 1 presents the dataset description, which was organized and cleaned to obtain the sample data for analysis.

Table 1: Dataset Description

Field Name	Definition
uid	User ID
author_id	Author ID
finish	Whether Watched to Completion
duration_time	Video Duration
user_city	User City
item_city	Author City
like	Whether Liked
real_time	Exact Release Time
item_id	Video ID
channel	Video Channel
music_id	Music ID
H. date	Hour, Day (of Release)

3.2. Data Preprocessing

(1) Data Sampling. To reduce training costs, data were sampled from the dataset for training purposes. Stratified sampling was employed to obtain a portion of the browsing information as training data, while ensuring a reasonable proportion of 'like' data. The result was that 2.5% of the data was used for training, indicated by the value 0.0252.

(2) Feature Extraction. Irrelevant fields such as channel, finish, H, and date were removed, retaining nine feature attributes, including user characteristics (uid, user_city), content characteristics (item_id, author_id, item_city, music_id, duration_time, real_time), and whether liked (like). The real_time field, containing string object corresponding to time values, was transformed into numerical values by converting it into time differences (in seconds) from a fixed point in time.

(3) Dataset Division. The prediction problem at hand is a binary classification model, with classification outcomes being either 'liked' or 'not liked'.

4. Model Construction

4.1. Introduction to the Basic Concept of the Model

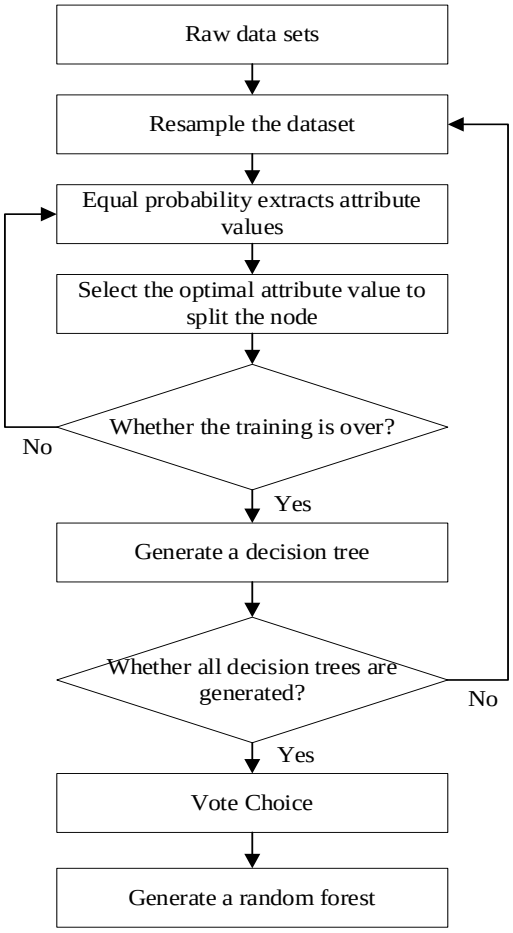

Figure 1. Random Forest Generation Diagram

Breiman[13] introduced Random Forest as an ensemble learning method, which has since achieved significant results across numerous fields. The method combines the concepts of both bagging and random feature selection, constructing multiple decision trees and integrating their results to produce the final prediction. The role of multiple decision trees is analogous to combining many nonlinear relationships to form a more intricate nonlinear relationship. Therefore, the Random Forest has several advantages such as high predictive accuracy, high tolerance to outliers and noisy data, and enhanced accuracy, stability, and interpretability. It has a strong advantage in handling datasets with complex features. The generation process of a Random Forest is depicted in Figure 1.

The prediction flow of the 'like' behavior during TikTok user browsing is illustrated in Figure 2, the middle part is the specific research process, the left part suggests the data sets needed to be used in different processes, and the right side describes the specific operation of the corresponding process in detail.

