Research on Customer Retention Prediction Model of VOD Platform Based on Machine Learning

Quansheng Zhao, Zhijie Zhao*, Liu Yang,Lan Hong, Wu Han
School of Computer and Information Engineering, Harbin University of Commerce, Harbin, Heilongjiang, 150028, China

***Abstract*—Advanced wireless technology and smart mobile devices allow users to watch Internet video from almost anywhere. The major VOD platforms are competing with each other for customers, slowly shifting from a "product-centric" strategic goal to a "customer-centric" one. At present, existing research is limited to platform business model and development strategy as well as user behavior research, but there is less research on customer retention prediction. In order to effectively solve the customer retention prediction problem, this study applies machine learning methods to video-on-demand platform customer retention prediction, improves the traditional RFM model to establish the RFLH theoretical model for video-on-demand platform customer retention prediction, and uses machine learning methods to predict the number of customer retention days. The Optuna algorithm is used to determine the model hyperparameters, and the SHAP framework is integrated to analyze the important factors affecting customer retention. The experimental results show that the comprehensive performance of the LightGBM model is better than other models. The total number of user logins in the past week, the length of video playback in the same day, and the time difference between the last login and the present are important features that affect customer retention prediction. This study can help companies develop effective customer management strategies to maximize potential customer acquisition and existing customer retention for maximum market advantage.**

Keywords—Video-on-demand platform; Customer Retention Forecast; RFM Model; Machine Learning; SHAP

I. INTRODUCTION

With the advent of the information age, market competition has become more and more intense, and enterprises have slowly changed from a "product-centric" strategic goal to a "customer-centric" one. In the case of limited resources, major e-commerce companies compete with each other for customers, the pursuit of a larger share of the market, to maximize their profits has become the development of each enterprise imperative.

Video-on-demand is developed in the 1990s, and is called "Video On Demand" in English, so it is also called "VOD". As the name suggests, it is a video-on-demand system that plays programs on demand by viewers. Watching video has become one of the most popular online activities and provides a huge market for various video content providers. Better networks, technological innovations and the availability of smart devices have changed the way people are entertained, allowing platforms to deliver services to viewers directly online via the Internet [1]. For any service to grow, it is crucial to understand the values and consumption habits of consumers. In real systems, there are users who stay for a long time and others who enter the system and leave soon after. Therefore, understanding user behavior and further predicting customer retention on VOD platforms can help platforms improve service quality and avoid customer churn. For users with short retention time, the platform can adopt ways such as trial membership and issuing coupons to let users experience the core services to retain them; while for users with longer retention time, the platform can appropriately provide more rewards or service updates to extend their retention time [2].

In the study for video-on-demand platforms, Zhang et al. take Tencent's over-the-top on-demand rule as an example to analyze the reasons for the model, summarize and analyze the problems that arise when current video-on-demand platforms innovate their profit models, and propose strategies to optimize their profit models and corresponding strategies [3]. Hu et al. studied user behavior and access patterns through a major ISP in Shanghai, China, focusing on comparing behavior and access patterns across platforms, and found that user migration across multiple platforms was common and highly influenced by the different characteristics of the platforms [4]. Köster et al. explore the relationship between social referrals, referral propensity, and the stickiness of video-on-demand websites, comparing consumers who are referred by social networks with consumers who arrived at the site through natural search or social media ads to understand the stickiness of the site. The results showed that consumers who were recommended via social referrals spent more time on the site, viewed more pages, and launched more videos than consumers who responded to social media ads, but less than consumers who went through natural search [5]. Wang et al. analyzed the differences between Chinese video platforms and U.S. video platforms, reflecting on the relationship between platform, market, and country. The historical factors and geographical characteristics that influence the operation, structure and governance of video platforms are explored [6]. Rahman et al. studied video consumer behavior from the perspective of VOD platforms. The analysis of the data revealed interesting features of video-on-demand platforms, such as the viewing habits and viewing patterns of different users and the correlation between user profile information (e.g., age, gender) and viewing habits, among others, and suggested that future user behavior could be predicted by learning from previous user behavior patterns[7].

The existing research by scholars on video-on-demand platforms is limited to two major aspects of platform business models and development strategies [6] and user behavior[7] studies, with few studies focusing on their customer retention predictions. To solve this problem, this paper proposes an improved RFM theoretical model based on user behavior and constructs a VOD customer retention prediction model based on machine learning to exploit the advantages of machine learning algorithms in prediction. Recent studies have shown that LightGBM, XGBoost, and CatBoost can show better performance in the face of complex, highly nonlinearly related data [8][9], and the Ensemble Learning (EL) algorithm integrates multiple machine learning models to form a model with stronger generalization and better backfitting capabilities[10]. Therefore, this paper uses LightGBM, XGBoost, CatBoost and other algorithms to construct customer retention prediction models for video-on-demand platforms. After the models are constructed, the accuracy of different models in this study scenario is compared by time complexity, R^2, MAE, RMSE indexes and the optimal prediction model is selected, and the prediction results are later analyzed using the SHAP interpretation method to explore the magnitude of the impact of different characteristics of users on the prediction values and the reasons behind them. The study found that a user's total number of logins in the last week, the length of the day's video playback and the time difference between the most recent login and the present were important features that influenced customer retention predictions.

