UDC 004.021–004.85-052

MACHINE LEARNING ALGORITHMS FOR PREDICTING USER BEHAVIOR

Latyshev A.T.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Perevyshko A.I. – Senior Lecturer at the Department of Foreign Languages

Annotation. This text is about machine learning algorithms for predicting user behavior. At the beginning, the author gives definition to user behavior prediction. After that, much attention is given to the mathematical bases, applications and limitations of each method. At the end, the author describes the advantages of using machine learning algorithms for predicting user behavior.

Keywords: user behavior prediction, machine learning, logistic regression, Random Forest, Gradient Boosting, ensemble learning, churn modeling, e-commerce analytics, predictive accuracy.

Introduction. User behavior prediction is the process of analyzing data about someone's actions and preferences to predict their future actions. This area of machine learning plays a crucial role in today's digital world, where personalization is becoming a key factor in the interaction between companies and their customers. Using machine learning algorithms optimizes resources and improves service quality. The use of such algorithms is widespread in e-commerce, social media, advertising and many other areas [1].

Main part. Many machine learning algorithms are used to predict user behavior. The main ones for predicting user behavior are Logistic Regression, Random Forrest and Gradient Boosting.

Logistic regression is a statistical technique used to model binary dependent variables. It is one of the simplest and most interpretable machine learning algorithms used for the classification of problems where the result takes two possible values such as yes or no, true or false. Logistic regression is based on linear regression, but applies a non-linear function, the sigmoid function, to predict probabilities. These probabilities are calculated using formula 1.

$$P(X) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j\right)}}, \quad (1)$$

where Y – a binary target variable (1 - action done, 0 - not done);

$X = (X_1, X_2, ..., X_p)$ – vector of user behavioral attributes;

β_0 – intersept (base probability);

β_j – weight of the j-th behavioral factor.

The model works on the principle of estimating the relationship between various behavioral parameters (time on site, number of clicks, etc.) and the probability of a target event. For example, it can show how an increase in product views affects the chances of a purchase.

Logistic regression is used to predict the likelihood of a user making a purchase based on his behavior on the site, such as session time, number of pages viewed and items added to cart. This method is also used to predict customer churn by analyzing the decrease in customer activity, frequency of service usage and history of interaction with support.

Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. More formally, for a k-dimensional random vector $X = (X_1,...,X_k)^T$ representing the real-valued input or predictor variables and a random variable Y representing the real-valued response [4].

This algorithm is particularly effective for predicting user behavior due to its robustness to overfitting and ability to handle high-dimensional data. The prediction is made by aggregating the

results of numerous decision trees, each trained on a random subset of features and data samples, as shown in formula 2:

$$Y = mode(\{T_k(X)\}_{k=1}^{K}), \quad (2)$$

where Y – predicted user action;

T_k – k-th decision tree in the ensemble;

X = (X_1, X_2,...,X_p) – vector of behavioral features;

K – total number of trees.

Random Forest analyzes hundreds of behavioral features, such as time spent per content type, interaction frequency and device usage patterns to predict which products or content a user is most likely to engage with, achieving recommendation accuracy of 85-92% on e-commerce platforms.

By evaluating complex feature interactions, such as transaction speed, location changes, device fingerprints, the model identifies anomalous behavior patterns with 30% higher precision than single decision trees, reducing false positives in payment systems. Gradient Boosting is a powerful machine learning technique that builds models sequentially, where each new model corrects the errors of previous ones, optimizing an arbitrary differentiable loss function. The final prediction is a sum of all weak learners (typically decision trees), as expressed in formula 3:

$$F(X) = \sum_{m=1}^{M} \gamma_m h_m(X), \quad (3)$$

where F(X) – final prediction score;

h_m – m-th weak learner (tree);

γ_m – learning rate for the m-th tree;

M – total number of boosting iterations.

XGBoost and LightGBM are modern implementations of gradient boosting optimized for big data and time series. XGBoost processes temporal behavioral data, including login frequency decay, support ticket patterns, feature usage trends, to predict churn risk 48 hours in advance, enabling proactive retention campaigns [3]. LightGBM models analyze real-time user engagement, such as click heatmaps, price sensitivity tests, cart abandonment history, to predict optimal discount levels that maximize conversion probability while preserving margin, boosting revenue by 12-18% in A/B tests.

Conclusion. Logistic regression does not handle nonlinear dependencies well. Random Forest is robust to overfitting, but less interpretable. It automatically selects important features and performs well with nonlinear data. Gradient Boosting usually shows the best accuracy by consistently correcting errors of previous trees, but requires careful tuning and more computational resources. The choice of a method depends on the problem which we want to solve logistic regression is suitable for a fast baseline solution, Random Forest for stable predictions without fine-tuning, and Boosting for maximizing accuracy. Machine learning algorithms for predicting user behavior open a wide range of opportunities for various industries, including e-commerce, medicine, social media and finances. Ultimately, machine learning is becoming not just a tool, but a key element in the development strategy of modern organizations [2, 5, 6].

References

1. *Breiman, L. Random Forest / L. Breiman // Machine Learning. — 2001. — Vol. 45, № 1. — Pp. 5–32.*

2. *Hosmer, D. W. Applied Logistic Regression / D. W. Hosmer, S. Lemeshow. — 3rd ed. — 2013. — 528 p.*

3. *Chen, T. XGBoost: A Scalable Tree Boosting System / T. Chen, C. Guestrin // Proc. of the 22nd ACM SIGKDD. — 2016. — Pp. 785–794.*

4. *Adele Cutler, D. Richard Cutler and John R. Stevens Random Forests // Ensemble Machine Learning: Methods and Applications – 2011.- Pp. 2-4.*

5. *Ofori, F., Maina, E. & Gitonga, R. (2020). Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review, Journal of Information and Technology, Vol. 4(1), 23-45*

6. *Daniyal M. Alghazzawi, Anser Ghazal Ali Alquraishee, Sahar K. Bardi, Syed Hamid Hasan ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. – 2023.*