Advances in Applied Mathematics 应用数学进展, 2024, 13(11), 4896-4907
Published Online November 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.1311471

Hans汉斯

基于Stacking集成学习的电信用户流失预测研究

于荣荣，冯　媛，玄金虎

青岛大学，数学与统计学院，山东　青岛

收稿日期：2024年10月19日；录用日期：2024年11月13日；发布日期：2024年11月21日

摘　要

随着信息化建设的迅速推进，电信市场趋于饱和，如何应对用户流失成为通信运营商亟待解决的问题。本文基于电信用户数据，对用户流失趋势进行了深入预测分析。首先，针对数据缺失进行了填补，并对特征进行编码和衍生，使用SMOTE与Tomek Link技术处理了数据不均衡问题。接着，本文使用随机森林、XGBoost、SVM、逻辑回归、AdaBoost和GBDT六种单一模型分别进行用户流失预测。为了提高预测的准确性和稳健性，本文采用了Stacking多模型融合的方式，模型对比结果表明，第二层模型选用SVM达到了最高的准确率(0.8645)，各项指标均优于单一模型。研究证明，Stacking集成模型在用户流失预测中具有较高的有效性，并通过分析识别了影响用户流失的关键因素，为电信运营商提供了减少客户流失的针对性建议，进而提升企业收益和利润。

关键词

用户流失预测，Stacking模型，电信运营商，机器学习

Research on Telecom Customer Churn Prediction Based on Stacking Ensemble Learning

Rongrong Yu, Yuan Feng, Jinhu Xuan

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Oct. 19th, 2024; accepted: Nov. 13th, 2024; published: Nov. 21st, 2024

文章引用: 于荣荣, 冯媛, 玄金虎. 基于 Stacking 集成学习的电信用户流失预测研究[J]. 应用数学进展, 2024, 13(11): 4896-4907. DOI: 10.12677/aam.2024.1311471

Abstract

With the rapid advancement of information technology, the telecommunications market is becoming increasingly saturated, making customer churn a critical issue that telecom operators must address urgently. This paper conducts an in-depth predictive analysis of customer churn trends based on user data from Telecom. Initially, missing data was imputed, and feature encoding and derivation were performed. The SMOTE and Tomek Link techniques were employed to address the problem of data imbalance. Following this, six individual models—Random Forest, XGBoost, SVM, Logistic Regression, AdaBoost, and GBDT—were used to predict customer churn. To improve the accuracy and robustness of the predictions, this study applied the Stacking ensemble learning approach. The model comparison results indicate that the second-layer model using SVM achieved the highest accuracy (0.8645), with performance metrics surpassing those of the individual models. The study demonstrates the effectiveness of the Stacking ensemble model in predicting customer churn and identifies the key factors influencing churn through detailed analysis. These findings provide telecom operators with targeted recommendations to reduce customer churn and enhance corporate revenue and profitability.

Keywords

Customer Churn Prediction, Stacking Model, Telecom Operators, Machine Learning

1. 引言

随着手机网络的普及和市场逐渐饱和，电信行业的竞争日益激烈。尤其在国家出台携号转网政策后，用户更换运营商变得更加便捷，这进一步加剧了用户流失的现象。用户流失直接影响电信运营商的收入，增加了营销成本，同时也给市场竞争力带来了挑战。因此，如何有效地预测用户流失，并采取相应的挽留策略，成为了电信运营商亟需解决的关键问题。

