World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3079

PREDICTING FOLLOWER GROWTH FOR SOCIAL MEDIA BLOGGERS AND MODELING USER FOLLOWING BEHAVIOR USING XGBOOST AND RANDOM FOREST ENSEMBLE LEARNING

ZhiYang Chen[1*], JiaXin Wu[2]
[1]*School of Mathematics and Physics, Southwest University of Science and Technology, Mianyang 621000, Sichuan, China.*
[2]*School of Law, Southwest University of Science and Technology, Mianyang 621000, Sichuan, China.*
[*]*Corresponding Author: ZhiYang Chen*

Abstract: This study addresses user interaction dynamics on social media platforms by constructing separate models for predicting new follower counts for bloggers and classifying users' targeted following behavior. Quantitative analysis of new follower counts reveals a high correlation between interaction behavior and follower growth, with correlation coefficients exceeding 0.89. By constructing an XGBoost model that utilizes second-order Taylor series expansion to optimize the objective function and incorporates 1- to 3-day lagged features to capture time-series patterns, the model demonstrated robust performance on the test set with an average absolute error of 20.71. For the user targeted following behavior classification prediction, a Random Forest model was selected to address the low-dimensionality and strong nonlinearity of the features. The study employed target encoding strategies to smooth user and blogger IDs, preventing overfitting, and defined interaction intensity formulas based on behavior weights. Experimental results show the Random Forest model achieved an AUC of 0.83 and an F1 score of 74.25% in user follow prediction tasks, effectively enhancing behavioral prediction accuracy. This study provides quantitative insights into user growth dynamics and micro-interaction patterns on social platforms through ensemble learning algorithms.

Keywords: Social media user behavior; XGBoost algorithm; Random forest model

1 INTRODUCTION

In the digital social era, interaction data between users and bloggers holds core operational value for platforms. Accurately predicting bloggers' influence growth trends and users' personalized follow preferences is crucial for optimizing content distribution and enhancing user retention[1-2]. This study focuses on two core tasks: quantitative prediction analysis of new follower counts for influencers and classification prediction of users' targeted following behavior. Previous research addressing these issues often encountered challenges such as complex nonlinear data relationships, sparse user behavior, and volatile time series fluctuations. The innovation in this section lies in verifying the stationarity of influencer growth sequences through ADF tests and extracting deep lag information through feature engineering. For user behavior prediction, we innovatively introduced target encoding techniques using smoothing coefficients to replace raw IDs. By constructing interaction intensity metrics between users and bloggers through behavioral weighting, we significantly enhanced model generalization capabilities. The overall research framework is as follows: First, historical interaction data was aggregated and cleaned, extracting multidimensional features including time, activity levels, and interaction frequency. Second, separate prediction systems were established: a regression model based on XGBoost and a binary classification model based on Random Forest. Finally, model accuracy is systematically evaluated using multiple metrics including MAE, RMSE, and confusion matrices[3].

2 XGBOOST-BASED MODEL FOR PREDICTING NEW FOLLOWERS OF BLOGGERS

2.1 Model Basis

Through the analysis of the features in dataset from https://51mcm.cumt.edu.cn/, we found that the ADF test results show: the test statistic = -20.373, and the p-value is less than 0.05. Taking Blogger B5 as an example, its number of followers fluctuates but has no obvious long-term trend. These indicate that the time series is stationary, and there is no need for complex time series models to handle trends or seasonality. Additionally, the correlation coefficients between interactive behaviors (watching, liking, commenting) and the number of new followers all exceed 0.89, meaning the regression model can effectively capture their correlations[4-5]. Correlation heatmap between the number of followers and interactive behaviors are shown in Figure 1.