II. Related Works

In the Internet industry, users who start using an application at a certain time and continue to use the application after a certain period of time are recognized as retained users. Customer retention measurement and analysis is a classic problem in various domains, such as online community platforms [11], telecommunication industry [12] and e-commerce [13]. Several works focus on the measurement and analysis of characteristics related to customer retention. Jiang et al. proposed a maximum entropy semi-Markov model to predict the customer life stages that need to be segmented for milk powder products, which is applicable to the case where the infant life stage transition is deterministic, but not to the case where there is no explicit life stage, such as video-on-demand systems[14].

In recent years, RFM models have also been widely used in customer behavior prediction studies. Marín et al. modeled user behavior based on the traditional proximity, frequency and currency (RFM) model to obtain a proximity, frequency, importance and duration (RFID) model of customer assessment from the perspective of customer-contact center interactions, and showed that the model can be generalized to any environment requiring classification or regression algorithms in any environment that requires classification or regression algorithms [15]. Perišić et al. proposed an extended framework of new proximity, frequency, and monetary value features for predicting user churn in the mobile gaming domain by combining features related to user lifecycle, intensity, and rewards, and indicated that the top five most important features of a multivariate churn prediction model include long-term and short-term frequency features, monetary, intensity, and lifecycle features[16]. Wei et al. first established an RFLP metric system for predicting MOOC user learning behavior and attrition by improving the RFM model in the business domain; secondly, histogram and chi-square tests were used to determine the characteristic variables affecting MOOC user attrition; finally, a MOOC user attrition prediction model was constructed by combining the Grouped Data Processing (GMDH) network as a post-processing information system [17]. Smaili et al. modified the model by adding diversity "D" as the fourth parameter, referring to the diversity of products purchased by a given customer, and the RFM-D based model was applied to the retail market to detect customer behavior patterns and the proposed model improved the quality of customer behavior prediction [18].

In order to effectively solve the customer retention prediction problem, this paper firstly improves the traditional RFM model in terms of extracting features and proposes the RFLH theoretical model suitable for customer retention prediction of VOD platform. After that, machine learning and other latest technologies are applied to the VOD customer retention prediction problem, and the advantages of machine learning algorithms in prediction are exploited to learn users' previous behavior patterns to predict their future behavior. And Optuna framework is used to optimize the model hyperparameters to improve the model prediction performance. Finally, the SHAP interpretation framework is combined with the interpretation analysis of different user behaviors in order to propose targeted suggestions and strategies for customer management in VOD platform.

III. Building a Theoretical Model of Customer Retention based on Improved RFM

RFM model is one of the most important methods to analyze the behavioral characteristics of customers, classifying them by three behavioral variables: proximity R (Recency), frequency F (Frequency), and value M (Monetary) [19]. The traditional RFM model predicts the future short-term behavior of customers through their past purchase behavior in e-commerce platforms. In the existing research, this model is mostly used for the comprehensive consideration of user activity, loyalty and consumption ability to further achieve user value identification and value group segmentation. It is generally believed that users with a shorter interval between recent consumptions and a larger number and amounts of recent consumptions have a higher recognition of products and services and therefore a lower tendency to churn; conversely, users with a longer interval between recent consumptions and a smaller frequency and amount of recent consumptions have a higher tendency to churn and a lower value to the platform. vod platform is different from ordinary e-commerce platforms, namely: most vod users in the viewing process does not generate actual consumption, but the two have a certain degree of relevance in some way.

Currently, the RFM model has been used in user churn prediction [13][20]. In this study, the RFM model is improved on the basis of RFM model to construct RFLH indicator system for vod user behavior and customer retention prediction, and the details are shown in Table I. In RFLH indicator system the indicator H(History) represents the number of days of historical retention, including the statistical characteristics of the number of days of historical retention in the last month, which is a more critical characteristic variable for accurate prediction of customer retention days. The improved RFM model is shown in Table I.

TABLE I. IMPROVED RFM THEORETICAL MODEL

Classification index	Indicator meaning
R(Recency of viewing)	The time interval between the viewer's most recent login and the observation point
F(Frequency of viewing)	The number of times the viewer logged in during a certain period
L (Length of viewing)	Viewer's viewing time and completion in a certain period
H (Days of historical retention)	The number of days the viewer retention in a certain period

IV. MACHINE LEARNING-BASED CUSTOMER RETENTION PREDICTION MODEL CONSTRUCTION

A. Machine Learning Algorithm Selection

In order to select the most suitable machine learning algorithm for customer retention prediction on Vod platform, this study constructs regression models based on three machine learning algorithms, LightGBM, Catboost and XGBoost, respectively, for comparison experiments, and selects the model with the best overall performance.

