Computer Science and Application 计算机科学与应用, 2026, 16(6), 313-320
Published Online June 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.166230

Hans汉斯

机器学习赋能保险反欺诈识别：基于随机森林模型的实证研究

刘　姿

中央财经大学保险学院，北京

收稿日期：2026年5月18日；录用日期：2026年6月22日；发布日期：2026年6月30日

摘　要

保险欺诈识别是保险公司风险管理与数字化转型进程中无法回避的核心议题。本文以阿里天池公开的金融数据分析保险反欺诈预测数据集为样本，构建“数据预处理－探索性分析－特征工程－模型比较－业务转化”的一体化研究框架，系统比较随机森林与XGBoost模型在欺诈识别任务中的表现。实证结果表明，训练样本中欺诈比例为25.86%，高额索赔、事故严重程度、投保至出险时间间隔及保费相关变量均具有显著的识别价值。经参数调优与阈值校准后，随机森林模型的AUC约为0.723，召回率为0.691，能够在较高的欺诈捕获能力与可接受的误报水平之间取得务实平衡。进一步基于模型输出，本文提出涵盖核保审核、理赔调查与风险定价的三级风险分层机制，为业务落地提供可操作路径。研究表明，机器学习模型可有效充当保险反欺诈中的风险排序与资源配置工具，但人工复核、样本扩充与持续反馈机制仍是其不可或缺的实务补充。

关键词

保险反欺诈，随机森林，XGBoost，机器学习，风险分层

Machine Learning-Enabled Insurance Fraud Detection: An Empirical Study Based on a Random Forest Model

Zi Liu

School of Insurance, Central University of Finance and Economics, Beijing

Received: May 18, 2026; accepted: June 22, 2026; published: June 30, 2026

文章引用: 刘姿. 机器学习赋能保险反欺诈识别：基于随机森林模型的实证研究[J]. 计算机科学与应用, 2026, 16(6): 313-320. DOI: 10.12677/csa.2026.166230

Abstract

Insurance fraud detection sits at the heart of contemporary risk management and digital transformation in the insurance industry. Drawing on the publicly available insurance anti-fraud prediction dataset from Ali Tianchi, this study develops an integrated analytical framework spanning data preprocessing, exploratory analysis, feature engineering, model comparison, and business implementation. Random Forest and XGBoost classifiers are compared for fraud identification. The empirical results show that fraudulent observations account for 25.86% of the training sample, and that high claim amounts, accident severity, the interval between policy binding and incident occurrence, and premium-related variables carry substantial predictive power. After parameter tuning and threshold calibration, the Random Forest model attains an AUC of approximately 0.723 and a recall of 0.691, striking a workable balance between fraud capture and the false-positive burden. Building on the model output, the paper proposes a three-tier risk-control mechanism that supports underwriting review, claim investigation, and risk-based pricing. The findings indicate that machine learning is well-suited to serve as a risk-ranking and resource-allocation instrument in insurance anti-fraud practice, with human review and continuous feedback remaining indispensable complements.

Keywords

Insurance Fraud Detection, Random Forest, XGBoost, Machine Learning, Risk Stratification

Copyright © 2026 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

保险欺诈已成为保险业风险治理与数字化转型进程中无法回避的核心问题。随着保险产品线上化、理赔流程自动化以及客户数据维度的持续扩展，传统依赖人工经验与固定规则的反欺诈方式逐渐暴露出识别滞后、误判率较高、调查资源分散等结构性不足[1] [2]。尤其在车险等高频理赔场景中，欺诈行为常表现为金额夸大、事故类型伪装、投保后短期出险等复合特征，单一规则体系难以覆盖此类复杂模式。