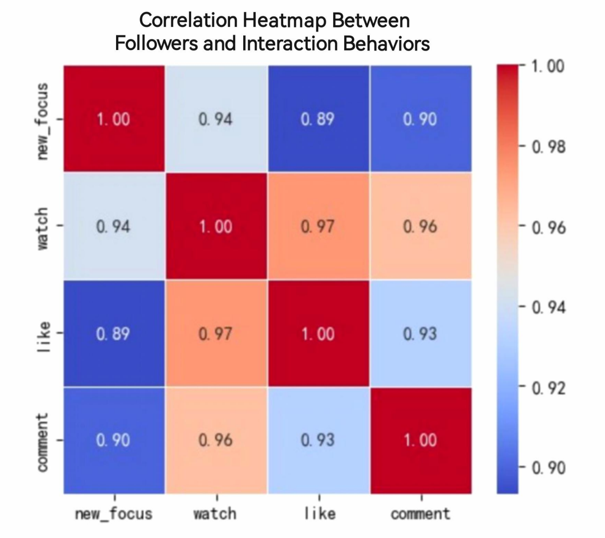

Figure 1 Correlation Heatmap between the Number of Followers and Interactive Behaviors

2.2 Feature Engineering

2.2.1 Aggregating features by blogger and date

Group the data by Blogger ID and date, count the number of different user behaviors (watching, liking, commenting, new following), and convert the original user behavior data into aggregated data at the daily and blogger levels to highlight the user interaction situation of each blogger per day[6-7].

2.2.2 Adding time features and lag features

Calculate "days_since_start" to represent the number of days from the start date of the data to the current date, and "day_of_week" to represent the day of the week for the current date. These time features help the model capture patterns and periodicity in the time series, such as variations in user activity on different days of the week.

Create lag features for the past 1-3 days for each interaction indicator (watching, liking, commenting). These lag features can reflect the continuity and trends of user behaviors, helping the model learn the impact of interaction situations in the past few days on the current number of new followers.

Meanwhile, since some bloggers only have data for one or two days, creating lag features will result in missing values. These rows are deleted to ensure the integrity and accuracy of the data for subsequent model training[8-9].

2.3 Establishment of XGBoost Model

Due to the complex nonlinear relationships between user behaviors (watching, liking, commenting) and the number of followers, we use the splitting rules of the XGBoost tree model for modeling.

2.3.1 Construction of objective function

The objective function of XGBoost consists of a loss function and a regularization term, and the predicted value is obtained by minimizing the objective function:

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t)+\mathrm{constant} \tag{1}$$

Where n is the number of samples, $l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)$ is the loss function that measures the difference between the predicted value and the true value; $\hat{y}_i^{(t-1)}$ is the predicted value of the i-th sample by the model in the $(t-1)$-th iteration, $f_t(x_i)$ is the predicted value of the i-th sample by the t-th decision tree, and $\Omega(f_t)$ is the regularization term used to control the complexity of the model and prevent overfitting.

2.3.2 Taylor series expansion

To simplify the calculation of the objective function, we use the second-order Taylor expansion to approximate the objective function. For a differentiable function g(x), the second-order Taylor expansion at x_0 is:

$$g(x)\approx g(x_0)+g'(x_0)(x-x_0)+\frac{1}{2}g''(x_0)(x-x_0)^2 \tag{2}$$

Expand the loss function $l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)$ in the objective function to the second order around $\hat{y}_i^{(t-1)}$:

$$l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)\approx l\left(y_i,\hat{y}_i^{(t-1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t(x_i)^2 \tag{3}$$

Where $g_i=\partial l(y_i,\hat{y}_i^{(t-1)})/\partial\hat{y}_i^{(t-1)}$ is the first derivative, and $h_i=\partial^2 l(y_i,\hat{y}_i^{(t-1)})/\partial(\hat{y}_i^{(t-1)})^2$ is the Hessian matrix of the second derivative. Ignoring the constant term $l(y_i,\hat{y}_i^{(t-1)})$, the objective function can be approximated as:

$$\mathrm{Obj}^{(t)}\approx\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t(x_i)^2\right]+\Omega(f_t) \tag{4}$$

2.3.3 Tree structure and model integration

In each iteration, a new decision tree $f_t(x)$ is learned. A decision tree can be represented as a mapping q(x) that maps the input sample x to a leaf node of the tree, and each leaf node has a weight w_j. Then $f_t(x)=w_{q(x)}$. Substitute it into the approximate objective function and group the samples according to the leaf nodes they belong to.

