基于可解释机器学习模型的电信行业客户流失预测研究

王圣节，张庆红
（新疆财经大学统计与数据科学学院，新疆 乌鲁木齐 830012）

摘　要：在电信行业中，客户流失的准确预测对于相关企业维持市场竞争力和增加收益至关重要。为此提出一个结合CatBoost算法和SHAP（shapley additive explanations）模型的客户流失预测框架，旨在提高预测的准确性，同时增强模型的可解释性。利用新疆某通信公司的实际营业数据，通过数据预处理及特征工程，构建预测模型，选取5种主要关键性能指标评估模型性能。实验结果显示，所提出模型在选取的评价指标上均优于当前主流机器学习预测模型。最后引入SHAP框架增强模型可解释性，揭示影响客户流失的关键因素，并提供具体的因素影响程度，为电信企业制定针对性的客户保留策略提供了科学依据。

关键词：机器学习；CatBoost算法；SHAP；预测模型；电信行业
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024166

Research on telecom industry customer churn prediction based on explainable machine learning models

WANG Shengjie, ZHANG Qinghong
College of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi 830012, China

Abstract: In the telecom industry, accurate prediction of customer churn is crucial for the companies involved to maintain market competitiveness and increase revenue. To this end, a customer churn prediction framework combining CatBoost algorithm and SHAP model was proposed, aiming to improve the accuracy of prediction and enhance the interpretability of the model. Using the actual business data of a communication company in Xinjiang, the prediction model was constructed through data preprocessing and feature engineering, and five major key performance indicators were selected to evaluate the model performance. The experimental results show that the proposed model outperforms the current mainstream machine learning prediction models in all the above evaluation indicators. Finally, the SHAP framework was introduced to enhance the model interpretability, reveal the key factors affecting customer churn, and provide the specific influence degree of the factors, which provided a scientific basis for telecommunications enterprises to formulate targeted customer retention strategies.

Key words: machine learning, CatBoost algorithm, SHAP, predictive model, telecom industry

收稿日期：2024−04−17；**修回日期：**2024−05−21
基金项目：国家自然科学基金资助项目（No.72164034）
Foundation Item: The National Natural Science Foundation of China (No.72164034)

0 引言

电信行业作为国家基础设施的关键部分，正面临市场饱和以及激烈竞争的双重挑战。我国现有的四大电信运营商——中国移动、中国电信、中国联通和中国广电在移动通信快速进步的大背景下，同全球通信公司面临的情况一样，都遭遇客户流失率不断攀升和新用户获取成本急剧上涨的问题。特别是在过去的5年中，新客户和新发卡量的增速明显放缓。在这样的市场环境中，维护现有客户基础并减少流失成了运营商的当务之急。与寻找新客户相比，保持已有的优质客户成本更低，尤其是在运营商之间的竞争异常激烈时，客户流失对企业盈利的负面影响尤为显著。

客户流失预测在电信行业中扮演着举足轻重的角色。电信运营商不仅要努力保留现有客户，还需要加强客户关系管理（customer relationship management，CRM）的力度。维持现有客户的忠诚度尤为关键，因为在市场饱和以及竞争白热化的情况下，客户可能随时选择转投其他服务提供商。为此，电信运营公司纷纷着手开发相应程序，以及时识别并挽留这些客户。毕竟，与吸引新客户相比，维持现有客户的成本要低得多。具体来说，吸引新客户所涉及的广告、人力资源和折扣等成本，可能是维持现有客户成本的5~6倍。因此，仅须略加关注，便可有效地识别和逆转客户流失的趋势，无须进行复杂的投入。

为了精确满足保留客户的需求，电信运营公司必须构建一个准确且高性能的模型。这个模型不仅要能识别出流失客户，还要能深入挖掘流失背后的原因。这样，公司才能制定出有效的客户挽留措施。此外，该模型还应运用先进技术，预测未来可能出现客户流失的时间点，以便公司提前做好应对准备。

随着大数据领域的蓬勃发展，数据挖掘和机器学习等解决方案为分析此类问题提供了有力支持。这些技术能够深入挖掘数据，揭示客户流失的潜在原因，从而帮助CRM部门制定更具针对性的策略，最大化公司的利润。同时，它们也可用于设计客户挽留方案，有效降低流失客户比例。

现有研究表明，电信运营商的主要目标是利用大量的电信数据来识别有价值的流失客户。然而，现有模型存在几个限制，这些限制在实际环境中对这个问题构成了严重障碍。电信行业产生的大量数据中包含缺失数值，这导致预测模型的预测结果不佳。为了解决这些问题，使用数据预处理方法消除数据中的缺失值和异常值，这能改善模型性能并正确分类数据。机器学习算法，尤其是如CatBoost这样的先进算法，具备额外能力能够处理和分析庞大的数据集，这对于拥有庞大客户基础的电信行业来说至关重要。CatBoost算法尤其适用于处理类别数据，能够提高模型训练的速度和预测的准确性，从而使企业能够有效利用大量历史数据来训练模型。此外，结合SHAP（shapley additive explanations）解释模型，不仅能预测客户流失，还能了解哪些因素最影响预测结果，为电信企业提供有针对性的客户保留策略。这种方法的应用展现了技术创新的跃进，同时也体现了新质生产力在电信行业的实践价值。

综上，本文结合CatBoost算法和SHAP模型提出一种电信行业客户流失预测模型。使用新疆某通信公司的真实数据集进行验证，所提出的流失预测模型使用信息检索指标进行评估。通过使用5种性能评价指标——准确率（accuracy）、精确率（precision）、召回率（recall）、F1值和AUC值来计算流失预测模型的准确性。本研究的目标是调查现有的机器学习和数据挖掘技术，并提出一个用于客户流失预测、识别流失因素并提供保留策略的模型，为相关部门提供对应的参考意见。

1　相关研究

在通信公司的客户流失预测领域，多种方法已被广泛应用，其中大部分研究集中于机器学习和数据挖掘技术。现有的相关工作主要分为两种：一种是应用单一数据挖掘方法提取知识；另一种则是比较多种模型以预测客户流失。通过运用机器学习、数据挖掘及其混合技术，这些研究支持企业识别、预测并留住潜在流失的客户，从而辅助决策制定和客户关系管理。