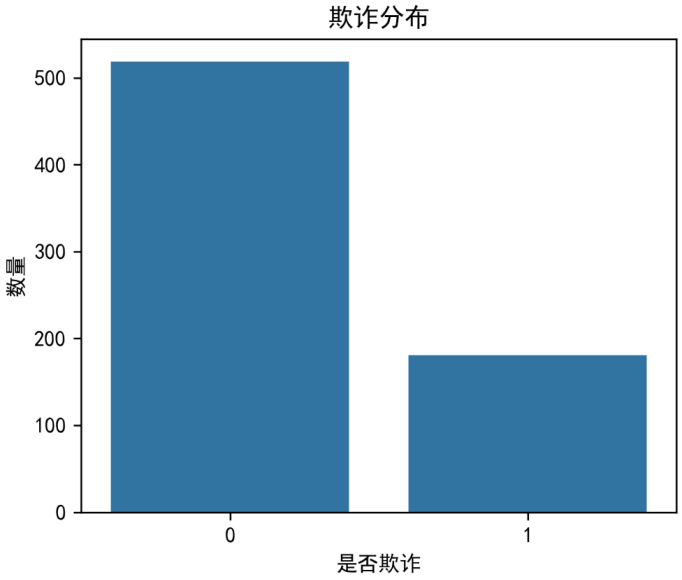
基于上述背景，本文以阿里天池公开的金融数据分析保险反欺诈预测数据集为样本，构建“数据预处理–探索性分析–特征工程–模型比较–业务转化”的研究框架。与单纯追求预测精度的技术报告不同，本文同时关注模型在保险业务中的可解释性与可落地性：一方面，通过随机森林与 XGBoost 模型的横向比较，评估机器学习方法在欺诈识别中的有效性(见图 1)；另一方面，结合风险分层、核保流程与理赔调查策略，提出可嵌入保险公司风险管控体系的应用方案。

2. 数据来源与研究设计

本文采用阿里天池学习赛“金融数据分析保险反欺诈预测”数据集[3]。数据集由 train.csv 与 test.csv 两个文件构成，涵盖客户基本信息、保险政策信息、事故信息、索赔信息及欺诈标签等字段，共含 38 个原始特征。其中训练集包含 700 条样本，测试集包含 300 条样本。训练集中非欺诈样本 519 条，占比 74.14%；欺诈样本 181 条，占比 25.86%。样本欺诈标签分布详见图 2，相关变量结构见表 1。样本呈现一定程度的类别不均衡，但尚未达到极端失衡水平，可在不进行重采样的前提下直接用于有监督建模[4]。

Figure 1. Research framework for insurance fraud detection
图 1. 保险反欺诈研究框架

考虑到保险反欺诈应用的实际目标并非单纯提升总体准确率，而是尽可能减少漏识别的欺诈案件，本文在模型评价环节同时关注 AUC、召回率、精确率以及阈值选择。相较于“准确率优先”的一般分类任务，保险欺诈检测更强调召回能力：一笔漏报通常意味着真实赔付损失，而误报则可通过后续人工复核加以筛除，二者的成本结构并不对称。



注：训练集中非欺诈样本 519 个，欺诈样本 181 个。

Figure 2. Distribution of fraud labels in the sample
图 2. 样本欺诈标签分布

Table 1. Data structure and variable processing
表 1. 数据结构与变量处理说明

变量类别	主要字段/处理方式
客户基本信息	年龄、性别、教育水平、客户关系时长等
保险政策信息	年度保费、免赔额、CSL、伞式限额等
事故与索赔信息	事故类型、出险时间、索赔金额及区间变量
目标变量	fraud = 0/1

3. 数据预处理与特征工程

在数据清理方面，本文首先设定 30%作为缺失字段的剔除阈值。训练集中，property_damage 与 police_report_available 两个变量的缺失率分别为 37.00%与 35.20%，且与目标变量的关联较弱，故予以剔除；policy_id 为样本编号类字段，对预测不具实质贡献，亦同步删除。对缺失比例较低的类别变量，采用众数填补以保持原始分布；对 collision_type 等缺失比例较高但具有业务含义的类别变量，则单独设置“Unknown”类别，以保留缺失本身所携带的信息。

在变量处理方面，本文将 policy_bind_date 与 incident_date 转换为日期格式，并以两者之差计算投保至出险的间隔天数；将 age、customer_months、policy_annual_premium、injury_claim、property_claim、total_claim_amount 及 vehicle_claim 等变量按业务区间离散化，以刻画不同年龄层、客户关系周期与索赔金额区间中的风险差异。对二元或有序变量采用标签编码，对无序多分类变量采用独热编码，以适配随机森林与 XGBoost 等基于树结构的模型。