电信运营商积累了大量的用户数据，这为预测用户流失提供了重要的基础。通过对用户历史数据的分析，可以建立预测模型，精准识别出可能流失的用户，并找出影响流失的主要原因[1]。利用这些信息，电信运营商可以制定更具针对性的营销策略，提高用户的满意度和忠诚度。国内外方法大多是基于模拟数据和实际数据，依据不同用户的特征建立的，而后将新用户的画像特征、消费特征等个人特征带入模型进行流失率预测或者量化评分。周支立等人通过对流失用户的特征进行探索，分析了不同类型流失用户之间的联系，并从特征和消费行为的角度对流失用户与正常用户进行了描述和分析，试图找出用户流失的规律[2]。黄宝凤等人提出了基于特征工程的方法，并通过组合模型创建新的衍生变量，结果表明这种方法能够在一定程度上提升模型的预测性能[3]。随着机器学习算法的发展，研究人员在流失用户预测领域提出了各种方法，主要采用传统的机器学习技术。虽然这些方法在一定程度上取得了成果，但也存在局限性，因此研究人员正在努力探索更准确、可靠的预测方法，以推动该领域的发展。

本文使用电信用户数据，结合机器学习算法，通过数据处理、特征选择以及 Stacking 集成学习的方法，构建了一个更为准确的用户流失预测模型。该模型比传统机器学习模型更加准确，可以更加有效预测用户流失，还能识别影响流失的关键变量，帮助运营商更好地制定挽留措施，从而提高市场竞争力，

减少用户流失对运营商带来的负面影响。

2. 数据和方法

2.1. 数据集

本文实验数据源于 iDataScicence 平台上的某电信公司的数据集，该数据集公开可用，包括 5986 条数据，每条数据代表一名用户并包含 20 个特征属性。特征属性由用户的基本信息、服务信息和合同信息三大块组成。其中基本信息包括性别、是否退休、是否已婚等，服务信息包括是否开通电话服务、是否开通电视服务等使用的服务，合同信息包括合同类型、签订年限、费用等情况，具体数据介绍见表 1。

Table 1. Telecom user data description
表 1. 电信用户数据介绍

变量名称	变量解释	变量名称	变量解释
customerID	客户 ID	OnlineBackup	是否激活了在线备份服务
gender	性别	DeviceProtection	客户是否有设备保险
SeniorCitizen	是否退休	TechSupport	是否已连接技术支持服务
Partner	是否已婚	StreamingTV	是否已连接流媒体电视服务
Dependents	是否有家属	StreamingMovies	是否已激活流媒体影院服务
tenure	注册月数	Contract	合同类型
PhoneService	是否已连接电话服务	PaperlessBilling	是否使用无纸化计费
MultipleLines	是否已连接多条电话线	PaymentMethod	付款方式
InternetService	客户端的 Internet 服务提供商	MonthlyCharges	每月付款金额
OnlineSecurity	是否已连接在线安全服务	TotalCharges	服务期内支付的总金额

2.2. 数据处理

首先处理缺失值，发现仅有 TotalCharges 存在缺失值，查看其他数据发现是因为其 tenure 为 0，即刚注册新用户，所以将总金额赋值为月度付款金额。

Table 2. Assignment rules for tenure_year
表 2. tenure_year 赋值规则

tenure (注册月数)	tenure_year
[0, 12]	1
(12, 24]	2
(24, 36]	3
(36, 48]	4
(48, 60]	5
(60, 72+]	6

然后将类别型特征转换为数值型特征，对数据进行编码，将 Yes 替换为 1，No 替换为 0，对于 Contract、InternetService、PaymentMethod 这类含有多种结果但无序的变量，选择 One-Hot 编码，避免数值排序的错误解读同时避免引入不必要的权重偏差。

接下来是特征衍生，通过原有的数据进行新特征的构建。针对各项服务开通数量构造新变量 NumServices，用于统计每个用户开通的服务数。对于 tenure 构造新特征 tenure_year，规则如表 2 所示。

将客户 ID 删除后我们特征变量变为 28 个，客户流失比例如图 1 所示，其中流失用户数量为 1587 占比 26.51%，数据样本分布不平衡，在后续预测中，即使所有都预测为非流失用户，其正确率都将达到 73%左右，因此我们需要对数据进行处理。