The objective function can be rewritten as:

$$\mathrm{Obj}^{(t)}\approx\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j} h_i+\lambda\right)w_j^2\right]+\gamma T \tag{5}$$

Where T is the number of leaf nodes of the tree, and I_j is the set of samples belonging to the j-th leaf node[10].

In each iteration, a new decision tree $f_t(x)$ is trained and added to the previous model:

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+\eta f_t(x_i) \tag{6}$$

Where η is the learning rate, which is used to control the contribution of each tree and prevent overfitting.

By continuously iterating, a series of decision trees are trained, and the final model is the weighted sum of these decision trees:

$$\hat{y}_i=\sum_{t=1}^{M}\eta f_t(x_i) \tag{7}$$

Where M is the number of decision trees.

2.3.4 Model training and validation

Fit the model using the training set data. Use 80% as the training set, and the remaining 20% as the test set. Meanwhile, calculate the MAE and RMSE for the predicted number of new followers on the test set. Daily new follower trend of Blogger B5 is shown in Figure 2.

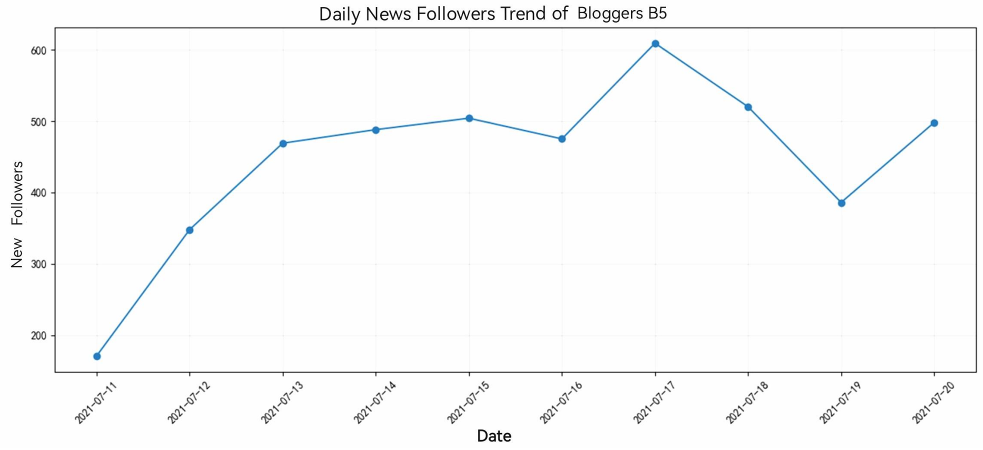

Figure 2 Daily New Follower Trend of Blogger B5

Use MAE to reflect the average absolute deviation between the predicted value and the true value:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{8}$$

Use RMSE, which is more sensitive to large errors, to measure the volatility of the predicted value:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{9}$$

The calculated MAE=20.71 and RMSE=29.85. Scatter plot of predicted values vs. true values is shown in Figure 3.