有关流失预测的研究可分为如下两类。

第一类是相关学者基于公开可用的数据集提出流失预测模型。Brandusoiu 等[1]应用支持向量机（support vector machine，SVM）将具有不同核函数的支持向量机应用于公开数据集，该数据集包含 3 333 个用户和 21 个变量。在这项研究中，多项式核函数 SVM 的准确率为 88.56%。在 Sabbeh[2]的研究中，在同一数据集进行了 10 种不同的机器学习算法模拟实验，并进行了比较研究。结果显示，随机森林算法和 AdaBoost 算法在数据集的 10 倍交叉验证中获得了 96.39% 的最高准确率。在同一数据集上，Jain 等[3]分别使用 Logistic Regression 和 Logit Boost 进行了预测分析，取得了 85.238 5% 的准确率，并得到了 Logistic 回归优于 Logit 提升的结论。同样是在该数据上，Stehani 等[4]比较了逻辑回归（logistic regression）、SVM、朴素贝叶斯（naive Bayes）、AdaBoost、*K*NN（*K*-nearest neighbours）和随机森林 6 种不同算法的性能指标，并且使用随机森林算法取得 89% 的准确率。Khamlichi 等[5]在一个包含 5 000 个样本的公开数据集上应用了不同的独立机器学习技术，如随机森林、XGBoost、SVM、决策树、逻辑回归和 *K*NN。其中，XGBoost 的准确率为 95%，F 值为 80%，表现较好。Pamina 等[6]在研究中使用了另一个包含 7 043 名订阅者的公开可获得的数据集，并应用了 *K*NN、XGBoost 和随机森林算法。他们在该数据集上利用 XGBoost 获得了 79.8% 的准确性。Tang 等[7]在研究中使用了一个包含 7 043 个样本的公开数据集，应用了 *K*-means 聚类技术和 XGBoost 算法。首先他们在训练集上使用了 *K*-means 聚类技术，然后分别使用 XGBoost 对聚类组进行训练，通过这种方式，将 XGBoost 单独使用时获得的 79% 的准确性提高到混合模型中的 81%，并且对比了独立模型和混合模型的性能指标。在 Wu 等[8]的研究中，考虑了 3 个不同的公开数据集，并对每个数据集分别进行了逻辑回归、决策树、随机森林、朴素贝叶斯、AdaBoost 和多层感知器（MLP）等各种机器学习算法的比较研究。他们还利用 *K*-means 聚类技术为每个数据集提出了客户分割框架。Zhang 等[9]使用 3 家主要通信公司的数据，并利用 Fisher 判别方程和逻辑回归分析构建了电信客户流失预测模型。根据结果，可以得出回归分析构建的电信客户流失模型具有更高的预测准确度（93.94%）和更好的效果。Lalwani 等[10]应用逻辑回归、朴素贝叶斯、SVM、随机森林、决策树和 XGBoost 等，最终评估结果显示，AdaBoost 和 XGBoost 分类器在准确率上表现最佳，分别为 81.71% 和 80.8%；而在 AUC 得分上，AdaBoost 和 XGBoost 分类器均达到最高的 84%；其中 XGBoost 分类器的 AUC 得分为 84%，在所有分类器中居首位。汪明达等[11]提出利用机器学习和投票相结合的混合模型来预测电信领域的流失用户，在公众数据集 AdaBoost 和梯度提升（gradient boosting）一次投票分类后 AUC 值能够达到 0.677 1。

第二类的学者针对真实数据集进行了深入研究。在 Owczarczuk 等[12]的研究中，采用逻辑回归模型和决策树模型来分析波兰移动运营商提供的包含 122 098 位客户记录的数据集，并使用提升曲线（lift curve）进行模型评估。类似地，Senthan 等[13]在斯里兰卡电信行业的本地数据集上进行了独立模型的比较研究，使用了决策树、

神经网络、逻辑回归、随机森林、XGBoost、AdaBoost和SVM，并发现XGBoost算法表现良好，准确率达到82.90%。此外，Sulikowski等[14]也在波兰的本地数据集上进行了研究，旨在通过使用共线性测试来确定流失预测的潜在因素，以确定电信行业客户流失行为的主要影响因素。Ullahi等[15]采用分类和聚类技术来识别电信行业的流失客户以及导致客户流失的因素，并最终通过随机森林取得88.63%的准则率。Shrestha等[16]对尼泊尔一个主要电信行业的本地数据集进行分析，该数据集包含52 332条客户记录，其中46 204条为未流失客户，6 128条为流失客户。并在该数据集上应用XGBoost算法，获得的准确率和F1值分别为97%和88%。

深入分析现有文献发现，电信行业客户流失预测的差异主要体现在数据集的选择和预测方法上。尽管现有研究中的机器学习方法在性能上表现较为出色，但在多个预测性能指标上仍有提升的潜力。此外，大部分现有研究倾向于使用如XGBoost、随机森林和决策树等所谓的“黑箱”机器学习模型，如文献[17]使用XGBoost和SHAP方法解释该电信流失预测模型，但CatBoost模型在某些场景表现更加优异。虽然这些模型展现了较高的性能，但它们普遍缺乏可解释性。文献[18]使用了CatBoost算法处理电信行业数据中的不平衡问题，但缺乏可解释性，模型仍然是黑箱模型，这一点不利于客户关系管理部门根据模型预测制定有效的客户保留策略。

针对以上问题，本文利用中国移动新疆分公司的实际营业数据作为研究案例，通过对数据进行清洗、异常值处理、缺失值处理、特征转换及独热编码等预处理步骤，构建了一个性能优化的预测模型。研究工作主要聚焦于以下两个核心领域。

（1）建立CatBoost算法驱动的客户流失预测模型。本文利用CatBoost算法对电信行业的客户流失进行预测。通过将模型的准确性、精确度、召回能力、F1值及AUC值等核心评估指标与其他现有主流机器学习模型进行对比，旨在评估并证明所提模型的性能优势。

（2）利用SHAP模型增强预测模型的可解释性。在确保CatBoost模型预测性能的基础上，本文进一步引入SHAP分析，旨在揭示影响客户流失的关键因素。通过深入分析这些因素，为电信企业在制定客户保留策略及防止客户流失的决策制定过程中提供科学依据和参考。

通过以上研究框架，本文不仅提出了一个高效准确的客户流失预测模型，且通过SHAP模型的应用，进一步提升了模型的可解释性，为电信行业提供了一个有效的决策支持工具，有助于企业在激烈的市场竞争中维护客户基础，优化服务质量。

2 客户流失预测模型构建

模型流程如图1所示。

2.1 客户流失模型构建

假定通信数据集 $D=\left\{\left(x_k,y_k\right)\right\}_{k=1}^{n}$，$x_k=\left(\boldsymbol{x}_k^1,\cdots,\boldsymbol{x}_k^m\right)$，数据集共有$m$个特征向量，涵盖了客户的手机品牌、通话时长、套餐使用、年龄等信息，$y_k\in\{0,1\}$代表预测的目标变量。其中，x_k表示数据集中的第k个样本，而y_k则对应第k个样本的实际标签值。