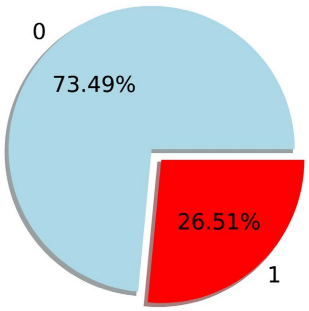

Figure 1. Customer churn rate
图 1. 客户流失比例

为解决少数类样本过少导致的分类不平衡问题，主要是通过过采样跟欠采样两种技术进行均衡样本。SMOTE (Synthetic Minority Over-sampling Technique)是一种过采样技术，通过在少数类样本之间生成新的点(即通过插值)来扩展少数类样本的分布，这使得少数类样本在特征空间上更加分散，从而有助于模型更好地学习[4]。Tomek Links 是一种欠采样技术，主要用于移除数据集中那些“容易混淆”的样本。Tomek Links 的定义为：在不同类别的样本对之间存在彼此为最近邻的关系[5]。如果样本 A 属于多数类，而样本 B 属于少数类，且这两个样本为最近邻，则称这两个样本构成了一个 Tomek Link。在这种情况下，Tomek Links 会删除多数类样本(即 A)，从而优化数据集。如果仅使用 SMOTE 技术进行扩充，可能会因为分类边界不清晰导致模型在训练时表现不佳，因此我们采用 SMOTE 和 Tomek Link 结合的技术，先用 SMOTE 技术进行扩充，然后使用 Tomek Links 删除容易混淆的多数类样本，使得分类决策边界更加清晰，最终得到我们的数据分别为 4399 和 4124。

2.3. 分类方法

2.3.1. SVM

SVM (Support Vector Machine)通过寻找一个最佳的超平面来最大化不同类别之间的间隔，从而实现分类的目的[6]。该算法在高维空间中表现优异，并且能够有效处理非线性分类问题。当样本数据线性可分时，选择惩罚参数，构造凸二次规划问题，寻找两类样本的最优分类超平面；当样本数据线性不可分时，通过非线性变换可以将数据特征转化到高维特征空间，使其线性可分。

2.3.2. 随机森林分类器

随机森林是一种集成学习方法，由许多独立的决策树组成，每棵树在训练时都会从训练数据集中随机选择一个子集(使用 Bootstrap 方法，即有放回抽样)，并在这些样本上进行训练[7]。随机森林将所有树的结果进行投票表决，选择出现次数最多的类别作为最终分类结果。此外，在构建每棵树时，每个节点

的分裂也只在随机选择的一部分特征上进行选择。这两个随机性的引入，使得随机森林具有较高的泛化能力和鲁棒性。

2.3.3. XGBoost

XGBoost (Extreme Gradient Boosting)是一种基于 GBDT 的模型，该算法通过在梯度提升框架中改进损失函数、正则化、切分点搜索和并行结构，显著提高了模型训练的速度[8]。与传统的 GBDT 方法不同，XGBoost 采用了目标函数的二阶泰勒展开技术，这使得在计算过程中，每个数据点只需要计算一阶和二阶导数，从而有效地提升了算法的并行处理能力。目标函数值越小，则树结构越好。

2.3.4. AdaBoost

自适应提升算法(AdaBoost)是一种通过调整训练样本权重来学习多个弱学习器，并将其组合成一个强学习器的集成算法[9]。AdaBoost 可以使用任意单模型作为弱学习器。其主要步骤为：

1) 初始化样本权重：每个样本的初始权重相同，通常为 1/N，其中 N 是样本总数。

2) 训练弱分类器：在每一轮迭代中，根据当前样本权重，训练一个弱分类器。

3) 调整权重：如果某个样本被错误分类，则增加其权重，使得下一轮训练的弱分类器更关注这个样本。如果某个样本被正确分类，则降低其权重。

4) 计算分类器的权重：根据弱分类器的分类错误率，计算该分类器的权重。错误率越低，该分类器的权重越高。

5) 最终分类：所有弱分类器的加权投票决定最终的分类结果。

2.3.5. GBDT

GBDT (Gradient Boosting Decision Tree)，即梯度提升决策树，是一种迭代的机器学习算法。它通过构建一系列弱学习器(即决策树)，并将它们的预测结果相加得到最终输出[10]。这种算法将决策树和集成学习思想有效结合，以提高预测准确性：