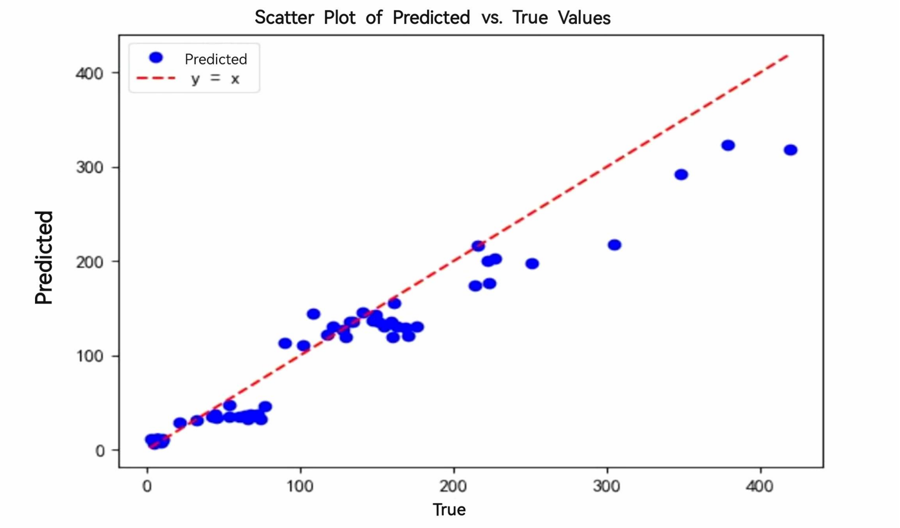

Figure 3 Scatter Plot of Predicted Values vs. True Values

2.4 Model Solution

Table 1 Results

Ranking	1	2	3	4	5
Blogger ID	B21	B5	B15	B60	B13
Number of new followers	438	418	316	244	223

Results are shown in Table 1.

3 RANDOM FOREST-BASED SYSTEM FOR PREDICTING AND RECOMMENDING NEW USER FOLLOWERS

3.1 Model Preparation

First, analyze the dataset. It is noted that although the sample size of dataset is large, the user behavior features only include four aspects: watching, liking, commenting, and following, resulting in a small sample dimension. Therefore, it is not suitable for deep learning models that require a large number of sample features, so RNN and LSTM algorithm models are excluded. Instead, we choose the random forest model, which can handle nonlinear relationships and has strong robustness, to solve and predict this problem.

Second, we need to process the data in dataset to extract the required feature data: 1) the total number of views, likes, comments, and follows for each user; 2) the number of views, likes, comments, and whether they follow each blogger for each user, which are integrated as the training set. Then, based on the known user behaviors in dataset as the test set, we infer the possibility of users having new follows and the corresponding follow objects. Some results are shown in Figures 4 below:

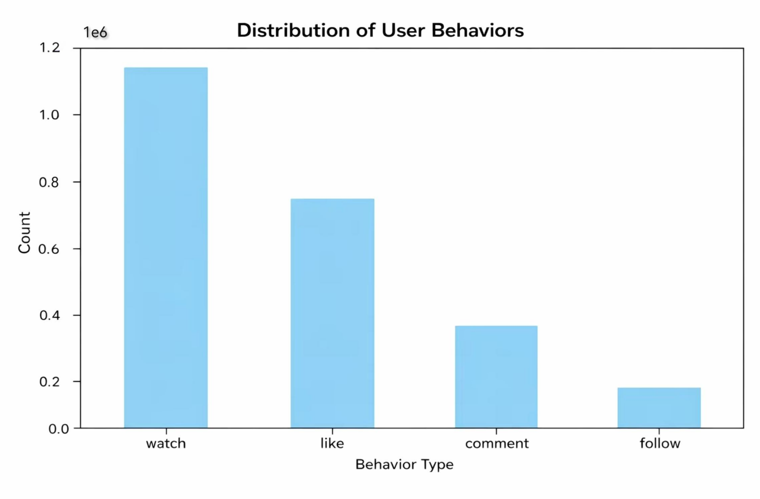

Figure 4 Statistics of Total User Behaviors

Furthermore, conduct in-depth analysis of these data, count user behaviors by the day of the week and hour corresponding to the date, and record the number of daily active and inactive users, as well as the total number of daily follows by users.