本文所采用的CatBoost[19]算法，是一种在梯度提升决策树（gradient boosting decision tree，GBDT）框架下，以对称决策树为基学习器的一种优化实现。该算法的主要优势在于其能够高效地处理类别特征，同时解决了预测偏移和梯度偏差的问题，从而显著降低过拟合的风险，并提升模型的泛化能力和预测准确性。在处理离散特征时，CatBoost采用了一种名为GreedyTS的策略[20]。该方法的核心思想是在构建决策树时，将

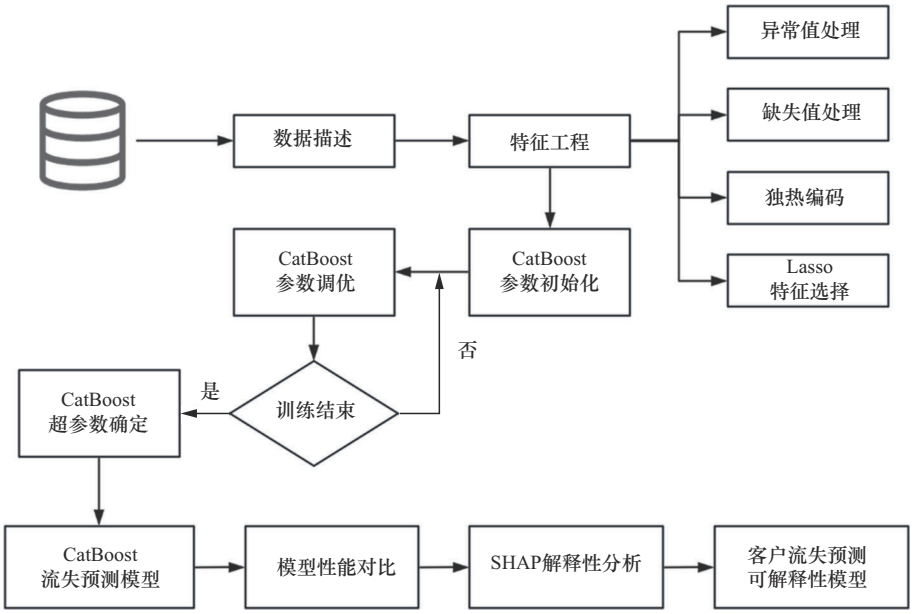


图1　模型流程

标签的平均值作为节点分裂的准则。具体来说，该策略的计算式表达如下。

$$x_{i,j}=\frac{\sum_{i=1}^{n}[x_{k,j}=x_{i,j}]\cdot Y_k+\lambda\cdot P}{\sum_{i=1}^{n}[x_{k,j}=x_{i,j}]\cdot Y_k+\lambda} \tag{1}$$

其中，$x_{i,j}$代表第j个特征的第i类值，Y_k为相应的标签值，分子是指所有第j个特征的第i类值所对应的标签值的总和，而分母则是第j个特征的第i类值的数量[21]，P是加入的先验项，而λ是一个大于0的权重系数。引入先验分布项，可以有效减少数据噪声的影响，并缓解低频类别数据对模型的负面影响。

2.2　SHAP 解释性方法

虽然基于CatBoost算法的客户流失预测模型能够达到较高的性能表现，但这种性能的提升不可避免地引入了模型作为“黑盒”的问题，从而降低了其可解释性。为了解决这一问题，引入SHAP模型。SHAP模型基于博弈论中的Shapley值概念，构建了一个强有力的框架，用于解释机器学习模型的预测结果。它的核心目的在于通过量化每个特征对预测结果的贡献，解释模型对单个样本的预测输出[22]。这样的方法旨在提高模型预测的透明度，确保模型的决策过程对研究人员和对非技术背景的用户都是可理解的。SHAP方法不仅可以揭示每个样本中各个特征对预测结果的影响力，还能区分这种影响的正负方向。由Lundberg等[23]提出的SHAP方法，通过解释Shapley值作为加性特征归因方法（additive feature attribution method），将模型预测值分解为特征值贡献的线性组合，进一步增强了模型预测的可解释性。函数表达式如下。

$$f(\boldsymbol{x})=\phi_0+\sum_{k=1}^{M}\phi_i x_i \tag{2}$$

其中，$f(x)$代表针对输入$\boldsymbol{x}$的模型预测值，M为特征总数，x_i表示输入特征向量$\boldsymbol{x}$中第i项特征的值，ϕ_i代表第i个特征对模型预测值的贡献度，亦即SHAP值，ϕ_0表示模型输出的基准值，即在没有任何特征贡献时的预测值。SHAP方法的精髓在于为每个特征ϕ_i计算其贡献值，这些值体现了在其他特征值固定的情况下，第i项特征对模型预测的平均影响。此方法不仅揭示了各个特征对预测结果的影响程度，还能指出这种影响是正是负。

3 特征工程及处理

3.1 数据概况

采用来自新疆某通信公司的真实营业数据集，去除敏感信息后样本总数为100 000，特征总数为43，数据集中有部分缺失值，数据集特征见表1。

3.2 特征工程

（1）异常值处理

针对所使用数据集中存在的异常值问题，使用四分位距（interquartile range，IQR）方法计算出数据集的四分位数间距，然后针对列表中的每个字段应用IQR方法来检测极端值。如果发现极端值，函数会使用盖帽法对其进行处理，即将超出上下界的值分别替换为上下界。实验中发现有15个特征含有异常值，特征变量的箱线图如图2所示，本研究使用盖帽法处理异常值，并合理使用最大值或最小值替换异常值。

（2）缺失值处理及独热编码

部分变量与目标变量IS_CHURN热力图如图3所示，目标变量是否流失（IS_CHURN）和信控标识（CREDIT_FLAG）、入网时间（NEW_DATE）、城乡标识（COUNTRY_FLAG）等变量的相关系数较高，说明以上变量的变动可能对是否流失生较大影响。