4. 探索性数据分析

描述性分析显示，样本平均年龄为 38.65 岁，年龄范围为 19~63 岁，主要集中于 30~50 岁区间；平均客户关系时长为 205.30 个月(约 17 年)，并与年龄变量呈现高度正相关，相关系数达 0.92。年度保费均值为 1246.55 元，分布区间为 411.66~2004.59 元，整体呈右偏特征，大部分样本集中于 1000~1500 元区间。

从欺诈模式来看，高额索赔与欺诈标签之间存在较为明显的正向关联。车辆索赔金额与欺诈的相关系数为 0.165，总索赔金额与欺诈的相关系数为 0.155；追尾事故与欺诈亦存在一定相关性，相关系数为 0.116。进一步的金额区间分析显示，总索赔金额位于 60000~70000 元区间、财产索赔金额位于 19000~20000 元区间时，欺诈风险相对更高，提示欺诈行为可能集中于“收益较高但不易触发强审查”的金额区间。车辆索赔金额与欺诈标签的关系见图 3。

5. 模型构建与实证结果

本文按 8:2 比例将数据划分为训练集与验证集，分别为 560 条与 140 条。在模型比较阶段，本文分别构建随机森林与 XGBoost 两种集成学习模型[5] [6]。XGBoost 模型经网格搜索后确定如下超参数(模型参数与核心评价指标见表 2)：特征采样比例 1、学习率 0.05、最大深度 6、随机树数量 200、子采样率 1。随机森林模型经参数调优后，在最终建模阶段选取随机树数量 100、最大深度 8、叶节点最小样本量 10、最大特征比例 0.2。

为提升超参数调优过程的透明度与可复现性，本文采用基于 5 折交叉验证的网格搜索(GridSearchCV)方法进行参数选择。具体而言，将训练集随机划分为 5 个互斥子集，每次以其中 4 份用于模型拟合、1 份用于验证，循环 5 次后以平均 AUC 作为评价标准；对随机森林模型，搜索空间设定为：随机树数量∈

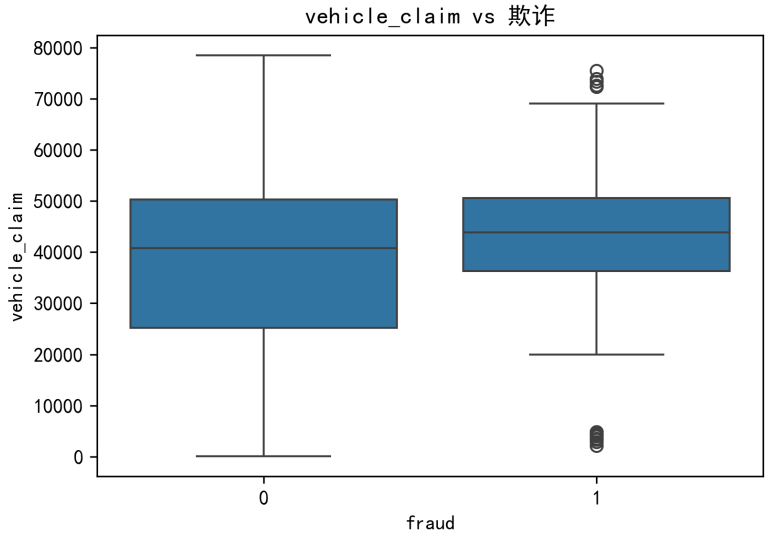


Figure 3. Relationship between vehicle claim amount and fraud label
图 3. 车辆索赔金额与欺诈标签的关系

Table 2. Model parameters and core evaluation results
表 2. 模型参数与核心评价结果

模型/步骤	关键参数或指标	结果/说明
XGBoost	学习率 0.05，最大深度 6，树数量 200	用于与随机森林进行基准比较
随机森林	树数量 100，最大深度 8，叶最小样本量 10，最大特征数量 0.2	最终用于深入分析
阈值选择	阈值 0.35	兼顾召回率与精确率
模型表现	AUC≈0.723，召回率 0.6909，精确率 0.4401	具备可用的风险排序能力

{100, 200, 300}、最大深度∈{6, 8, 10, 不限}、叶节点最小样本量∈{5, 10, 20}、最大特征比例∈{0.2, 0.3, sqrt}；对 XGBoost 模型，搜索空间设定为：学习率∈{0.05, 0.1, 0.2}、最大深度∈{4, 6, 8}、随机树数量∈{100, 200, 300}、子采样率∈{0.8, 1.0}、特征采样比例∈{0.8, 1.0}。最终选定的参数组合为上述网格中平均验证 AUC 最高且模型复杂度可控的方案，从而在偏差–方差权衡之间取得平衡。

