

基于WOA-XGBoost模型对电商用户流失预测研究

左 恣

南京信息工程大学数学与统计学院, 江苏 南京

收稿日期: 2025年6月4日; 录用日期: 2025年6月23日; 发布日期: 2025年7月23日

摘 要

在电商竞争日益激烈的当今时代, 用户流失问题对企业发展影响重大, 防止客户流失是任何企业都需要重点解决的一个问题, 因此预测流失用户一直是学术界研究的热点。本文基于抖音电商用户相关信息数据集, 分别构建了决策树、随机森林、XGBoost模型对用户是否流失进行预测, 测试集上的预测结果表明XGBoost模型在AUC值、召回率、F1指标效果最优。随后, 鉴于XGBoost模型的性能易受参数影响, 为进一步提升其预测精度, 引入鲸鱼优化算法(WOA)对XGBoost的关键参数, 如学习率、最大树深度、子采样率等进行全局寻优。在优化过程中, 采用五折交叉验证法确保模型的泛化能力, 避免过拟合现象。实验结果显示, 经WOA优化后的XGBoost模型, 相比原模型, 在AUC指标上提升了1.20%, 准确率提升1.05%, F1值增长了0.97%, 在电商用户流失预测任务中展现出更强的预测能力和更高的可靠性, 为电商平台精准识别流失用户、制定有效挽留策略提供了更优的技术方案。

关键词

用户流失预测, XGBoost, 鲸鱼优化算法, 支持向量机, 决策树

A Study on E-Commerce User Churn Prediction Based on the WOA-XGBoost Model

Min Zuo

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Jun. 4th, 2025; accepted: Jun. 23rd, 2025; published: Jul. 23rd, 2025

文章引用: 左恣. 基于 WOA-XGBoost 模型对电商用户流失预测研究[J]. 电子商务评论, 2025, 14(7): 1892-1901.
DOI: 10.12677/eci.2025.1472384

Abstract

In the current era of increasingly fierce e-commerce competition, the problem of user churn has a significant impact on the development of enterprises. Preventing customer churn is a key issue that any enterprise needs to address. Therefore, predicting churn users has always been a hot topic in academic research. This paper uses the data set of TikTok e-commerce user-related information to construct decision tree, random forest, and XGBoost models respectively to predict whether users will churn. The prediction results on the test set show that the XGBoost model has the best performance in terms of AUC value, recall rate, and F1-measure. Subsequently, considering that the performance of the XGBoost model is vulnerable to parameter influences, in order to further improve its prediction accuracy, the Whale Optimization Algorithm (WOA) is introduced to globally optimize the key parameters of XGBoost, such as learning rate, maximum tree depth, subsample rate, etc. During the optimization process, the five-fold cross-validation method is adopted to ensure the generalization ability of the model and avoid overfitting. The experimental results show that, compared with the original model, the XGBoost model optimized by WOA has an increase of 1.20% in the AUC index, a 1.05% increase in accuracy, and a 0.97% enhancement in F1-value. It demonstrates stronger predictive ability and higher reliability in the task of e-commerce user churn prediction, providing a better technical solution for e-commerce platforms to accurately identify churn users and formulate effective retention strategies.

Keywords

User Churn Prediction, XGBoost, Whale Optimization Algorithm, Support Vector Machine, Decision Tree

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着电子商务的迅猛发展，用户增长红利逐渐趋缓，行业竞争进入存量博弈阶段。据中国互联网络信息中心统计，截至 2023 年，我国电商用户规模增速已连续三年低于 5% [1]，如何有效维系现有用户、降低流失率成为企业可持续发展的核心挑战。用户流失不仅直接导致营收下滑，还会增加新客获取成本 [2]，因此精准识别潜在流失用户并实施干预策略具有显著的经济价值。

2. 相关研究

近年来，机器学习技术为用户流失预测提供了新思路。电信与金融行业因具备丰富数据和稳定业务流程，成为研究重点领域。早期探索中，朱世武等[3]在电信行业应用决策树模型分析用户流失，发现性别等个体属性具有显著影响，为后续模型构建提供了实证参考。Verbeke 等人[4]基于银行客户数据，系统比较了决策树、逻辑回归、支持向量机、集成学习等方法，指出随机森林与梯度提升树在处理高维、非线性特征时表现优异，且具备较强泛化能力。同时，作者提出了基于规则归纳的解释机制，提升模型可解释性，增强了企业实际应用的可行性。关于用户流失领域，Lemmens 与 Croux [5]将 Bagging 和 Boosting 策略应用于分类树，发现其在缓解样本不平衡问题、提高模型稳定性方面成效显著，尤其适用于高风险流失用户的早期识别。此外，Huang 等人[6]提出了一种结合深度学习与注意力机制的客户流失预测框架，