数据缺失情况如图4所示，缺失数据的处理按照特征的属性区别对待，进行独热编码，最终得到40个特征。

（3）数据稳健标准化

由于在真实数据集中，数据中会出现离群值

表1 数据集特征

变量名称	变量解释	变量类型	变量名称	变量解释	变量类型
CUST_TYPE	客户类型	类别型	IS_PAY	进30天是否有缴费行为	类别型
CUST_TYPE2	客户类型2	类别型	QFJE	欠费金额	数值型
SALESNAME	套餐名称	类别型	QFZQ	欠费周期	数值型
HX_LATN_NAME	地州名称	类别型	IS_HY	是否有未到期合约	类别型
CREDIT_FLAG	信控标识	类别型	IS_BILLING_LAST	上月是否出账	类别型
NEW_DATE	入网时间	日期型	IS_ACTIVE_LAST	上月是否活跃	类别型
ZFK	主副卡标识	类别型	IS_YC	本月是否有溢出费用	类别型
PHONE_MODEL	终端机型	类别型	IS_YC1	上月是否有溢出费用	类别型
PHONE_PATTERN	终端制式	类别型	IS_TS	进一月是否有费用争议的投诉行为	类别型
COUNTRY_FLAG	城乡标识	类别型	ALL_FLUX_USE	近一月的流量总量（M）	数值型
NATION	族别	类别型	CV_FLUX_USE	近4周的流量使用量波动	数值型
CUST_AGE	年龄	数值型	FLUX_USE_HB1	本周及上周流量使用量环比	数值型
DLL	是否大流量套餐	类别型	FLUX_USE_HB2	上周及上上周流量使用量环比	数值型
VOICE_DUR	语音时长	数值型	FLUX_USE_HB3	上上周及上上上周流量使用量环比	数值型
BALANCE_CAS	现金预存款余额	数值型	CV_F_CHARGE	近半年账单收入波动	数值型
PHONE_BRAND	终端品牌	类别型	ALL_CALL_DUR	近一月的语音总量（Min）	数值型
IN_MONTHS	在网时长	数值型	CV_CALL_DUR	近4周的语音量使用量波动	数值型
YXCP	客户下有效产品数	数值型	CALL_DUR_HB1	本周及上周语音使用量环比	数值型
ARPU	近3个月计费收入	数值型	CALL_DUR_HB2	上周及上上周语音使用量环比	数值型
ARPU_CHANGE	ARPU值较上月变化	数值型	CALL_DUR_HB3	上上周及上上上周语音使用量环比	数值型
PACKAGEEXES_CHANGE	套餐值较上月变化	数值型	IS_CHURN	是否流失	类别型