在阈值选择方面，本文以“尽可能识别真实欺诈样本”为目标，对不同阈值下的召回率与精确率进行系统比较。结果显示，当阈值设定为 0.35 时，模型能够在较高召回率与可接受精确率之间实现平衡。基于该阈值，模型召回率为 0.691，精确率为 0.440，意味着模型可识别出约 69%的真实欺诈案件；同时仍存在一定误报，需与人工复核机制配合使用。

从 ROC-AUC 结果看，最终随机森林模型的 AUC 约为 0.723 (两模型 ROC 曲线对比见图 4，阈值敏感性见图 5，最终模型 ROC 曲线见图 6，特征重要性排序见图 7)，具备可用的区分能力。特征重要性分析进一步显示，事故严重程度评分以 0.368 的贡献度显著领先，成为最具判别力的核心特征；投保至出险时间间隔、保费比、年保费额与客户关系时长等变量也对模型判断起到辅助作用。上述结果与保险反欺诈实践中对“事故损失严重程度”及“短期投保即出险”风险模式的经验认识高度吻合。

随机森林模型自带的特征重要性指标只能反映特征对整体预测精度的平均贡献，难以解释模型在具体样本上的判别逻辑。为弥补这一不足并增强方法论的严谨性，本文进一步引入基于 Shapley 值的可解释性框架 SHAP (SHapley Additive exPlanations) [7]进行补充分析。SHAP 通过将每个特征对单个预测的边际

贡献以可加形式分解，既可在全局层面给出与特征重要性一致的稳健排序，又可在局部层面针对每一笔可疑赔案输出“哪些特征将其推向欺诈/非欺诈方向、推动幅度多大”的具体解释，从而满足保险业务中对模型决策可追溯性的实际要求。

全局 SHAP 分析显示，事故严重程度评分、投保至出险的间隔天数、车辆索赔金额与年保费比是推动模型预测向“欺诈”方向偏移的主要变量，与基于 Gini 不纯度的特征重要性排序高度一致，验证了模型识别逻辑的稳健性。在样本级解释上，本文随机抽取了若干被模型判定为高欺诈概率的样本进行案例分析：例如，某保单投保仅 45 天即发生大额索赔，SHAP 分解显示“投保至出险时间间隔”贡献了 +0.18 的概率推动，“事故严重程度”贡献 +0.12，“车辆索赔金额”贡献 +0.07，三者共同将该样本的欺诈预测概率从基线水平的 0.26 抬升至 0.71；与之相对，某长期保单的低额索赔样本则因“客户关系时长”

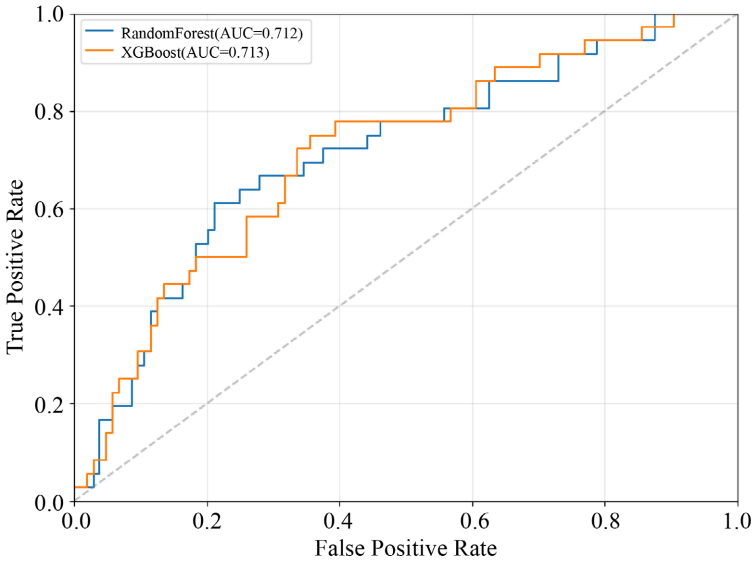


Figure 4. ROC Curve comparison between random forest and XGBoost
图 4. 随机森林与 XGBoost 模型 ROC 曲线比较