LightGBM is a distributed gradient boosting framework based on decision tree algorithm. Designed to provide a fast, efficient, low-memory, high-accuracy tool that supports parallelism and large-scale data processing, LightGBM achieves linear acceleration in data computation by reducing the memory use of data, reducing communication costs, and improving efficiency when multiple machines are parallel. Its advantages are faster training efficiency, low memory usage, and higher accuracy. The disadvantage is that it may grow deeper decision trees and produce overfitting[21].

CatBoost is a library of gradient boosting algorithms, which has the advantages of overcoming the gradient bias and effectively solving the problem of prediction bias, improving the accuracy of the algorithm, enhancing the generalization ability, and preventing the occurrence of overfitting phenomenon. Its disadvantages are that it requires a lot of memory and time for the processing of category features, and the setting of different random numbers has an impact on the model prediction results [22].

XGBoost, an efficient Gradient Boosting algorithm, integrates the idea of iteratively generating multiple weak learners and then adding up the prediction results of each learner to get the final prediction result, which has a better performance in structured data processing. The advantage is that the regularization term is added to prevent overfitting and parallel optimization is possible to improve the efficiency of the algorithm. The disadvantages are that it has too many parameters, the tuning parameters are too complex and it is only suitable for processing structured data [23].

LightGBM, CatBoost, and XGBoost are all Boosting algorithms in integrated learning and Boosting algorithms are widely used in the industry and can show better performance in the face of complex, highly nonlinear data [9], which is applicable to the data in this paper.

B. Optuna Tuning Framework

In order to optimize the performance of individual machine learning models and to make them comparable with each other. To set the hyperparameters of the models, the methods usually used are "grid search" or "random search". However, the "grid search" method requires more computational power and time due to the larger parameter space. The advantage of random search is that the search is fast, but it is easy to miss some important information and difficult to determine the distribution of the parameters. Therefore, for hyperparameter optimization, a software framework called Optuna automatic hyperparameter optimization is used in this experiment.

Optuna is a framework designed for automated and accelerated research, It has three advantages: (1) define-by-run API that allows users to construct the parameter search space dynamically, (2) efficient implementation of both searching and pruning strategies, and (3) easy-to-setup, versatile architecture that can be deployed for various purposes, ranging from scalable distributed computing to light-weight experiment conducted via interactive interface [24]. The operation procedure is as follows:

1) Determine the optimization direction, parameter type, value range and maximum number of iterations.

2) Enter the cycle.

3) Select a group of individuals evenly within the function that defines the range of parameter values.

4) The trimmer automatically terminates hopeless populations according to pruning conditions.

5) Calculate the individual objective function value of the unpruned population.

6) Repeat the above steps until the maximum number of iterations are reached and out of the loop.

7) Optimal solution of the output problem [25].

C. Model Performance Evaluation Metrics

In this study, the time complexity is the sum of training time and prediction time, and the higher time complexity will affect its practical application to some extent. A lower time complexity helps to accomplish fast and effective prediction [26]. Meanwhile, the model is evaluated by using three indicators: mean absolute error MAE, root mean square error RMSE and coefficient of determination R2, and the evaluation indicators are calculated as follows:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(y_{rt} - y_{pt})| \qquad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_{rt} - y_{pt})^2} \qquad (2)$$

$$R^2 = \frac{[\sum_{t=1}^{n}(y_{rt}-\overline{y}_r)\cdot(y_{pt}-\overline{y}_p)]^2}{\sum_{t=1}^{n}(y_{rt}-\overline{y}_r)^2\cdot\sum_{t=1}^{n}(y_{pt}-\overline{y}_p)^2} \qquad (3)$$

Where: n is the number of data; y_p is the predicted result; y_r is the true value; $\overline{y}_\mathrm{p}$ and $\overline{y}_\mathrm{r}$ are the average of the predicted and true results, respectively. In the regression prediction, the smaller the RMSE value and the closer the R^2 value is to 1, the higher the interpretability of the model. In this paper, the MAE, RMSE and R^2 are used to make a comprehensive comparison of the model prediction results and verify the prediction accuracy of the model.

V. Empirical Analysis of VOD Customer Retention Predictions

In this paper, the predicted customer retention is defined as the number of days a user logs in the next seven days. For example, if a user's prediction result on January 1 is equal to 3, it means that this user will visit the VOD platform for 3 days in the next 7 days (January 2~8). In this paper, we first pre-process the data to improve the data quality. The features are extracted according to the improved RFM model and combined with machine learning methods to build a video-on-demand platform customer retention prediction model. The specific process is shown in Fig. 1.