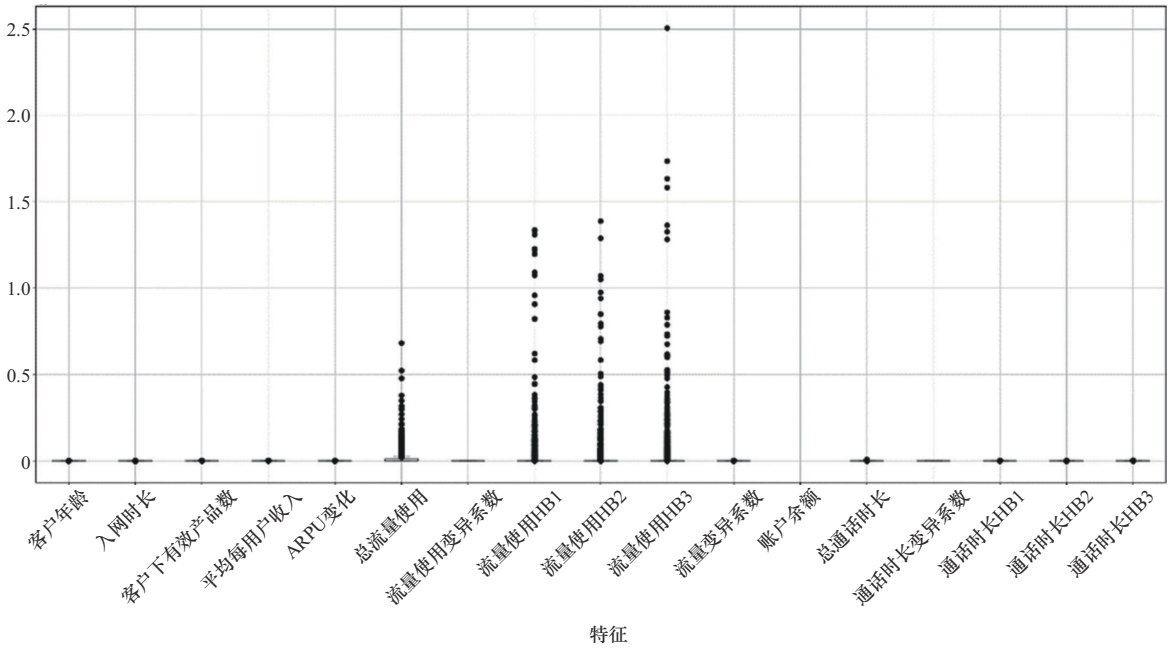


图2　特征变量的箱线图

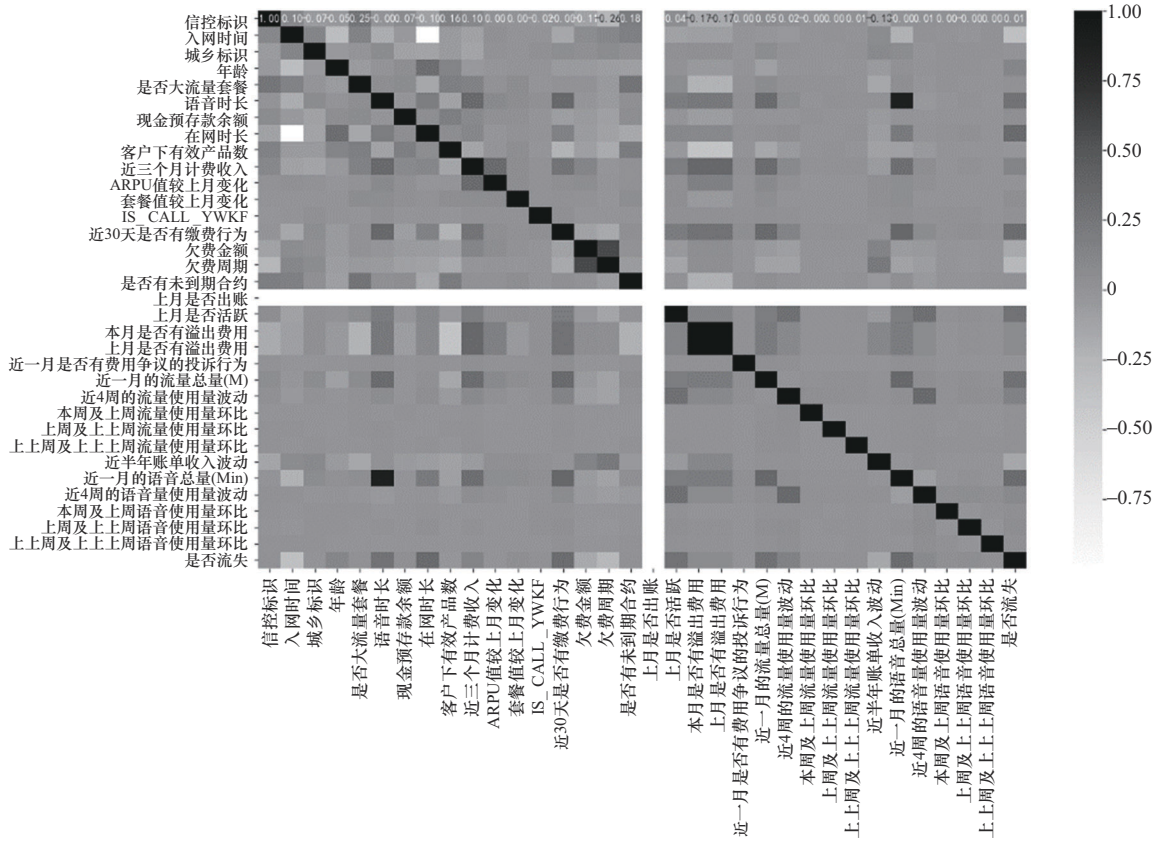


图3　部分变量与目标变量IS_CHURN热力图

问题，如果直接使用平均值和标准差进行标准化，那么离群值会对结果产生很大影响。因此使用稳健标准化（RobustScaler）这一更加稳健的中位数和四分位数范围来标准化数据，使得它对

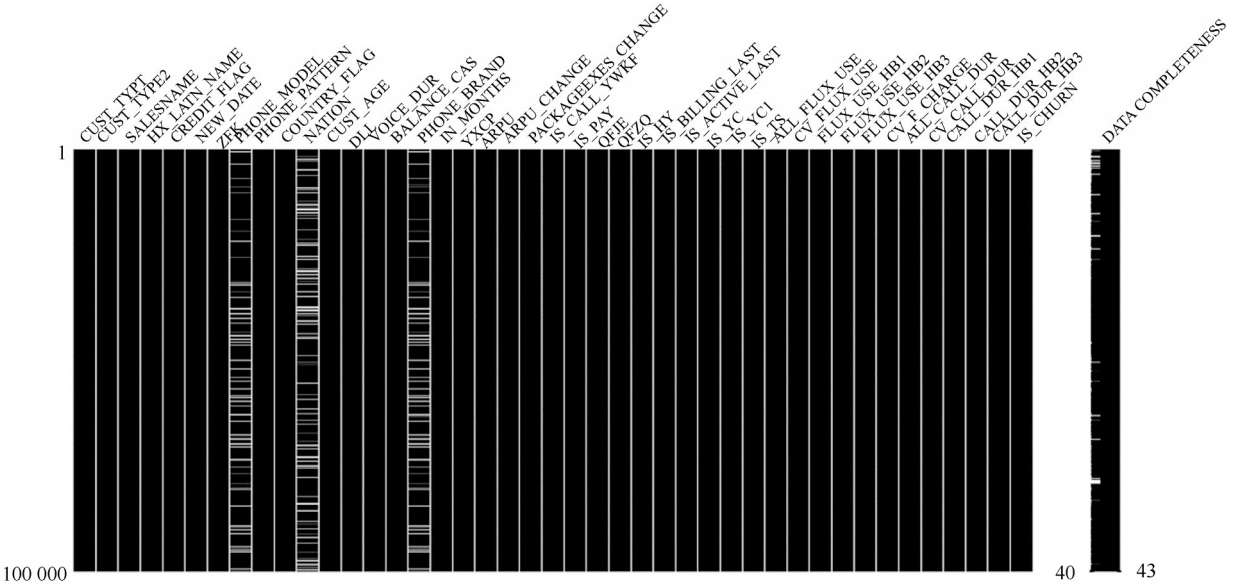


图4　数据缺失情况

离群值不那么敏感。计算式如下。

$$X_{\text{scaled}}=\frac{X-\text{median}(X)}{Q3(X)-Q1(X)} \tag{3}$$

其中，RobustScaler 主要移除中位数，并通过 IQR 来缩放数据。IQR 是第三四分位数（75th quantile）与第一四分位数（25th quantile）之差。因此，标准化后的数据的范围会在大多数数据的范围内。

4　实验及结果分析

4.1　实验环境及评价指标

实验环境为 Windows 11 操作系统，13th Gen Intel(R) Core(TM) i7-13700H，2.40 GHz，16 GB 机带 RAM，NVIDIA GeForce RTX 4060 Laptop GPU，采用 Python3.8 编程语言，通过 Jupyter 和 Pycharm 专业版编辑器进行仿真实验。

通常在研究客户流失率预测时，使用混淆矩阵展示客户流失预测结果。该矩阵详细地反映了各类别下，实例的正确分类与错误分类数。正确分类包括真阳性（true positive，TP）和真阴性（true negative，TN），错误分类则是假阳性（false positive，FP）与假阴性（false negative，FN）。虽然混淆矩阵能详细描述分类效果，但它本身并不能直接提供明确的结论，因此在实际应用场景中，通常辅助使用其他评估指标来全面评估模型的性能。本实验采用以下 5 项主要的分类评估指标：精确率（precision）、召回率（recall）、准确率（accuracy）、F1 值和 AUC 值。这些指标从不同角度反映了模型预测的质量和可靠性。各个指标的表达式如下所示，混淆矩阵见表 2。

表 2　混淆矩阵

真实结果	预测结果	
	流失	未流失
流失	TP	FN
未流失	FP	TN

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5}$$

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{FN}+\text{FP}+\text{TP}+\text{TN}} \tag{6}$$

$$\text{F1}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{7}$$

4.2　对比实验分析

将本文算法和当前电信行业其他 7 种主流机

器学习模型的预测结果性能进行对比分析，其中LR代表逻辑回归，RF代表随机森林，DT代表决策树，GBoosting代表梯度提升，模型性能指标对比见表3。

对比分析各模型的性能指标，结果揭示了CatBoost模型在准确率、精确率、召回率、F1值和AUC值等关键性能指标上均优于其他模型，显著展现了其在预测任务中的高效能。此外，LightGBM和XGBoost模型亦表现出较好的性能，紧随CatBoost模型之后。与之形成对比的是，逻辑回归和决策树模型在上述性能指标中的表现较弱，尤其在准确率和召回率这两个关键指标上更是如此。基于上述分析，可以得出结论，CatBoost模型在本文评估的多个评价指标上均表现出较高的性能。为了进一步验证这一结论，本文采用了以准确率为基准的10折交叉验证方法，并通过箱线图展示了验证结果，10折交叉验证箱线图如图5所示。