3.2 Model Establishment

Based on the random forest model algorithm, first simplify the problem into a binary classification problem, that is, only predict whether the user will have a new follow behavior:

$$y_{u,b,t}=\begin{cases}1 & \text{if user u follows blogger b at time t}\\ 0 & \text{otherwise}\end{cases} \tag{10}$$

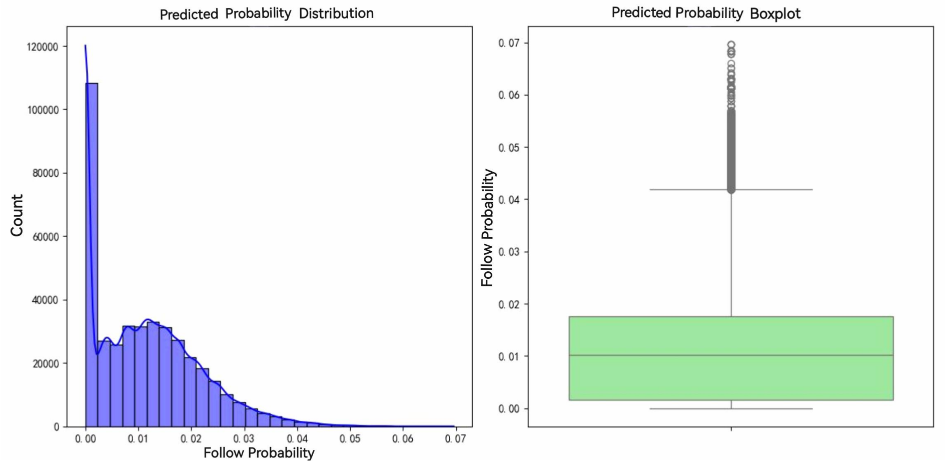

Figure 5 Follow Probability Prediction Chart

Follow probability prediction chart is shown in Figure 5. The model takes user historical behavior data (watching, liking, commenting, following), blogger features, time features, etc., as parameters to calculate the user's follow probability for a single blogger. According to the processed data in the model preparation, the historical follow rate of users/bloggers is used instead of the original ID to avoid overfitting. The specific target encoding is as shown in the following formula (11):

$$TE(u_i)=\lambda(n_i)\cdot\ mean(y|u_i)+(1-\lambda(n_i))\cdot\ mean(y) \tag{11}$$

Where: $\lambda(n_i)=\frac{n_i}{n_i+m}$ (smoothing coefficient, n_i is the number of occurrences of the user, m is a hyperparameter).

Second, set behavior weights according to user behavior features to obtain the interaction strength between users and bloggers:

$$\text{Interaction Strength}_{u,b}=\sum_{k\in K} w_k\cdot\ count_k(u,b) \tag{12}$$

Where:

K: Set of behavior types {watching=1, liking=2, commenting=3, following=4}.

w_k: Weight of behavior k.

$count_k(u,b)$: Number of times user u performs behavior k on blogger b.

The random forest reduces variance and improves generalization ability through voting among multiple decision trees. The random forest model randomly samples with replacement from the total samples to generate sufficient subsamples, trains all generated subsamples, integrates the training results of each subsample, and takes the final result with the highest probability as the training result of the total sample. Therefore, the number of random trees selected will determine the operational efficiency and accuracy of the model. The random forest reduces variance through Bootstrap Aggregating and random feature subsets, while a single decision tree itself is a model with low bias and high variance. As the number of trees increases, the variance of the model will further decrease, but the computational cost will increase. Mathematically, the generalization error of the random forest can be expressed as:

$$Error=Bias^2+Variance+Irreducible\ Noise \tag{13}$$

When the number of trees is sufficiently large, increasing the number of trees will not lead to overfitting. After experimental testing, we found that when the number of trees increases from 0 to 150, the model performance improves significantly; but after exceeding 150, the improvement range becomes smaller. Meanwhile, to improve model efficiency and accuracy, we perform dimensionality enhancement and denoising on the four characteristic behaviors of users based on factors such as user behavior features at different times and the real-time relationship between users and bloggers. User-blogger interaction heatmap is shown in Figure 6.