A. Data Processing and Feature Selection

iQIYI is China's and the world's leading high-quality video entertainment streaming platform, with more than 500 million users enjoying entertainment services on iQIYI every month. In this paper, we use the dataset provided by iQIYI AI competition platform, which contains video data, user personal information, user startup logs, user viewing and interaction behavior logs, etc. The detailed fields have various information such as the user's device type, device storage, device running memory, gender, age, education status, and occupational status.

The main natural attribute variables of Vod users are set in the paper, including gender, age group, education level, occupational status, and device information. The viewing behavior characteristics variables are classified based on indicators R, F, L, and H. The data of 15,000 users were randomly selected as samples. Based on the behavior records, the attribute variables and viewing behavior characteristic variables of each user were extracted separately. Each indicator was divided by the research as shown in the Table II, and a regression model was constructed to predict the number of days users would log in in the next seven days.

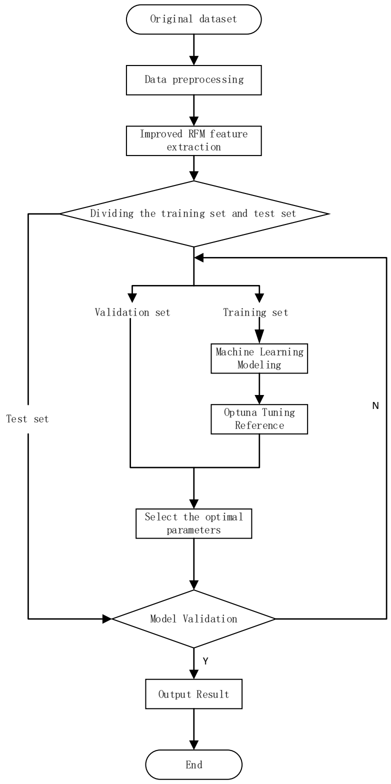

Fig. 1. Flow chart of customer retention prediction model for VOD platform.

TABLE II. Improved RFM Model

Variable Type	Indicators		Meaning
Natural Property Variables	X Properties	X1	Sex
		X2	Age
		X3	Education
		X4	Occupational Status
		X5	Device ram
		X6	Device rom
		X7	Device type
Watch behavioral characteristics variables	R(Recency of viewing)	R1	Login or not for the day
		R2	Time difference between last login and current
	F(Frequency of viewing)	F1	Total number of logins in a month
		F2	Total number of logins in the last week
	L (Length of viewing)	L1	Video playback duration for that day and each of the previous seven days
		L2	Number of video plays per day for the day and the first seven days
		L3	Video completion of the day and each day of the previous seven days
	H (Days of historical retention)	H1	Median number of historical retention days in a month
		H2	Average of the previous four weeks of historical retention days
		H3	Weighted average of the previous four weeks' historical retention days

B. Model Training

After the new dataset was constructed with a total of 953112 data, the missing values outliers in the new dataset were processed, such as filling the missing values in the data with 0, and removing the ones with more than 24h of playing time in a day. The processed data were tested for correlations of the feature variables by Pearson, Spearman, and Kendall three-class correlation coefficients, as shown in Fig. 2, 3, and 4. The test results showed that the selected feature variables were statistically correlated with the target variables.

After data processing, the last log-in time point of 15,000 users recorded in the data was subtracted by seven days as the test set, 80% of the remaining data was used as the training set, and 20% was used as the validation set to input three machine learning algorithms, LightGBM, CatBoost, and XGBoost, respectively, for model training.

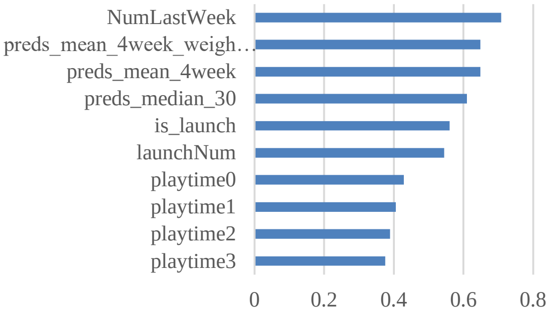

Fig. 2. Pearson correlation test graph (partial).

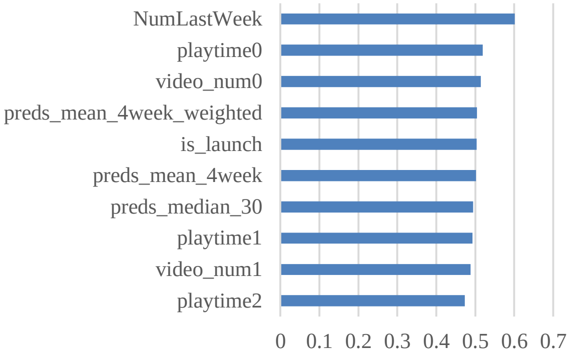

Fig. 3. Spearman correlation test graph (partial).