在本研究中，使用以准确率为基准的10折交叉验证箱线图，对模型性能进行了初步的对比分

表3　模型性能指标对比

模型	准确率	精确率	召回率	F1值	AUC值
CatBoost	0.852 9	0.844 8	0.865 1	0.854 8	0.926 4
LightGBM	0.849 1	0.841 0	0.860 9	0.850 8	0.922 3
XGBoost	0.846 5	0.839 5	0.856 7	0.848 0	0.921 3
RF	0.847 5	0.835 6	0.865 2	0.850 1	0.913 9
GBoosting	0.844 9	0.835 7	0.858 7	0.847 0	0.911 8
*K*NN	0.814 65	0.822 165	0.803 694	0.812 825	0.781 8
LR	0.760 5	0.787 9	0.713 1	0.748 5	0.818 8
DT	0.769 4	0.770 0	0.768 4	0.769 2	0.769 4

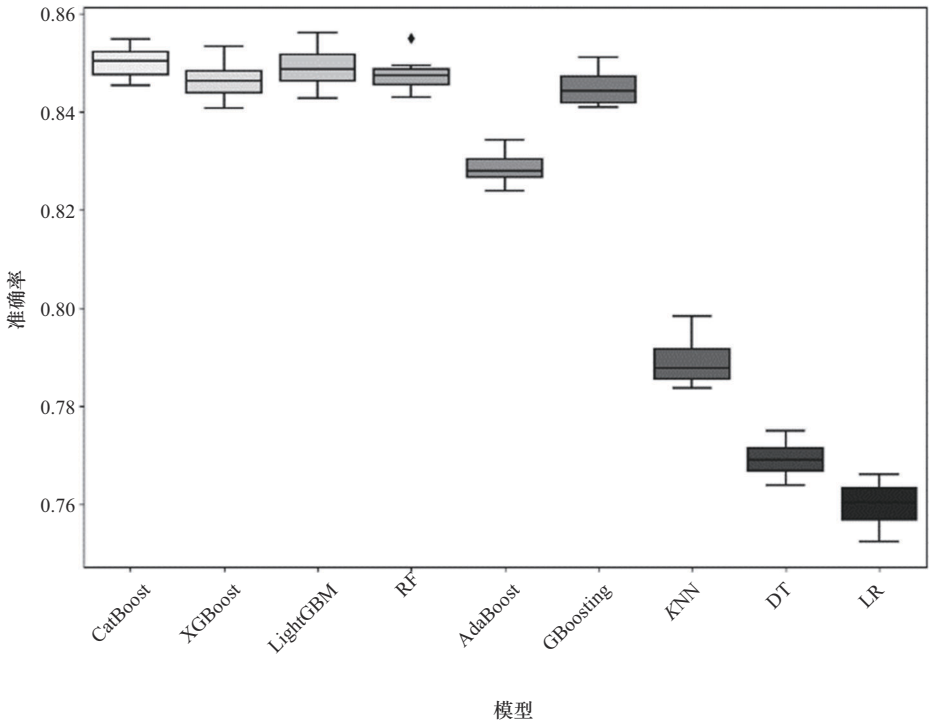
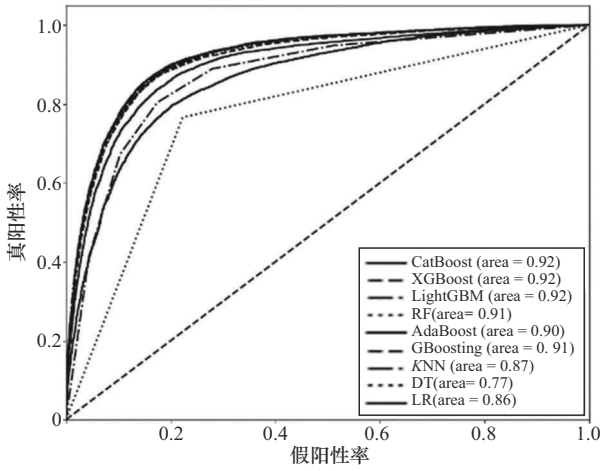
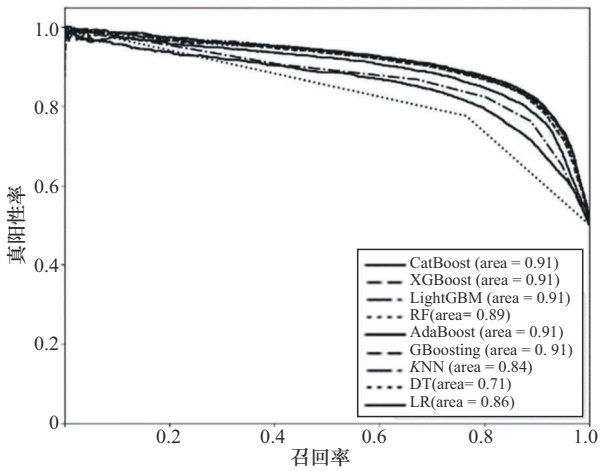


图5　10折交叉验证箱线图

析。结果显示，CatBoost模型在精度上的表现超越了其他比较模型。为了进一步深入分析各模型的性能表现，绘制了AUC（area under the curve）曲线和精确率—召回率曲线，CatBoost与主流机器学习性能比较如图6所示。这两种曲线提供了更全面的视角来评估和比较不同模型在客户流失预测任务上的表现，为选择最适合的模型提供了科学依据。



（a）ROC曲线



（b）精确率—召回率曲线

图6　CatBoost与主流机器学习性能比较

根据表2和图6中的实验数据，可以观察到CatBoost模型在精确率、准确率以及AUC值等关键性能指标上超过了其他比较模型。CatBoost模型之所以能够取得这样的成绩，主要得益于其对离散型特征采取的目标统计编码方法，该方法能有效减少数据中的噪声。此外，CatBoost通过解决梯度偏差和预测偏移的问题，显著降低了模型训练过程中的过拟合风险，进而提升了模型预测的整体性能。