可自动识别关键特征,提升模型的适应性与精准度。Nguyen & Mutum [7]则强调用户满意度与品牌忠诚度在流失预测中的中介作用,提出应结合行为数据与情感因素构建更全面的预测体系。随着模型能力的提升,研究者逐步将社交行为、消费路径等非结构化或高维信息纳入分析框架。王重仁与韩冬梅[8]创新地将社交网络因素融入 XGBoost 模型,发现用户互动频率、社群参与度等变量显著提升模型预测性能,验证了多源异质数据在刻画用户行为模式中的关键作用。

在电商领域,特别是像抖音电商这样的新兴平台,用户行为数据呈现出爆炸式增长,且具有高度的复杂性和多样性。传统的模型可能难以有效捕捉这些数据的特征,而集成学习和深度学习模型则展现出更大的潜力。然而,这些模型通常包含大量的参数,需要精心调优才能达到最佳性能。针对这点,本文以抖音电商用户相关信息数据集为研究对象,构建了多种机器学习模型,对用户是否流失进行预测。考虑到 XGBoost 模型的预测效果在很大程度上依赖于其超参数设置,本文进一步引入鲸鱼优化算法对 XGBoost 的关键参数调优。优化后的模型在准确性与稳定性方面均得到显著提升,验证了智能优化算法在提升流失预测模型性能方面的重要作用。

3. 模型建立与评价指标

3.1. XGBoost 算法原理

XGBoost 是极端梯度上升的决策树,用于分类、回归等任务,以高效、可扩展性著称。

对于第 t 棵树,第 t 个样本,其输入的特征向量可表示为 x_i ,对应的预测值 $\hat{y}$ 可通过式(1)进行计算。

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (1)$$

式子中 $f_t(x_i)$ 表示第 t 棵树的预测模型;则 $\hat{y}_i^{(t-1)}$ 表示第 $t-1$ 树的预测结果。

XGBoost 算法的目标函数为:

$$\text{obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^t \Omega(f_k) \quad (2)$$

式子中: obj 表示目标函数; $l(y_i, \hat{y}_i)$ 则是损失函数,值越小函数的精度越高; $\hat{y}$ 是预测值, y 则是真实值, Ω 是正则化项

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (3)$$

式子中: T 表示每个叶子节点的数目, ω 表示第叶节点的分数, λ 和 γ 则是正则化系数控制权重大小以及叶子数量。

通过对损失函数进一步的优化推导,得到如下形式的目标函数:

$$\text{obj}^{(t)} = \sum_{i=1}^n \left(l(y_i, \hat{y}_i^{(t-1)}) + f_t(x_i) \right) + \sum_{k=1}^t \Omega(f_k) \quad (4)$$

对目标函数运用二阶泰勒展开式进行近似处理,可推导出(5)式

$$\text{obj}^{(t)} \approx \sum_{i=1}^n \left(l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \sum_{k=1}^t \Omega(f_k) \quad (5)$$

$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ 是损失对 $\hat{y}_i^{(t-1)}$ 一阶导, $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ 是损失对 $\hat{y}_i^{(t-1)}$ 二阶导。

叶子结点 j 中所包含样本的一阶导总和可表示为:

$$G_i = \sum_{i \in I_j} g_i \quad (6)$$

叶子结点 j 中所包含样本的二阶导总和可表示为:

$$H_i = \sum_{i \in I_j} h_i \quad (7)$$

最优化函数为:

$$\text{obj}^{(t)} = \sum_{j=1}^T \left(G_j \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right) + \gamma T \quad (8)$$

运用目标函数, 对 ω_j 求导可得:

$$\omega_j = -\frac{G_j}{H_j + \lambda} \quad (9)$$

3.2. WOA

鲸鱼优化算法是一种模拟座头鲸捕食行为的群体智能优化算法。灵感来源于座头鲸的泡泡网狩猎行为。它的核心思想基于座头鲸的气泡网捕食策略, 通过模拟鲸鱼围捕猎物的行为来搜索最优解。WOA 主要包含以下三种行为机制:

3.2.1. 收缩包围圈

位置更新公式:

$$\mathbf{X}(t+1) = \mathbf{X}^*(t) - \mathbf{A} \cdot \mathbf{D} \quad (10)$$

其中: t 为当前迭代次数; $\mathbf{X}^*(t)$ 为当前最优解的位置向量;

$\mathbf{A}$ 与 $\mathbf{D}$ 为系数向量, 计算公式为:

$$\mathbf{A} = 2\mathbf{a} \cdot \mathbf{r}_1 - \mathbf{a}, \quad \mathbf{D} = |\mathbf{C} \cdot \mathbf{X}^*(t) - \mathbf{X}(t)| \quad (11)$$

式中: $\mathbf{D}' = |\mathbf{X}^*(t) - \mathbf{X}(t)|$, 表示当前位置与最优解的距离; $\mathbf{a}$ 控制收敛速度; $\mathbf{r}_1$ 为 $[0, 1]$ 内的随机向量。

3.2.2. 螺旋泡泡网攻击

座头鲸在逼近猎物时, 会沿螺旋路径上升, 同时收缩包围圈。

螺旋位置更新公式:

$$\mathbf{X}(t+1) = \mathbf{D}' \cdot e^{bl} \cdot \cos(2\pi l) + \mathbf{X}^*(t) \quad (12)$$

其中: b 为螺旋形状的常数参数, 控制螺旋的紧密程度; l $[-1, 1]$ 内的随机数, 决定螺旋的旋转方向和幅度。

3.2.3. 随机搜索猎物

当座头鲸未确定猎物位置时, 会随机选择一个当前最优解附近的位置进行搜索。

位置更新公式:

$$\mathbf{X}(t+1) = \mathbf{X}_r(t) - \mathbf{A} \cdot \mathbf{D}_r \quad (13)$$

其中: $\mathbf{X}_r(t)$ 为从当前种群中随机选择的一个位置; $\mathbf{D}_r = |\mathbf{C} \cdot \mathbf{X}_r(t) - \mathbf{X}(t)|$, 计算当前位置与随机个体的距离。当 $|\mathbf{A}| > 1$ 时, 算法倾向于全局搜索; 当 $|\mathbf{A}| < 1$ 时, 倾向于局部开发。

3.3. 模型评价指标

用户流失预测属于典型的二分类任务，模型的效果可借助混淆矩阵进行评估，其具体结构如表 1 所示。

Table 1. Confusion matrix table

表 1. 混淆矩阵表

	预测为正类	预测为负类
实际是正类	TP	FN
实际是负类	FP	TN

基于混淆矩阵能够算出准确率(Accuracy)、F1-score 以此来评判模型的优劣，下面为是这些指标的计算方法。

Recall 与 Precision 计算方式如下所示：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (14)$$

$$\text{F1} = \frac{2 * \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

式中 Recall 与 Precision 的计算公式为

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (16)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (17)$$

同时，可以利用混淆矩阵计算出假正率(FPR)和真正率(TPR)，并据此绘制 ROC 曲线。模型分类性能的优劣可通过曲线下方的 AUC (面积)进行评估。FPR 和 TPR 的计算公式如下所示。

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (18)$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (19)$$

4. 实验分析

4.1. 数据的来源以及预处理

本文选用的数据来源于阿里云公开数据中直播电商数据集，其中包含 5563 条用户相关信息，共 17 个预测用户流失相关指标，具体变量解释如表 2。

Table 2. Variable explanation table

表 2. 变量解释表

原始数据字段名	变量解释
Churn	用户是否流失
Tenure	用户在平台注册的时间
PreferredLoginDevice	用户常用的登录设备

续表

CityTier	用户所在城市的等级
WarehouseToHome	仓库到用户家的距离
MaritalStatus	婚姻状态
AgeGroup	年龄段
Gender	性别
HourSpendOnApp	用户在 App 上的平均使用时长
OrderAmountHikeFromlastYear	与去年相比，订单金额的增长幅度
DaySinceLastOrder	用户自上次下单以来已过去的天数
PreferedOrderCat	用户偏好的订单类别
NumberOfStreamerFollowed	用户关注的直播带货主播数量
SatisfactionScore	用户满意度评分
Complain	用户是否曾投诉
OrderCount	用户累计下的订单数量
CouponUsed	优惠券使用