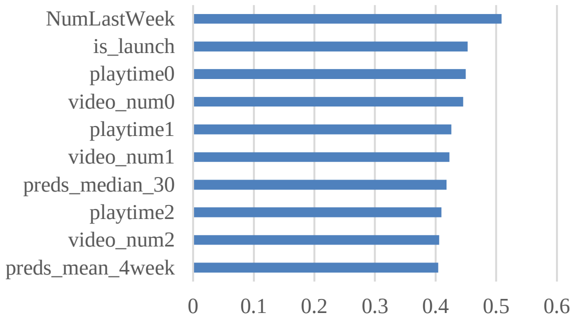

Fig. 4. Kendall correlation test graph (partial).

C. Hyperparameter Tuning

Hyperparameters are the framework parameters of machine learning models, and machine learning algorithms contain various hyperparameters that can be set to improve the accuracy of the model. To set a suitable set of hyperparameters, a hyperparameter tuning approach can be used, which objectively scans various attributes of the model hyperparameters and selects an optimal subset to improve the performance and effectiveness of learning, resulting in a model with optimal performance on a given dataset. For the machine learning-based customer retention prediction model, the Optuna optimization framework is used to tune the main hyperparameters of the prediction model, and the main hyperparameters obtained according to the Optuna optimization framework are shown in Table III.

TABLE III. MODEL'S MAIN HYPERPARAMETERS

Machine learning models	**Hyperparameter Name**	**Optimal hyperparameter values**	**Meaning and Impact**
LightGBM	objective	regression	Assign learning tasks and corresponding learning objectives.
	max_depth	8	The maximum depth of the tree model. The most important parameter for preventing overfitting, decisive for model performance and generalization ability
	learning_rate	0.01	shrinkage rate, choosing a relatively small learning rate can achieve stable and better model performance
	n_estimators	724	The number of iterations of boosting, generally speaking, the more iterations the better the model performance, but too many iterations will often lead to overfitting of the model and affect the training time of the model
	lambda_l1	8	L1 regularization parameter, used to adjust the control overfit
	lambda_l2	7	L2 regularization parameter, used to adjust the control overfit
	num_leaves	1231	Number of leaves on a tree, the larger the value of

			the maximum number of leaf nodes, the more accurate the model, but too large may be over-fitted.
	subsample	0.8	The sampling ratio of training samples can be used to accelerate training and handle overfitting.
	max_bin	247	max_bin indicates the maximum number of bins for the features. A smaller max_bin makes training faster, a larger max_bin makes the model more accurate, but too large a max_bin can lead to overfitting.
CatBoost	depth	5	The maximum depth of the tree model.
	learning_rate	0.01	Used for reducing the gradient step
	n_estimators	5526	The maximum number of trees that can be built when solving machine learning problems.
	l2_leaf_reg	9	Coefficient at the L2 regularization term of the cost function.
	max_bin	209	The number of splits for numerical features.
	bootstrap_type	Bernoulli	Defines the method for sampling the weights of objects
	subsample	0.78	Sample rate for bagging
XGBoost	max_depth	5	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
	learning_rate	0.01	Step size shrinkage used in update to prevents overfitting.
	n_estimators	2542	n_estimators indicates the number of integrated weak evaluators. The larger n_estimators, the better the learning ability of the model.
	alpha	7	L1 regularization term on weights. Increasing this value will make model more conservative.
	gamma	2	Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger is, the more conservative the algorithm will be.
	subsample	0.7	Subsample ratio of the training instances.

D. *Analysis of Experimental Results*

The trained model is tested with a test set, and the model is evaluated by time complexity, mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). The prediction results of the different models were compared and the results are shown in Table IV. The comparison of training time and prediction time and time complexity of different algorithms are shown in Fig. 5, 6 and 7.

TABLE IV. COMPARISON OF MODEL PREDICTION RESULTS

Machine learning models	**MAE**	**RMSE**	**R^2**
LightGBM	0.9910	1.3708	0.6174
XGBoost	0.9896	1.3710	0.6173
CatBoost	0.9882	1.3712	0.6172

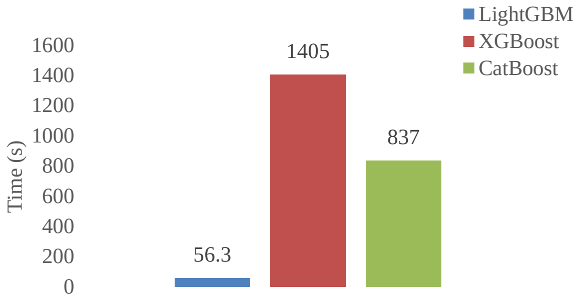

Fig. 5. Comparison of model training time.