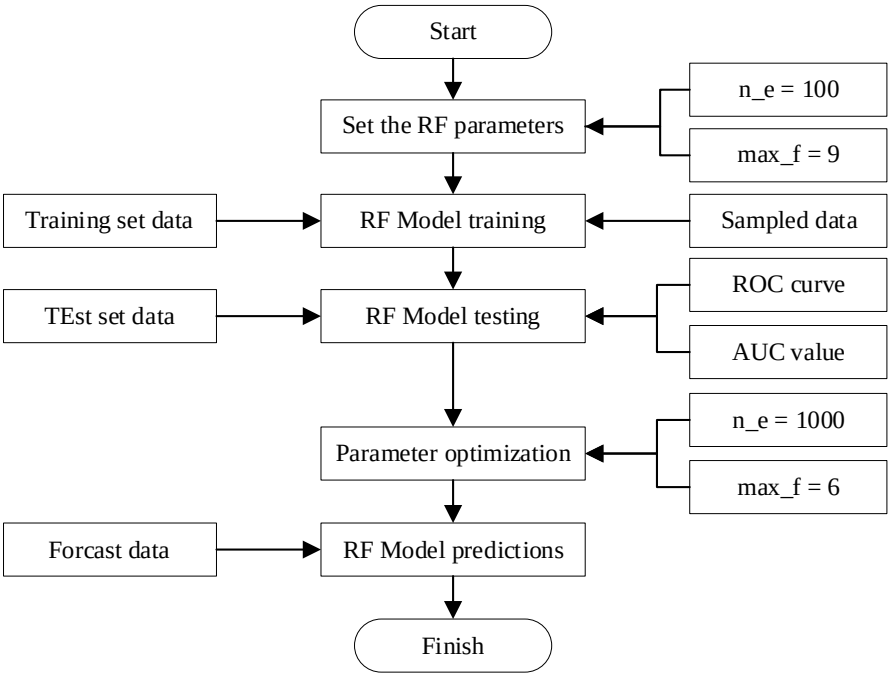

Figure 2. Prediction Process Diagram

4.2. Introduction to Evaluation Metrics

The ROC curve and the AUC value are two critical metrics for assessing the performance of binary classification models. The ROC curve provides an intuitive understanding of classifier performance at different thresholds, while the AUC value allows for a quantitative evaluation of the classifier's overall performance.

(1) The ROC curve (Receiver Operating Characteristic Curve) is a graphical tool that depicts a classifier's performance, showing the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds. The True Positive Rate is understood as the proportion of positive instances correctly identified (1 - the rate of missed diagnoses), thus a TPR closer to 1 is better. The False Positive Rate is the proportion of negative instances incorrectly identified as positive (rate of false diagnoses), and hence an FPR closer to 0 is preferable.

(2) The AUC (Area Under the Curve) represents the area under the ROC curve and measures classifier performance. An AUC value closer to 1 indicates better classifier performance, while an AUC value closer to 0 indicates poorer performance.

4.3. Model Pre-Training

The results of the model pre-training are shown in Figure 3. The Random Forest (RF) model yielded an AUC value of 0.541, with a prediction accuracy of 98.05% .