对变量缺失值进行剔除后剩下 3775 条数据，用户流失比例如图 1 所示，未流失用户数 3143 位，用户流失数 631 位，存在样本量不平衡。基于样本不平衡问题，通过 SMOTE 方法进行处理。具体而言，在将原始数据集划分为训练集和测试集之后(7:3)，使用 SMOTE 对训练集进行过采样，人工合成少数类样本，从而使训练集中各类别样本数量更加平衡。

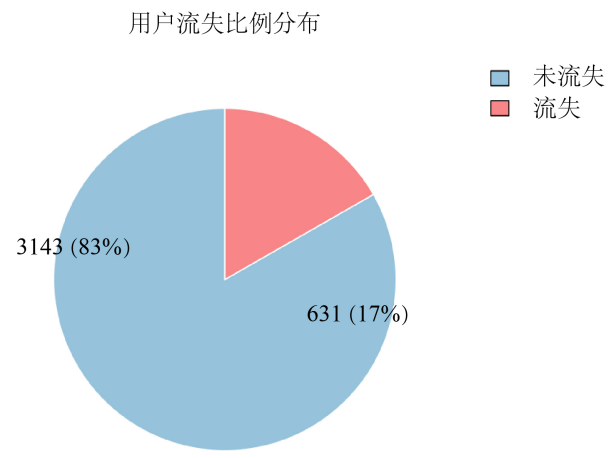


Figure 1. Distribution map of user churn situation
图 1. 用户流失情况分布图

为了探究各变量与用户流失之间的关系，本文绘制了变量相关性热力图。从图 2 可以看出，部分变量与用户流失具有较强的线性相关性，其中在平台注册时长与用户流失呈显著负相关($r = -0.33$)，表明用户使用时间越长，流失可能性越低；投诉次数与用户流失呈正相关($r = 0.25$)，提示投诉多的用户更易流失；此外，满意度得分也与用户流失负相关($r = -0.20$)，说明满意度较高的用户更可能留存。与此同时，相关性分析还发现部分特征之间存在较强的共线性，例如用户累计下的订单数量与优惠券使用的相关系

数高达 0.75，提示二者可能存在信息冗余。因此，在后续建模中，为避免多重共线性带来的干扰，本文优先保留与用户流失相关性较高的变量，并对高度相关的变量对进行合并或删除，从而提升模型稳定性与解释性。