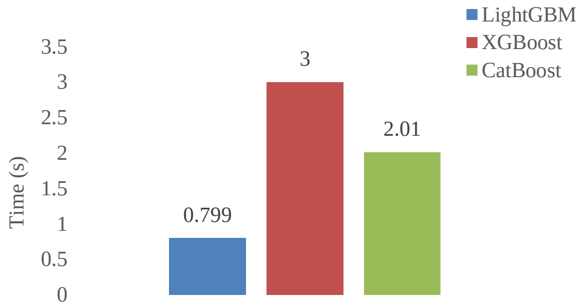

Fig. 6. Comparison of model prediction time.

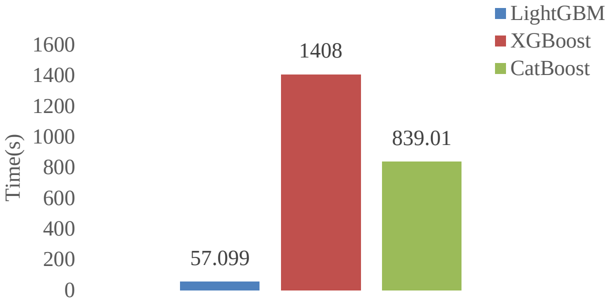

Fig. 7. Comparison of time complexity.

The experimental results show that all three tree-based machine learning models using Optuna can obtain MAE values less than 1, which means that the mean value of model prediction error is within 1 day, with CatBoost having the smallest MAE value of 0.9882. The RMSE metric is strongly influenced by outliers, and in this study the RMSE of all three tree models is around 1.371, with LightGBM has the smallest RMSE value of 1.3708. R² is a measure of the ability of the independent variable to explain the dependent variable. When the R² value is close to 1, the explanatory power of the independent variable is at a high level. Among all models, the strongest explanatory power is LightGBM, followed by XGBoost and CatBoost. For the prediction models constructed in this study, the prediction results of the three models are not very different and all can accurately predict the number of customer login days in the next seven days, but in terms of time complexity, LightGBM is much more efficient than the other two models. In summary, LightGBM was selected as the customer retention prediction model for the video-on-demand platform and further analyzed in this study.

VI. SHAP-based Model interpretation Analysis

SHAP (SHapley Additive ExPlanations) is a method for explaining the predictions of machine learning models. It is an additive explanatory model constructed by Lundberg et al [27] in 2017 inspired by cooperative game theory, which provides a unified way to understand how each feature contributes to the prediction and how the combination of features determines the final prediction. SHAP values are based on the concept of Shapley values in cooperative game theory and are used to fairly distribute the value generated by a set of individuals to each Individuals. Similarly, SHAP values distribute the predictions of a machine learning model to each feature, allowing us to see how each feature contributes to the prediction. SHAP values take into account feature interactions and the relationship between features and predictions, providing a more complete and accurate interpretation than other model interpretation techniques such as partial dependency graphs and feature importance values. SHAP interprets the predicted value of a model as the attributed value of each input feature SHAP interprets the model prediction as the sum of the attributed values (Shap Values) of each input feature, and the Shap Values are calculated as follows:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{in}) \quad (4)$$

Suppose the i-th sample is x_i and there are n features, the baseline of the whole model is the mean of all sample target variables, defined as y_{base}, where $f(x_{i1})$ is the Shapley value of the first feature of sample i, when $f(x_{i1})$ is greater than 0, this feature has a positive effect on the predicted value; when $f(x_{i1})$ is less than 0, this feature has a negative effect on the predicted value. In addition, the TreeSHAP framework developed by Lundberg specifically for tree models has greatly improved the operation efficiency [28].

SHAP Summary Plot combines feature importance with feature effect to reflect the overall positive and negative relationship between feature value and future login days of customers [29]. In order to further clarify the positive/negative relationship between each indicator and the target variable, this paper uses Tree SHAP for model interpretation. As shown in Fig. 8, the horizontal coordinate in the figure is SHAP value, each row represents a feature variable, and each point represents a sample. The redder the point color is, the larger the value of the feature itself, and the bluer the color is, the smaller the value of the feature itself. A positive value represents a positive impact, while a negative value represents a negative impact.

Fig. 8 shows the summary of SHAP features of LightGBM. From the figure, we can see that the three most important features in the model are the total number of logins in the recent week, the length of video playback on the same day, and the time difference between the most recent logins and the present. For the R(recency) feature, the customers who log in on the current day have a positive impact on the number of login days in the future, the greater the time difference between the latest login and the present, the greater the negative impact on the future login days; For the F(frequency) feature, the more the total number of logins in the past week and the total number of historical logins, the more accustomed users are to using the platform, and the higher the positive impact on the future login days of customers; For the L(length) feature, the more playback duration and number of the current day and the previous seven days, the higher the positive impact on the future login days of customers; while the higher the video completion degree of the current day and the previous seven days, the higher the negative impact on the future login days of customers; For the H(history) feature, the constructed historical retention days feature reflects the login characteristics of users to some extent. The mean, median and weighted mean of historical retention days all have an impact on the future login days of customers. The larger the mean and weighted mean of historical retention days, the higher the positive impact on the future login days of customers.