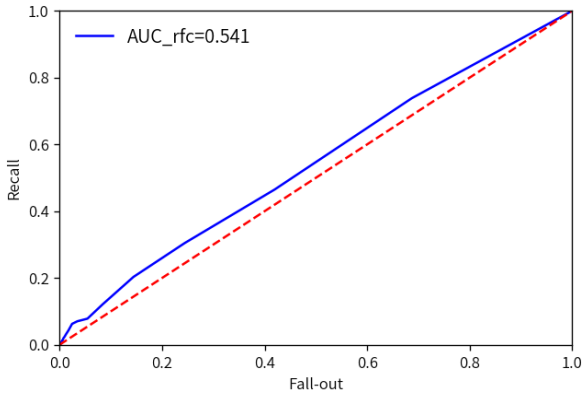

Figure 3. Model Pre-Training Results

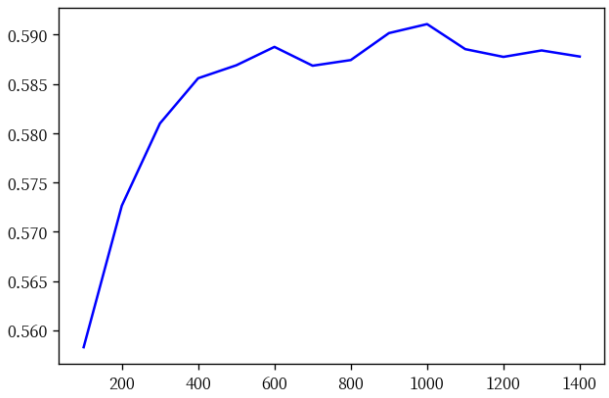

Figure 4. n_e Optimization Results

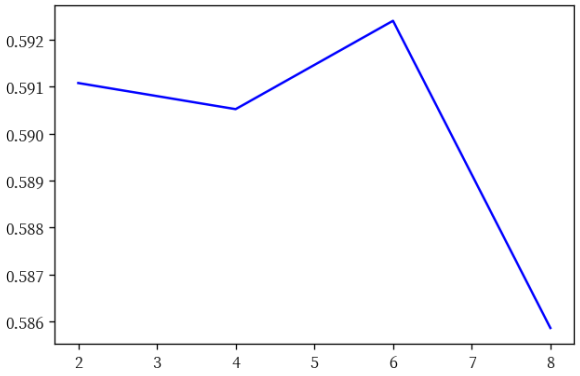

Figure 5. max_f Optimization Results

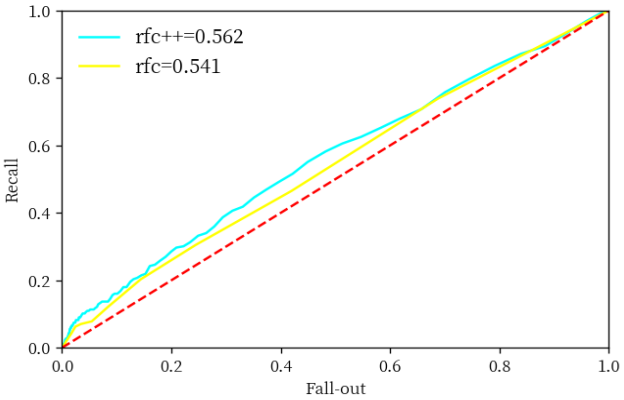

Figure 6. Before and After Optimization Comparison

5. Model Optimization

5.1. n_e Optimization

The parameter `n_estimators` represents the number of subsamples generated from the original dataset through bootstrap sampling, which corresponds to the number of decision trees. If `n_estimators` is too small, the model may underfit, whereas too large a value might not significantly improve the model. In the pre-training of the model, `n_estimators` was set to 100. As shown in Figure 4, the AUC value was optimized when `n_estimators` was increased to 1000, indicating the best model performance.

5.2. Max_f Optimization

With the `n_estimators` parameter nearing its optimum, the `max_features` parameter was optimized. This parameter represents the maximum number of features to consider when looking for the best split in constructing the decision trees, aiming to enhance the fitting capability of each submodel and thus the overall prediction accuracy of the model. As demonstrated in Figure 5, an optimal AUC value was achieved when the number of features `max_features` was set to 6, indicating the best model fit.

5.3. Model Training

Figure 6 illustrates that after optimization, the AUC value and the model's performance have noticeably improved, with an accuracy rate reaching 99.07%. Although the prediction accuracy is exceptionally high, the proportion of non-liked data is significant, implying that a model predicting all instances as non-liked would still achieve an accuracy of 99.03%, suggesting certain limitations in the data.