$$F_M(x)=\sum_{m=1}^{M}T(x;\Phi_m)$$

决策树用 $T(x;\Phi_m)$ 表示，其中 Φ_m 为决策树的参数，M 为决策树的个数。

2.3.6. Stacking 集成学习模型

Stacking 是一种集成学习技术，近年来在人工智能领域获得了广泛的关注。这种方法通常包括多个层级的嵌套模型[11]。在第一层，称为特征学习层，多个不同的模型如支持向量机、随机森林和逻辑回归等被并行使用。这些模型的预测输出被合并成新的特征集，供最终层的模型使用。最后一层模型则利用这些特征集进行训练，以便对给定的数据标签作出最终预测，形成一个基本的 Stacking 集成学习架构。图 2 是其学习示意图。

其中第一层的模型 1-n 是比较灵活的，可供我们选择的较多，然后每个模型输出值作为模型 m 的输入，模型 m 选择也较为灵活，然后多个模型的输出最终汇总成一个输出。以本文二分类为例，将训练集分别导入前 n 个模型，我们得到每个模型判断是否流失的概率，然后将所有生成的概率放进模型 m 中进一步分类，然后得到我们最终的结果。

Stacking 方法在第一层采用高拟合度的训练模型，如 XGBoost 和随机森林，以确保新生成的特征集能够有效地代表原始训练数据的特征。第一层包括多种不同的模型，每种模型的计算原理不同，旨在自动且有效地提取数据中的非线性特征。但是可能导致过拟合，因此通常在第二层采用较为简单的模型进行预测。

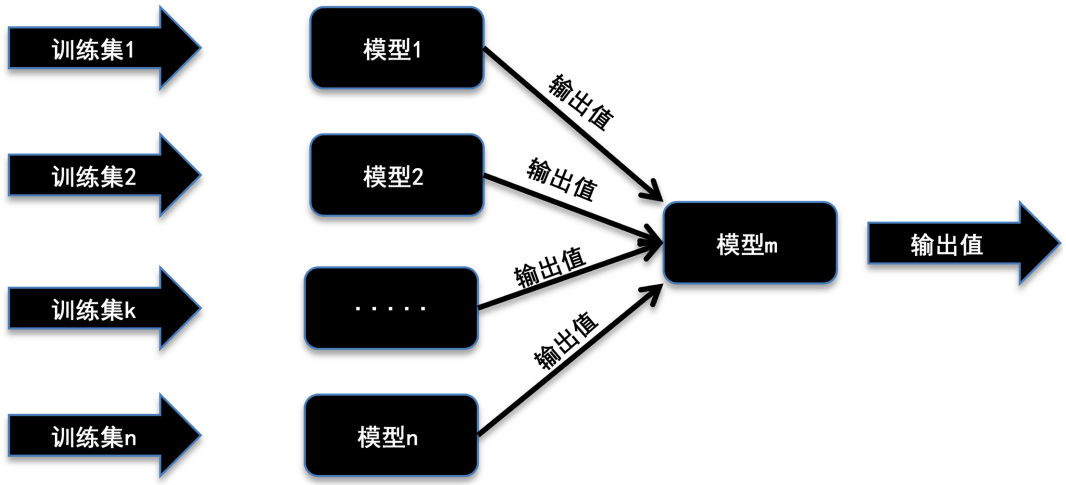

Figure 2. Stacking ensemble learning framework
图 2. Stacking 集成学习框架

2.4. 评价指标

令：真正的阴性，即被正确识别为阴性的个体；False Negatives (FP)：假阳性，即被错误识别为阳性的实际阴性个体；True Positives (TP)：真正的阳性，即被正确识别为阳性的个体；False Negatives (FN)：假阴性，即被错误识别为阴性的实际阳性个体[10]。则有：