As for the differences in personal characteristics, the overall effect on the predicted results is little. As shown in the figure, customers with jobs are more likely to retain than those without jobs. The device type has no or positive impact on the future login days; the RAM and ROM of the device will have an impact on the performance of the device. In this experiment, the larger the ram and rom values are, the higher the positive impact on the future login days of the customers; the gender has a mixed impact on the prediction results; the younger the age is, the higher the positive impact on the future login days of the customers.

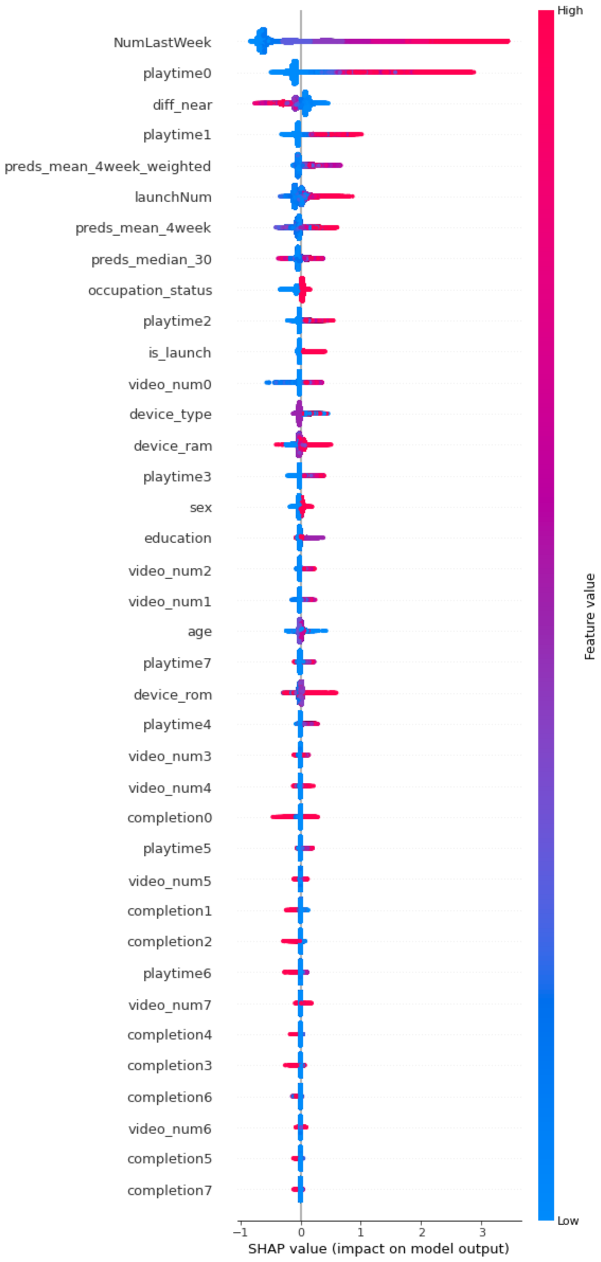

Fig. 8. LightGBM SHAP feature summary chart.

For companies, for users whose model predicts 0 logins and have not logged in for a long time, reduce the investment in this area in the future marketing process. For customers who are predicted to log in in the future, the focus should be on the total number of logins in the past week, the length of video playback on the same day and the time difference between the last login and the present. Companies can establish a membership system as well as a point system to obtain points through sign-ups to send members, sign-ups to receive points as well as accumulate viewing hours and complete tasks, and transform them into premium privileges that can be enjoyed to further strengthen user stickiness.

VII. Conclusion and Outlook

The rapid development of information technology has made watching videos one of the most popular online activities and provides a huge market for various video content providers. vod customer retention prediction model can accurately predict the number of days users will log in in the future, which is crucial for vod platform to retain customers and increase its core competitiveness in the market. This study proposes an improved RFLH customer retention prediction model based on the traditional RFM model based on a machine learning customer retention prediction research model, and empirically tests the model with real data from the iQIYI platform, constructs the prediction model by three different machine learning algorithms, and applies Optuna to select the optimal hyperparameters of the model to improve the prediction accuracy of the model. The experimental results show that the Vod customer retention prediction model based on machine learning proposed in this study is effective, and LightGBM is the most efficient compared with the other two models with similar accuracy and has high practicality. Finally, the main factors of different features affecting customer retention are analyzed in conjunction with the SHAP model, and the feature variables that have a greater impact on the prediction model are found, and the overall positive and negative relationships between the feature variables and customer retention are carved. This study remedies the shortage of customer retention prediction research in video-on-demand platforms to a certain extent. Meanwhile, the RFLH index system and SHAP explanatory model in the paper can provide better decision support for Vod platform to understand the real situation of users and thus retain them. In this study, the analysis of personal characteristics is derived from the output of the model. In the future, personal characteristics and behaviors can be further combined to explore the influence of users' personal characteristics on customer retention, and more factors affecting customer retention in video-on-demand platforms can be introduced into the prediction model, as well as other prediction techniques can be applied to further improve the accuracy of prediction.