6. Conclusion

This paper constructed an effective prediction model based on the Random Forest model through the analysis of 'like' behavior on TikTok. The experimental results indicate that the model has high accuracy and stability in predicting 'like' behavior on TikTok. By predicting user 'like' behavior, the platform can provide better personalized recommendations, enhancing user engagement and satisfaction.At the same time, the application of the model is universal, and it can provide a reference for the prediction of other entertainment platforms.However, predicting user behavior is a dynamic process that requires constant model updates and optimization. In future research, we aim to explore more advanced machine learning algorithms and feature engineering methods to improve model performance and provide more intelligent recommendation services for the platform. It is important to note that predicting user 'like' behavior does not equate to manipulating user actions. The platform's

primary intent should be to offer better services and experiences, not to misuse data advantages for inappropriate purposes. Furthermore, we will also focus on user privacy protection issues to ensure that the platform respects and safeguards users' privacy rights while providing quality services, achieving mutual growth for both the platform and its users.

References

[1] Lü Wenbao. Analysis of Characteristics of Highly Liked Short Videos on Mainstream Media's TikTok Accounts: A Case Study of COVID-19 Prevention and Control Videos by Three Mainstream Media Outlets[J]. Media, 2021, (15): 53-55.

[2] Zhang Yu, Zhao Haobo. A Multidimensional Study of 'Like' Behavior in Social Media[J]. Journal for News Buffs, 2017, (08): 8-11. DOI: 10.16017/j.cnki.xwahz.2017.08.003

[3] Hu Xian, Wu Jiang, Liu Kaiyu, et al. Study on the Factors Influencing 'Like' Social Interaction Behavior: Based on the Context of WeChat Moments[J]. Information Science, 2020, 38(01): 36-41. DOI: 10.13833/j.issn.1007-7634.2020.01.006.

[4] Lu Zeyu, Zha Xianjin, Yan Yalan. Study on the Influence Mechanism of 'Like' Behavior for Intelligent Recommendation Content in Social Media Environments[J]. Modern Information, 2023, 43(02): 42-55.

[5] Yan Xinmiao, Sun Taolan, Lu Yuhang, et al. A Machine Learning-based Prediction Model for Dental Caries in 12-year-old Children in Sichuan Province[J]. West China Journal of Stomatology, 2023, 41(06): 686-693.

[6] Zhang Fangfeng, Ni Ran. Research on Online User Consumption Prediction Model Based on SOMTE Balance[J]. Mathematics in Practice and Theory, 2020, 50(15): 49-59.

[7] Zhong Minhui, Zhang Runa, Yu Chan, et al. Construction and Validation of a Predictive Model for Postpartum Depression Risk[J]. Journal of Nursing, 2023, 38(15): 76-81.

[8] Hu Yanfang, Xiong Wen, Gao Wei. A Method for Predicting User Churn in Online Games Based on the Spark Platform[J]. Computer Engineering and Science, 2022, 44(10): 1730-1737.

[9] Shen Jinglei, Yu Huiqun, Fan Guisheng, et al. Design and Implementation of a Recommendation System Based on Random Forest Algorithm[J]. Computer Science, 2017, 44(11): 164-167+186.

[10] Wu Lihua, Wang Zhaojun, Chi Qinghua, et al. Prediction of Opportunistic Connections in Mobile Internet Based on User Behavior[J]. Journal of Wuhan University (Engineering Edition), 2019, 52(01): 89-94.

[11] Gui Jinyong, Li Shengjun, Gao Jianhu, et al. A Random Forest Prediction Method for Gas Saturation Based on Feature Variable Expansion[J/OL]. Lithologic Reservoirs, 1-11 [2024-01-06].

[12] Wang Xiaozhi, Mou Fengyun, Zhang Yongchuan, et al. Short-term Traffic Flow Prediction Using Taxi GPS Trajectory Data: A Case Study of the Jiefangbei District in Chongqing City[J]. Science Technology and Engineering, 2023, 23(28): 12265-12274.

[13] BREIMAN L. Random forests ［J］. Machine Learning, 2001, 45(1): 5-32.