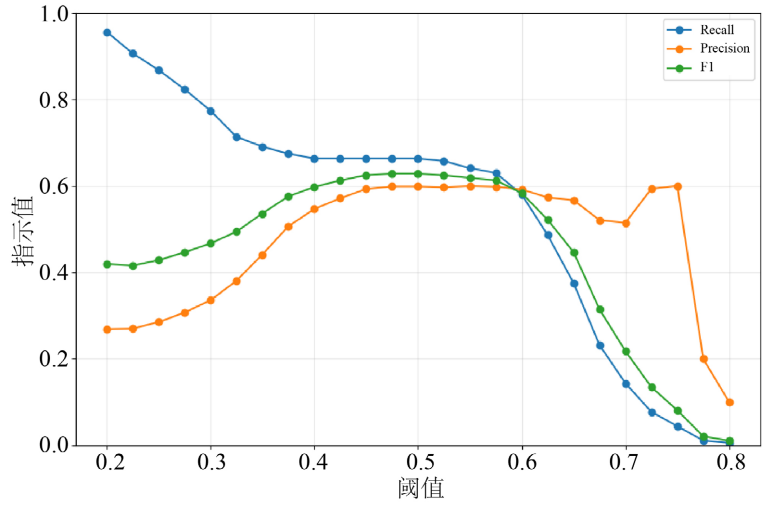


Figure 5. Model evaluation metrics across different thresholds
图 5. 阈值变化下模型评价指标波动