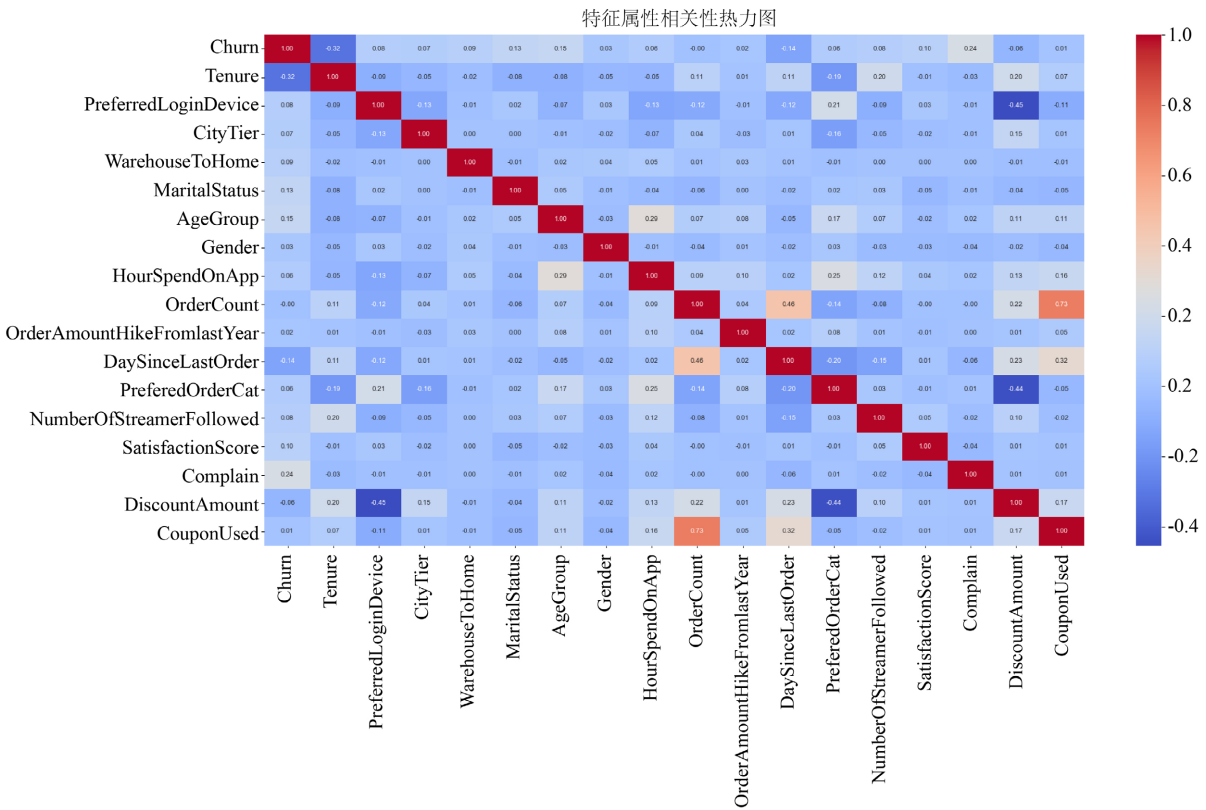


Figure 2. Correlation heatmap of feature attributes
图 2. 特征属性相关性热力图

4.2. 实验结果分析

为了预测用户留存情况，本文分别构建了决策树(Decision Tree)、支持向量机(Support Vector Machine, SVM)和极端梯度提升(XGBoost)三种模型。三模型预测效果评价指标如图 3，图 4 以及表 3 所示。

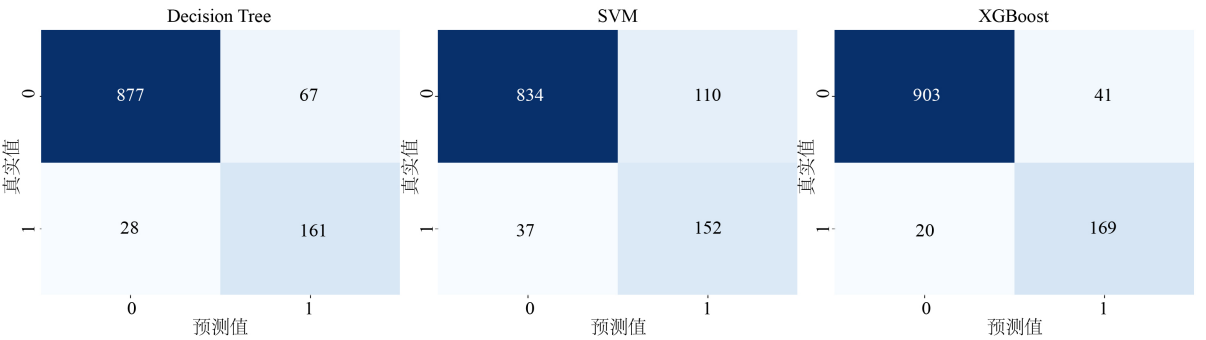


Figure 3. Confusion matrix diagram of three models
图 3. 三大模型混淆矩阵图

Table 3. Comparison of indicators of three models
表 3. 三大模型指标对比

	召回率	准确率	F1	AUC 值
决策树	0.8519	0.9162	0.7722	0.8904
SVM	0.8042	0.8703	0.6741	0.9266
XGBoost	0.8942	0.9462	0.8471	0.9683