4.3　SHAP可解释性分析

模型的可解释图如图7所示，显示了本文研究所提出的基于CatBoost和SHAP模型的摘要和自带的特征重要性的排序。

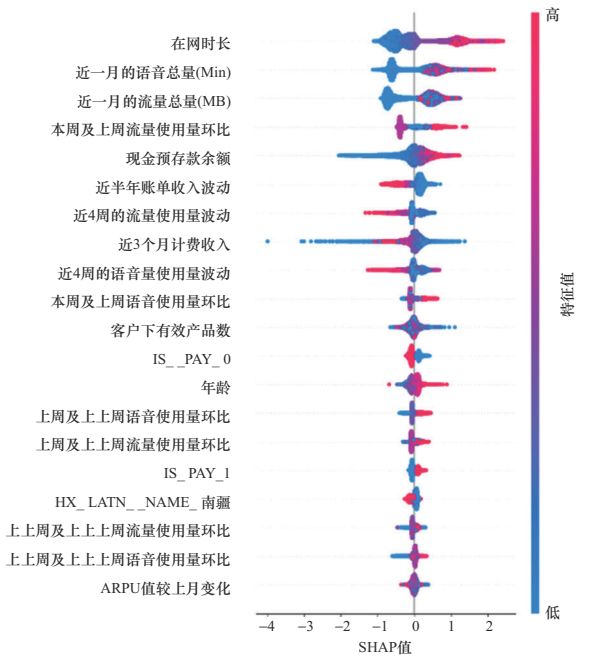
如图7（a）所示，特征1（年龄）、特征12（现金预存款余额）、特征5（近3个月计费收入较上月的变化）、特征7（近4周的流量使用量波动）对用该公司的客户是否流失有关键影响，且对模型的特征较为重要。

特征1（年龄）：对客户流失的影响表现在不同年龄段的客户可能有不同的消费习惯、忠诚度和需求。例如，年轻客户可能更倾向于尝试新的电信服务或优惠，而老年客户可能更看重稳定性和服务质量。年龄也可能与客户的生命周期阶段相关，如年轻人可能因经济能力增强而更换更高级的服务，而老年人可能因退休等因素减少电信消费。

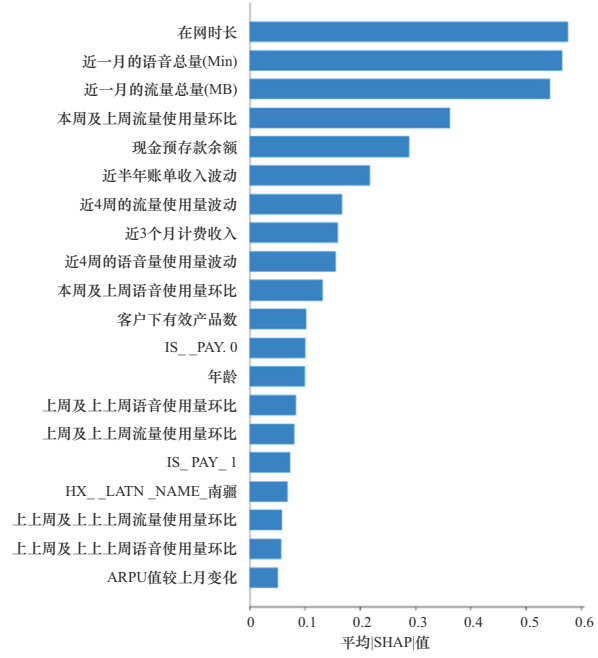
特征12（现金预存款余额）：预存款余额反映了客户的财务状况和对电信服务的投入程度。余额较低的客户可能更容易因经济压力或其他服务商的优惠而流失。另一方面，高余额的客户可能对当前服务较为满意，或者有长期使用的打算，因此流失风险较低。

特征5（近3个月计费收入较上月的变化）：计费收入的变化直接反映了客户消费行为的波动。如果收入下降，可能意味着客户减少了通话时间、短信发送或数据使用量，这可能是流失的前兆。收入增加则可能表明客户对当前服务满意，正在增加使用量或升级服务，流失风险较低。

特征7（近4周的流量使用量波动）：流量使用量的波动可能反映了客户对移动数据服务的需求变化。如果使用量显著下降，可能意味着客户



（a）SHAP摘要



（b）CatBoost特征重要性排序

图7　模型的可解释图

找到了其他满足其需求的方式，如使用 Wi-Fi，或者转向其他运营商。使用量稳定或增加则表明客户对当前的移动数据服务较为依赖，流失风险较低。

而从图7（b）看出在网时长（IN_MONTHS）、近一月的语音总量（ALL_CALL_DUR）/Min、近一月的流量总量（ALL_FLUX_USE）/MB 对客户是否流失具有重要的影响。

进一步地，本研究选择了在网时长（IN_MONTHS）、近一个月的通话总时长（ALL_CALL_DUR）/Min 和近一个月的数据流量总量（ALL_FLUX_USE）/MB 这 3 个关键特征，进行特征依赖性分析。SHAP 特征依赖分析如图 8 所示，横轴表示特征的数值，而纵轴代表对应特征的 SHAP 值。从图 8 中可以明显观察到，随着这 3 个特征值的增长，SHAP 值同样呈增长趋势，这表明这 3 个因素在推动客户流失方面起到了积极的作用。

5　结束语

在当前竞争激烈的电信市场中，客户流失预测是客户关系管理的一个重要问题，通过识别类似的客户群体，并为相应群体提供有竞争力的优惠/服务，从而留住有价值的客户。本文针对电信行业客户流失问题，提出一种基于 CatBoost 算法的客户流失预测模型，并使用 SHAP 方法对模型进行解释性分析。该模型以新疆某通信公司数据为基础，通过数据预处理和特征工程，建立一个基于 CatBoost 算法的预测模型，通过对已有文献所提出的主流机器学习模型做对比实验，验证本文算法在客户流失预测的性能优越性，融入 SHAP 可解释性方法，结合实际业务对模型进行全局和局部解释性分析以及特征依赖性分析，并通过对比不同模型的重要性排序，综合分析出了影响客户流失的关键因素，为通信公司提供了应对客户流失挑战的创新解决方案。展现了在数字经济时代，如何通过科技创新推动服务模式变革，以更好地适应市场需求，实现客户价值最大化。在未来工作中，将针对模型的特征选择等方面，提高模型精度，为决策部门提供更进一步的建议。