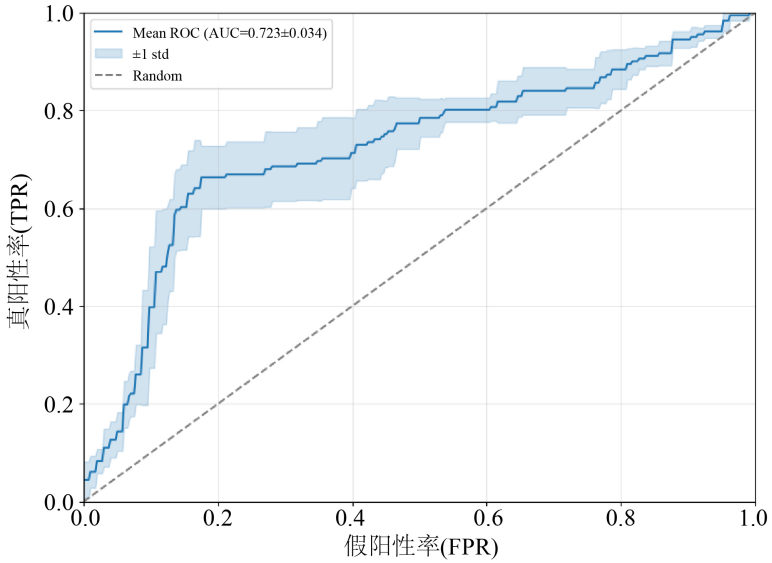


Figure 6. ROC curve of the final model
图 6. 最终模型 ROC 曲线

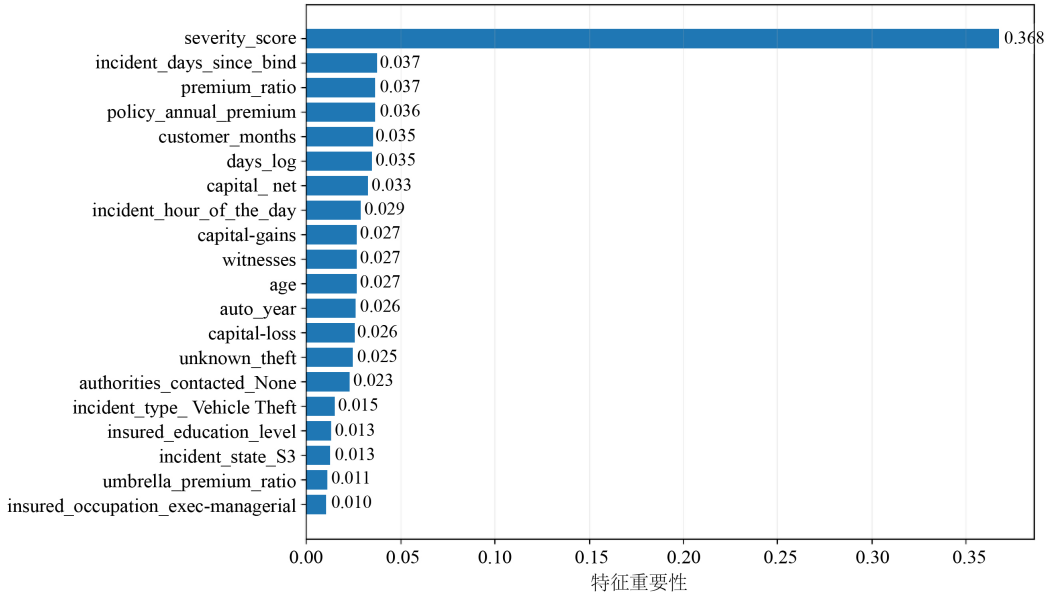


Figure 7. Feature importance ranking of the random forest model
图 7. 随机森林模型特征重要性排序

和“年保费”的负向贡献被判为低风险。上述案例及解释表明，模型并非依赖单一特征作出“黑箱”判断，而是在多个业务变量的协同作用下形成可被业务人员理解的判别路径，可作为人工复核与稽核环节的有力辅助工具。

6. 业务应用与风控实施路径

基于模型输出的欺诈概率，本文建议构建三级风险分层机制。对预测概率不低于 0.70 的高风险客户，应启动强化核保与重点调查流程，包括增加财务核查、补充事故材料、设置更高免赔额或调整保额上限；

对预测概率位于0.40~0.70之间的中风险客户，执行标准审核并在投保后一定期间内进行重点跟踪；对预测概率低于0.40的低风险客户，可提供快速通道与简化核保服务，以降低运营成本并优化客户体验。具体三级风险分层处置策略见表3。

在系统实施层面，可将模型嵌入核保与理赔业务系统，构建“自动风险评分–智能分案–人工复核–反馈学习”的闭环机制。模型不宜替代人工判断，而应作为风险排序与资源配置的辅助工具：其核心价值在于协助保险公司将有限的调查资源优先配置至高风险案件，进而提升调查效率并降低漏赔风险。

Table 3. Three-tier risk-control mechanism based on model output
表3. 基于模型输出的三级风险管控机制

风险等级	预测概率区间	建议处置策略
高风险	$p \geq 0.70$	强化核保、重点调查、补充材料、差异化免赔额或保额上限
中风险	$0.40 \leq p < 0.70$	标准审核、重点跟踪、必要时追加人工复核
低风险	$p < 0.40$	快速通道、简化核保、续保优惠与客户体验优化

7. 结论与不足

本文基于保险反欺诈数据构建机器学习识别框架，结果表明随机森林模型在样本不均衡且欺诈模式相对复杂的情境下具备可用的风险区分能力。模型AUC约为0.723，召回率为0.691，可为保险公司开展风险分层、核保审核与理赔调查提供辅助性决策依据。特征重要性分析进一步表明，事故严重程度、投保至出险时间以及保费相关变量是识别欺诈风险的关键因素。

本文仍存在以下三方面不足：第一，样本规模相对有限，模型的泛化能力需在更大规模的真实业务数据上进一步验证；第二，部分图表与指标来源于竞赛数据的处理结果，正式投稿前仍需补充变量口径界定、代码复现说明与稳健性检验；第三，本文尚未引入社交网络、图神经网络或非结构化文本与图像信息，后续研究可结合团伙欺诈识别与多模态数据进一步提升模型性能。

参考文献

[1] Bolton, R.J. and Hand, D.J. (2002) Statistical Fraud Detection: A Review. *Statistical Science*, **17**, 235-255. https://doi.org/10.1214/ss/1042727940

[2] Phua, C., Lee, V., Smith, K., *et al*. (2010) A Comprehensive Survey of Data Mining-Based Fraud Detection Research. arXiv:1009.6119. https://arxiv.org/abs/1009.6119

[3] 阿里天池. 金融数据分析保险反欺诈预测数据集[DB/OL]. https://tianchi.aliyun.com/competition/entrance/531994/information, 2025-11-20.

[4] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[5] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[6] Chen, T. and Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[7] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 4765-4774.