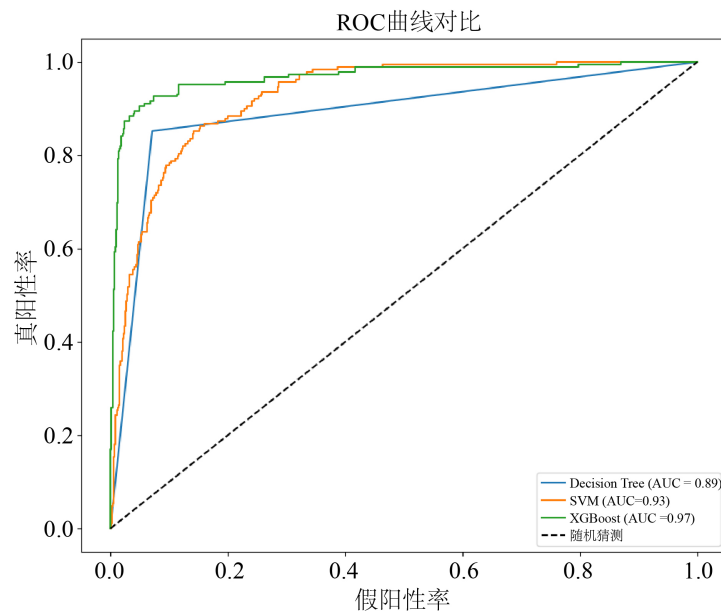


Figure 4. ROC curve diagram of three models
图 4. 三大模型 ROC 曲线图

通过对比各模型在测试集上的混淆矩阵、准确率、召回率、F1 值等指标，并结合 ROC 曲线下的 AUC 值进行综合评估，结果显示 XGBoost 模型的整体预测效果最优。

基于此，进一步引入鲸鱼优化算法对 XGBoost 模型的超参数进行调优，以提升模型性能。在模型参数优化过程中，本研究采用鲸鱼优化算法对 XGBoost 模型的五个核心超参数进行调优，分别为 $n_estimators$ (范围 100~300)、 max_depth (范围 3~7)、 $gamma$ (范围 0~0.5)、 $subsample$ (范围 0.7~1.0) 和 min_child_weight (范围 1~3)，优化过程中设置鲸鱼种群数量为 30，最大迭代次数为 300。最终 WOA 寻找的最佳参数为 $n_estimators$: 221、 max_depth : 7、 $gamma$: 0.056、 $subsample$: 0.868、 min_child_weight : 3、 $learning_rate$: 0.1。

根据图 5，表 4 可知优化后的模型相较于未优化的 XGBoost 模型在各项性能指标上均有提升，表明鲸鱼优化算法有效地提高了模型的分类能力。

图 6 的 SHAP 分析结果揭示了 WOA-XGBoost 模型中各特征对客户流失预测的贡献度。在网时长 (Tenure) 是客户流失预测的核心驱动因素：低值用户 (蓝色) 的 SHAP 值显著右偏，预示高流失风险 (如新客户)，而高值用户 (红色) 强烈抑制流失，体现用户黏性的关键作用。关注主播数 (NumberOfStreamerFollowed) 和折扣力度 (DiscountAmount) 的提升显著降低流失概率，而投诉记录 (Complain) 则成为最强风险信号 (SHAP 正向偏移)。消费活跃度指标如订单总数 (OrderCount) 增加、最近下单间隔 (DaySinceLastOrder) 缩短均与流失负相关，而月费用提升 (MonthlyCharges) 在部分用户中触发流失敏感度。应用使用时长

(HourSpendOnApp)和仓库距离(WarehouseToHome)的影响呈现非线性：适度使用降低风险，但过度使用可能预示不满；距离增加则线性推高流失率。值得注意的是，人口属性(年龄、性别、婚姻状态)位于重要性底部且 SHAP 值聚集于零附近，表明业务行为数据比静态属性更具预测力。

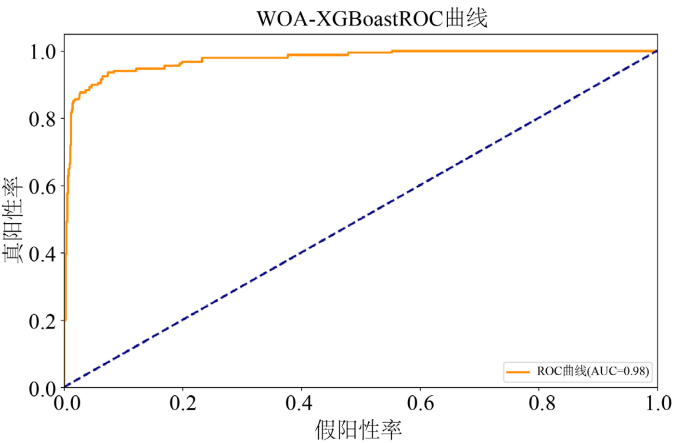


Figure 5. ROC curve diagram of WOA-XGBoost
图 5. WOA-XGBoost ROC 曲线图

Table 4. Comparison chart of optimized model indicators
表 4. 优化后模型指标对比图

	召回率	准确率	F1	AUC 值
WOA-XGBoost	0.8991	0.9561	0.8553	0.9799
XGBoost	0.8942	0.9462	0.8471	0.9683
指标提升	0.55%	1.05%	0.968%	1.20%