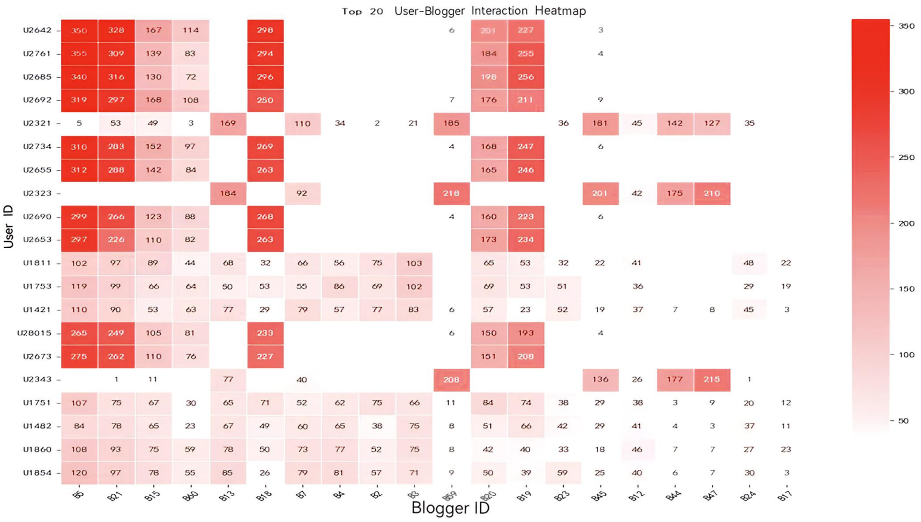

Figure 6 User-Blogger Interaction Heatmap

After determining the number of random trees, formulate the corresponding prediction function to train and predict the decision trees:

$$P(y_{u,b,t}=1|x_{u,b,t})=\frac{1}{T}\sum_{t=1}^{T} h_t(x_{u,b,t}) \tag{14}$$

Where:

$x_{u,b,t}$: Feature vector (user, blogger, interaction features).

T: Number of trees in the random forest.

$h_t(\cdot)$: Prediction function of the t-th decision tree.

After training the decision trees, calculate feature importance. A high importance indicates that the feature has a strong ability to distinguish the target variable:

$$\Delta Gini=Gini_{parent}-\left(\frac{N_{left}}{N}Gini_{left}+\frac{N_{right}}{N}Gini_{right}\right) \tag{15}$$

Where: $Gini=1-\sum_{k=1}^{K} p_k^2$, p_k is the proportion of category k.