准确率(Accuracy)：

$$\text{Accurary} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

精确率(Precision)：

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

特异性(Specificity)：

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

召回率(Recall)：

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

ROC 曲线和 AUC：ROC (Receiver operating characteristic)曲线图横轴为 FPR (False positive rate)，FPR 越大，在预测类别为正类时，实际类别为负类的数据越多。纵轴为 TPR (True positive rate)，TPR 越大，在预测类别为正类时，实际类别为正类的数据越多。即期望 TPR = 1，FPR = 0。故 ROC 曲线越靠拢(0, 1)点，越偏离 45˚对角线越好。AUC (area under the curve)指 ROC 曲线下方的面积，期望 AUC 值越大越好。

3. 电信流失预测与因素分析

3.1. 特征选择

依次分析不同特征与流失率的关系，根据上文已知整体的流失率是 26.525%，图 3 展示了几个典型特征的客户流失率类别柱状图。

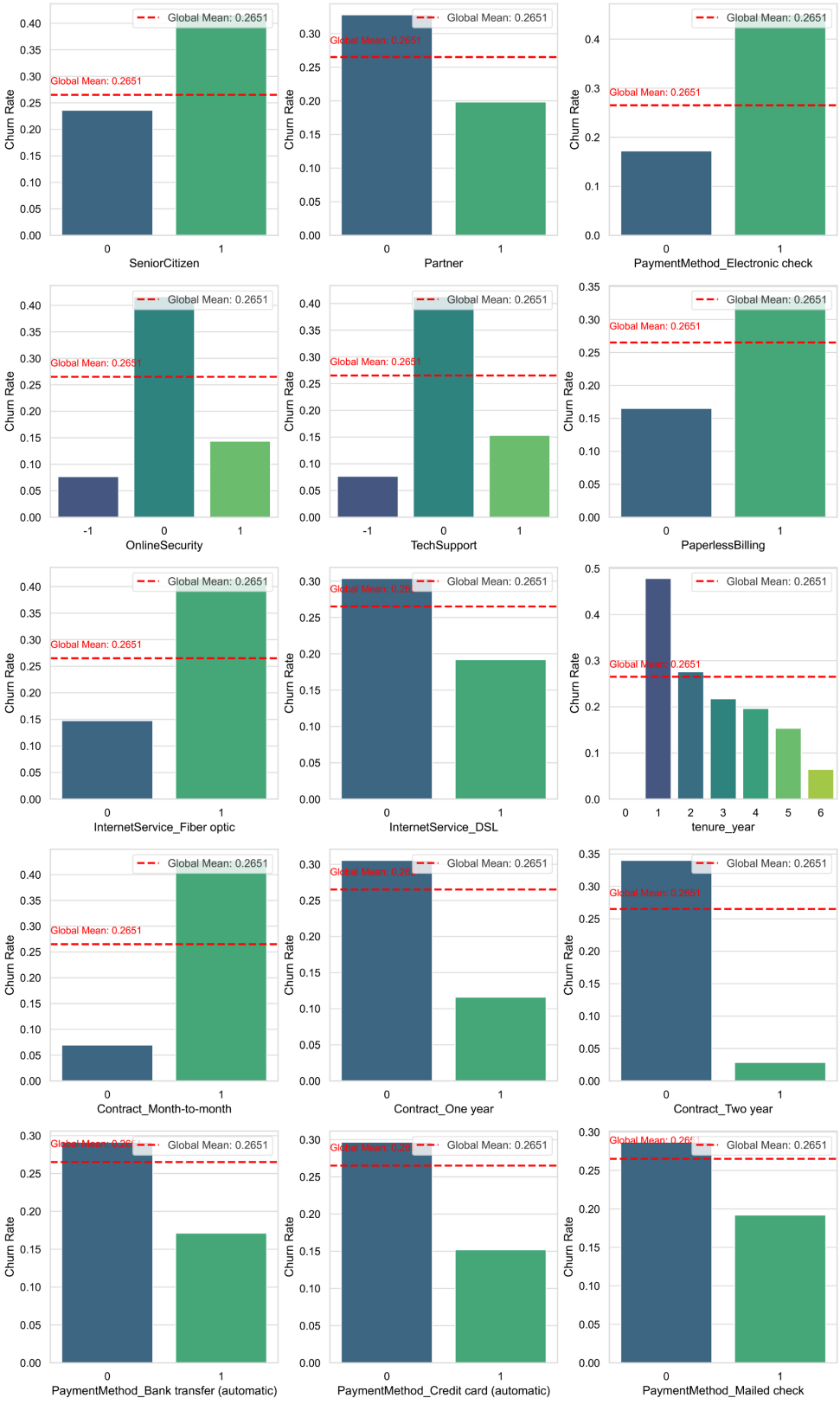

Figure 3. Bar chart of customer churn rate categories
图 3. 客户流失率类别柱状图

从图 3 中可以看出，老年客户(SeniorCitizen)以及选择月度合同(Contract_Month-to-month)的客户流失率显著高于全局平均值，表明这两类群体是流失的高风险人群。使用电子支票(Electronic check)支付的客户流失率也较高，而使用银行转账或信用卡自动支付的客户流失率较低。此外，拥有伴侣(Partner)和家属(Dependents)的客户流失率较低，表明家庭支持可能有助于客户留存。对于服务特征，未使用在线安全(OnlineSecurity)、技术支持(TechSupport)等服务的客户流失率明显高于使用这些服务的客户，表明提供更多增值服务可以有效降低客户流失率。同时，随着客户服务年限(tenure_year)的增加，流失率逐渐下降，反映出新客户在初期几年内更容易流失。因此，企业可以通过引导客户选择长期合同、提供更多增值服务和优化支付方式等策略，来减少客户流失风险。