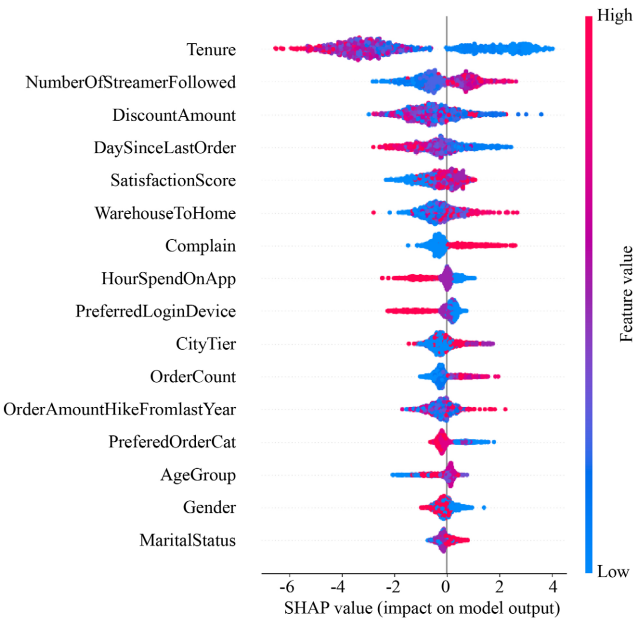


Figure 6. WOA-XGBoost SHAP value analysis
图 6. WOA-XGBoost SHAP 值分析

5. 结论与展望

本文基于抖音电商用户行为数据,构建了决策树、随机森林和 XGBoost 模型对用户流失进行预测,结果表明 XGBoost 模型在多项评估指标上表现优越。进一步地,为克服 XGBoost 模型在超参数选择方面的局限性,本文引入鲸鱼优化算法(WOA)对其关键参数进行全局优化,并采用五折交叉验证提升模型的稳定性和泛化能力。实验结果表明,WOA-XGBoost 模型在 AUC 值、召回率及 F1 值等指标上均优于原始模型,显著提升了流失用户的识别能力。

未来研究可在以下几个方面展开:一是进一步引入更多维度的用户行为特征,如情感倾向、社交互动等,以丰富模型的输入变量;二是尝试结合深度学习模型或集成多模型框架,以进一步提升预测性能;通过持续优化算法与挖掘更深层次的用户行为模式,有望为电商平台制定更精准的用户挽留策略提供坚实的数据支撑和技术保障。

参考文献

- [1] 中国互联网络信息中心(CNNIC). 中国互联网络发展状况统计报告[EB/OL]. 北京: 中国互联网络信息中心, 2023. <https://www.cnnic.cn/NMediaFile/2023/0908/MAIN1694151810549M3LV0UWOAV.pdf>, 2025-07-17.
- [2] Reichheld, F.F. and Sasser, W.E. (1990) Zero Defections: Quality Comes to Services. *Harvard Business Review*, **68**, 105-111.
- [3] 朱世武, 崔嵬, 谢邦昌. 移动电话客户流失数据挖掘[J]. 数理统计与管理, 2005, 24(1): 62-68.
- [4] Verbeke, W., Martens, D., Mues, C. and Baesens, B. (2012) Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques. *Expert Systems with Applications*, **38**, 2354-2364. <https://doi.org/10.1016/j.eswa.2010.08.023>
- [5] Lemmens, A. and Croux, C. (2006) Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, **43**, 276-286. <https://doi.org/10.1509/jmkr.43.2.276>
- [6] Huang, C., Ke, S. and Tsai, C. (2019) A Deep Learning Model for Customer Churn Prediction Based on Attention Mechanism. *Expert Systems with Applications*, **129**, 93-103.
- [7] Nguyen, B. and Mutum, D.S. (2012) A Review of Customer Churn in the Mobile Telecommunications Industry. *Marketing Review*, **12**, 327-351.
- [8] 王重仁, 韩冬梅. 基于社交网络分析和 XGBoost 算法的互联网客户流失预测研究[J]. 信息技术与网络安全, 2017, 36(23): 58-61.