Acknowledgment

This work was supported by the 2021 Harbin University of Commerce Teacher "Innovation" project support project.

References

[1] Gupta, G., & Singharia, K. (2021). Consumption of OTT media streaming in COVID-19 lockdown: Insights from PLS analysis. Vision, 25(1), 36-46.

[2] Wang Y, Guo Y, Chen Y. Accurate and early prediction of user lifespan in an online video-on-demand system[C]//2016 IEEE 13th International Conference on Signal Processing (ICSP). IEEE, 2016: 969-974.

[3] X. Q. Zhang, & Wang X. (2022). Reflection on the profit model innovation of network video platform -- Taking Tencent's advance on demand adding rules as an example. Youth Journalist (4), 2.

[4] Yan, H., Lin, T. H., Gao, C., Li, Y., & Jin, D. (2018). On the understanding of video streaming viewing behaviors across different content providers. IEEE Transactions on Network and Service Management, 15(1), 444-457.

[5] Köster, A., Matt, C., & Hess, T. (2021). Do all roads lead to Rome? Exploring the relationship between social referrals, referral propensity and stickiness to video-on-demand websites. Business & Information Systems Engineering, 63, 349-366.

[6] Wang, W. Y., & Lobato, R. (2019). Chinese video streaming services in the context of global platform studies. Chinese Journal of Communication, 12(3), 356-371.

[7] Rahman, S., Mun, H., Lee, H., Lee, Y., Tornatore, M., & Mukherjee, B. (2018, October). Insights from analysis of video streaming data to improve resource management. In 2018 IEEE 7th International Conference on Cloud Networking (CloudNet) (pp. 1-3). Ieee.

[8] Zhou, Y., Liu, Y., Wang, D., & Liu, X. (2021). Comparison of machine-learning models for predicting short-term building heating load using operational parameters. Energy and Buildings, 253, 111505.

[9] X. Z. Wen & Z. M. Ren.(2023) Predicting the Order Delivery Time of E-commerce Platform Based on the Temporal and Spatial Features of Regional Distribution Center. Operations Research and Management Science., in press.

[10] Fu, B., He, X., Yao, H., Liang, Y., Deng, T., He, H., ... & He, W. (2022). Comparison of RFE-DL and stacking ensemble learning algorithms for classifying mangrove species on UAV multispectral images. International Journal of Applied Earth Observation and Geoinformation, 112, 102890.

[11] Lamrhari, S., El Ghazi, H., Oubrich, M., & El Faker, A. (2022). A social CRM analytic framework for improving customer retention, acquisition, and conversion. Technological Forecasting and Social Change, 174, 121275.

[12] Sudharsan, R., & Ganesh, E. N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science, 34(1), 1855-1876.

[13] Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458-475.

[14] Jiang, P., Zhu, Y., Zhang, Y., & Yuan, Q. (2015, August). Life-stage prediction for product recommendation in e-commerce. In Proceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1879-1888).

[15] Marín Díaz, G., Galán, J. J., & Carrasco, R. A. (2022). XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company. Mathematics, 10(20), 3896.

[16] Perišić, A., & Pahor, M. (2021). RFM-LIR feature framework for churn prediction in the mobile games market. IEEE Transactions on Games, 14(2), 126-137.

[17] Wei Ling, & Xinyue Guo. (2020). Using adapted RFM and GMDH algorithms to predict MOOC user attrition rate. Distance Education in China, (9), 39-43.

[18] Smaili, M. Y., & Hachimi, H. (2023). New RFM-D classification model for improving customer analysis and response prediction. Ain Shams Engineering Journal, 102254.

[19] Jackson, R., & Wang, P. (1994). Strategic database marketing. McGraw Hill Professional.

[20] Wu, Z., Jing, L., Wu, B., & Jin, L. (2022). A PCA-AdaBoost model for E-commerce customer churn prediction. Annals of Operations Research, 1-18

[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3149-3157).

[22] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[23] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[24] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

[25] Gao, S., Xu, J., Dan, W., Li, Q., & Huang, Y. (2021, November). Research on Optimal Control of Fractional Order PI λ D μ Parameters of SCR Denitrification System. In 2021 3rd International Conference on Industrial Artificial Intelligence (IAI) (pp. 1-6). IEEE.

[26] Xiao Qian, Zhipeng Jiao, Yunfei Mu, Wenbiao Lu, &Hongjie Jia. (2021). LightGBM Based Remaining Useful Life Prediction of Electric Vehicle Lithium-Ion Battery under Driving Conditions. Transactions of China Electrotechnical Society, 36(24), 5176-5185.

[27] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[28] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

[29] Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 136, 105405.