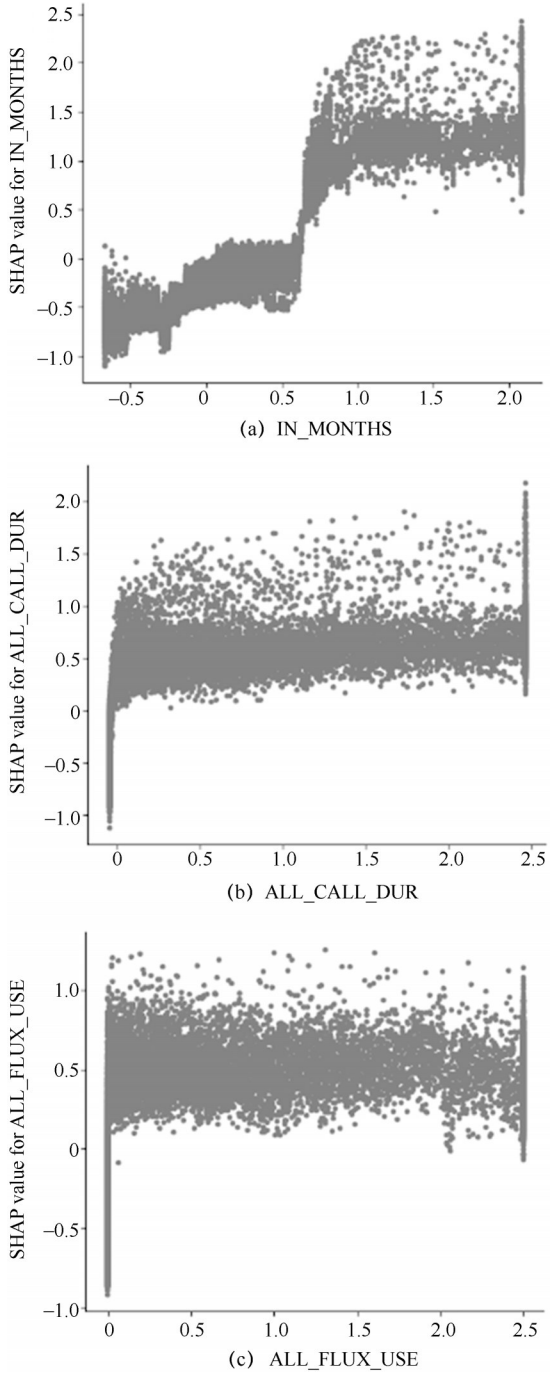


图8 SHAP特征依赖分析

参考文献：

[1] BRANDUSOIU I, TODEREAN G. Churn prediction in the telecommunications sector using support vector machines[J]. Margin, 2013, 1(1):19-22.

[2] SAHAR F. Machine-learning techniques for customer retention: a comparative study[J]. International Journal of Advanced Computer Science and Applications, 2018, 9(2): 273-281

[3] JAIN H, KHUNTETA A, SRIVASTAVA S. Churn prediction in telecommunication using logistic regression and logit boost[J]. Procedia Computer Science, 2020, 167: 101-112.

[4] STEHANI S, KARUNYA N, RANJAN D R J B, et al. Customer churn reasoning in telecommunication domain[C]//Proceedings of the 2020 International Conference on Image Processing and Robotics (ICIP). Piscataway: IEEE Press, 2020: 1-5.

[5] KHAMLICHI F I, ZAIM D, KHALIFA K. A new model based on global hybridization of machine learning techniques for "customer churn prediction" [C]//Proceedings of the 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). Piscataway: IEEE Press, 2019: 1-4.

[6] PAMINA J, RAJA B, SATHYABAMA S, et al. An effective classifier for predicting churn in telecommunication[J]. Jour of Adv Research in Dynamical & Control Systems, 2019 (11): 221-229.

[7] TANG P. Telecom customer churn prediction model combining K-means and XGBoost algorithm[C]//Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). Piscataway: IEEE Press, 2020: 1128-1131.

[8] WU S L, YAU W C, ONG T S, et al. Integrated churn prediction and customer segmentation framework for telco business [J]. IEEE Access, 2021 (9): 62118-62136.

[9] ZHANG T Y, MORO S, RAMOS R F. A data-driven approach to improve customer churn prediction based on telecom customer segmentation[J]. Future Internet, 2022, 14(3): 94.

[10] LALWANI P, MISHRA M K, CHADHA J S, et al. Customer churn prediction system: a machine learning approach[J]. Computing, 2022, 104(2): 271-294.

[11] 汪明达, 周俏丽, 蔡东风. 采用混合模型的电信领域用户流失预测[J]. 计算机工程与应用, 2019, 55(24): 214-221, 270.
WANG M D, ZHOU Q L, CAI D F. User churn prediction in telecom domain using hybrid model[J]. Computer Engineering and Applications, 2019, 55(24): 214-221, 270.

[12] OWCZARCZUK M. Churn models for prepaid customers in the cellular telecommunication industry using large data marts [J]. Expert Systems with Applications, 2010, 37(6): 4710-4712.

[13] SENTHAN P, RATHNAYAKA R, KUHANESWARAN B, et al. Development of churn prediction model using XGBoost - telecommunication industry in Sri Lanka[C]//Proceedings of the 2021 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS). Piscataway: IEEE Press, 2021: 1-7.

[14] SULIKOWSKI P, ZDZIEBKO T. Churn factors identification

from real-world data in the telecommunications industry: case study[J]. Procedia Computer Science, 2021(192): 4800-4809.

[15] ULLAH I, RAZA B, MALIK A K, et al. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector[J]. IEEE Access, 2019(7): 60134-60149.

[16] SHRESTHA S M, SHAKYA A. A customer churn prediction model using XGBoost for the telecommunication industry in Nepal[J]. Procedia Computer Science, 2022(215): 652-661.

[17] PENG K, PENG Y. Research on telecom customer churn prediction based on ga-xgboost and shap[J]. Journal of Computer and Communications, 2022, 10(11): 107-120.

[18] OLIVEIRA G X C. Addressing data imbalance in customer churn prediction: A novel approach for telecommunications companies[D]. Lisboa: Dissertação de mestrado, ISCTE-Instituto Universitário de Lisboa, 2023.

[19] PROKHORENKOVA L, GUSEV G, VOROBEV A, et al. CatBoost: unbiased boosting with categorical features[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 6639-6649.

[20] 贾潇瑶. 融合CatBoost和SHAP的乳腺癌预测及特征分析[J]. 计算机与现代化, 2023(10): 32-38.
JIA X Y. Breast cancer prediction and feature analysis model based on CatBoost and SHAP[J]. Computer and Modernization, 2023(10): 32-38.

[21] 马朔, 李钊, 赵军. 基于CatBoost用信预测模型的TreeSHAP解释性研究[J]. 计算机系统应用, 2023, 32(3): 338-344.
MA S, LI Z, ZHAO J. Research on interpretative TreeSHAP based on CatBoost's credit utilization prediction model[J]. Computer Systems and Applications, 2023, 32(3): 338-344.

[22] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 4768-4777.

[23] 廖彬, 王志宁, 李敏, 等. 融合XGBoost与SHAP模型的足球运动员身价预测及特征分析方法[J]. 计算机科学, 2022, 49(12): 195-204.
LIAO B, WANG Z N, LI M, et al. Integrating XGBoost and SHAP model for football player value prediction and characteristic analysis[J]. Computer Science, 2022, 49(12): 195-204.

[作者简介]

王圣节（1997– ），男，新疆财经大学博士生，主要研究方向为机器学习及其应用、深度学习。

张庆红（1973– ），女，博士，新疆财经大学教授、博士生导师，主要研究方向为统计学。