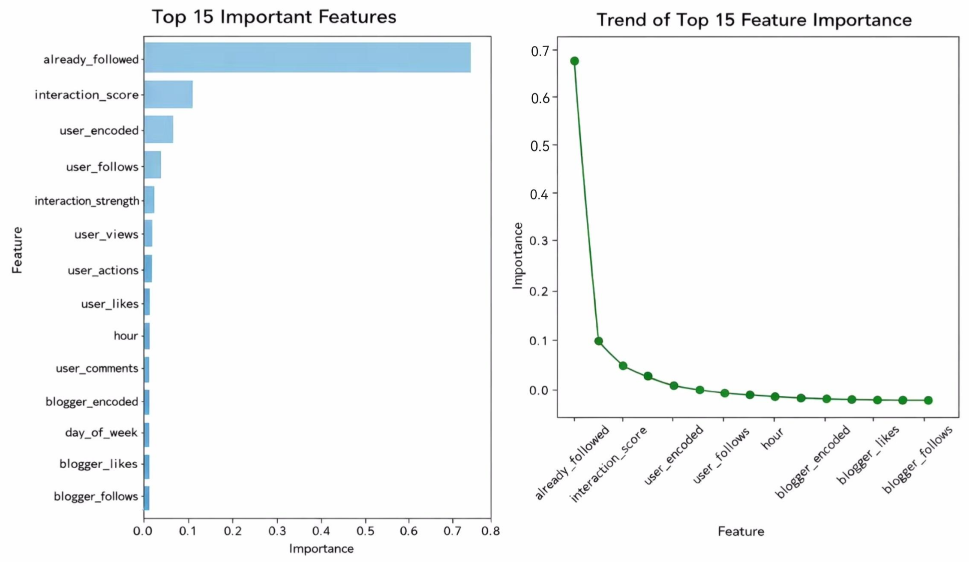

Figure 7 Top 15 Feature Importance

Top 15 feature importance is shown in Figure 7. To analyze and enhance the accuracy of the model, we use 4 core classification indicators: Accuracy, Precision, Recall, and F1-score. We use TP to represent the number of actual follow behaviors correctly predicted by the model, TN to represent the number of actual non-follow behaviors correctly predicted by the model, FP to represent the number of actual non-follow behaviors incorrectly predicted as follows by the model, and FN to represent the number of actual follow behaviors missed by the model to construct a confusion matrix (Table 2):

Table 2 Confusion Matrix

	Predicted not follow	Predicted follow
Actually follow	FN	TP
Actually not follow	TN	FP

Then, Accuracy, Precision, Recall, and F1-score are:

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{16}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{17}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{18}$$

$$\text{F1}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{19}$$

Second, based on the above analysis results, further construct the ROC curve and calculate the AUC score (as shown in Figure 8, AUC = 0.5 indicates random guessing, and >0.8 indicates the model has discriminative ability) [6] to enhance the effectiveness of the model threshold. Among them, the True Positive Rate (TPR) is Recall, and the False Positive Rate (FPR) is calculated as follows:

$$\text{FPR}=\frac{\text{FP}}{\text{FP+TN}} \tag{20}$$

$$\text{AUC}=\int_0^1 \text{TPR}\cdot\ \text{dFPR} \tag{21}$$

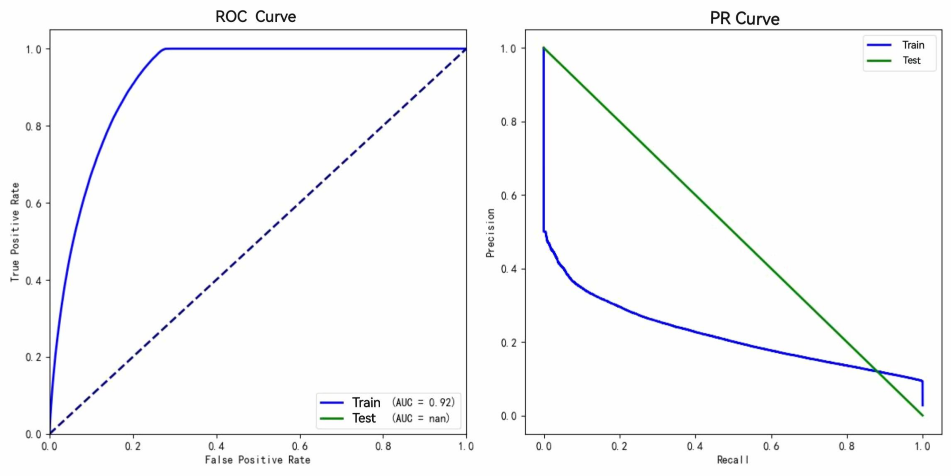

Figure 8 ROC and PR Curves

3.3 Model Solution

The solution results of the new followed blogger IDs of the specified users on July 22, 2024 are shown in Table 3 below:

Table 3 Results

User ID	U7	U6749	U5769	U14990	U52010
New followed blogger IDs	B27/B7/B23	B24/B53/B25	B44/B59/B47	B12/B23/B4	B15/B21/B8

4 CONCLUSIONS

Through deep mining of social platform interaction data, this study successfully establishes two-dimensional behavioral prediction models. In the quantitative prediction of new follower counts for bloggers, the XGBoost algorithm captures nonlinear correlations between viewing, liking, commenting, and new follower acquisition, enabling high-precision estimation of follower growth for specific dates. In the classification prediction of targeted user following behavior, the ensemble learning mechanism of the Random Forest algorithm addressed data sparsity issues, enabling accurate identification of potential follow targets based on users' historical behavior. Although the models demonstrated high accuracy, room for improvement remains. The current Random Forest model relies primarily on manually designed lag features to capture time-series characteristics, failing to fully explore the dynamic patterns of user interest evolution over time. Furthermore, when processing massive samples, the parameter tuning process for random forests is complex and computationally intensive. Future research should focus on incorporating deep learning sequence models like LSTM or Transformer to directly handle temporal streams of user behavior data. Additionally, refining user behavior clustering analysis could enable differentiated parameter configurations for distinct user clusters, thereby enhancing the model's predictive flexibility and accuracy under extremely sparse data conditions.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Su J, Luo X, Li H, et al. Unlocking the Power of User Tie Strength: A Multistudy on Cross-Platform Content Sharing Behaviors. Journal of the Association for Information Systems, 2025, 26(5): 1457-1484. DOI: 10.17705/1JAIS.00942.

[2] Wan S, Yang S, Fu Z. Focus on user micro multi-behavioral states: Time-sensitive User Behavior Conversion Prediction and Multi-view Reinforcement Learning Based Recommendation Approach. Information Processing and Management, 2025, 62(2): 103967-103967. DOI: 10.1016/j.ipm.2024.103967.

[3] Meier Y, Krämer C N. A longitudinal examination of Internet users' privacy protection behaviors in relation to their perceived collective value of privacy and individual privacy concerns. New Media & Society, 2024, 26(10): 5942-5961. DOI: 10.1177/14614448221142799.

[4] Saura R J, Marqués P D, Soriano R D. Privacy concerns in social media UGC communities: Understanding user behavior sentiments in complex networks. Information Systems and e-Business Management, 2023, 23(1): 1-21. DOI: 10.1007/s10257-023-00631-5.

[5] E I X W, Zhang Q. The online website privacy disclosure behavior of users based on concerns-outcomes model. Soft Computing, 2022, 26(21): 11733-11747. DOI: 10.1007/s00500-022-07369-1.

[6] Park M, Chai S. Establish a Model for Detecting Fake News Using Machine Learning; Focusing on User Behaviors and Social Networks. Proceedings of the Korean Society for Intelligent Information Systems Conference, 2022.

[7] Yang Q, Gao L. Research on the Influence of Privacy Concern and Relationship Quality on User Value Co-Creation Behavior of Tourism Platform. Journal of Social Science and Humanities, 2022, 4(4). DOI: 10.53469/jssh.2022.4(04).18.

[8] Meriem A. The Internet user relational orientation: towards a better understanding of online browsing behaviour. Journal of Marketing Management, 2021, 37(13-14): 1374-1408. DOI: 10.1080/0267257x.2021.1917644.

[9] Jennifer G, Jennifer H, Megan P, et al. The mediating effect of meditation and physical activity behaviors on the associations of COVID-19 related worry, attention to news, and stress with mental health in mobile app users in the United States: Cross-sectional survey. JMIR Mental Health, 2021, 8(4): e28479. DOI: 10.2196/28479.

[10] Lauren C, Colette M. A qualitative systematic review of Early Intervention in Psychosis service user perspectives regarding valued aspects of treatment with a focus on cognitive behavioural therapy. The Cognitive Behaviour Therapist, 2021, 14: e31. DOI: 10.1017/s1754470x2100026x.