3.2. 简单预测模型的评价与分析

在排除掉不相关特征后，数据集被划分为训练集和测试集，其中 80%为训练集。使用训练集进行 k 折交叉验证，并分别应用随机森林、GBDT、AdaBoost、XGBoost、SVM 和逻辑回归分类模型来预测用户流失。通过网格搜索找到每个模型的最优参数。随后，表 3 对这六个模型在数据集上的五项评价指标进行了综合对比。

Table 3. Performance metrics of six models on customer churn prediction
表 3. 六种模型在客户流失预测中的性能指标

模型	准确率	精确率	召回率	特异性	AUC
随机森林	0.853916344	0.856344562	0.853916344	0.822159091	0.927287712
GBDT	0.836167954	0.837183134	0.836167954	0.818465909	0.918618302
AdaBoost	0.812995957	0.81482555	0.812995957	0.786647727	0.894795489
XGBoost	0.842917125	0.844034189	0.842917125	0.824147727	0.925309445
SVM	0.832208902	0.832423954	0.832208902	0.829261364	0.91153334
逻辑回归	0.831914894	0.832325355	0.831914894	0.824147727	0.918217932

我们对比了六种常见分类模型在用户流失预测任务中的表现，包括随机森林、GBDT、AdaBoost、XGBoost、SVM 和逻辑回归。结果显示，各模型在准确率、精确率、召回率、特异性和 AUC 等评价指标上表现相对均衡。其中，随机森林、XGBoost 和 GBDT 在准确率和 AUC 方面表现尤为突出，特别是随机森林和 XGBoost 的 AUC 均超过 0.925，表明这两种模型在区分正负类样本时具有较高的辨别能力。同时，SVM 和逻辑回归在特异性上表现略优于其他模型，特异性均为 0.82，显示它们在识别负类样本时有一定优势。相比之下，AdaBoost 的表现相对较弱，特异性仅为 0.79，AUC 也低于其他集成模型，说明其在该任务中的区分能力不足。总体而言，各模型的表现相对均衡，但没有单一模型在所有指标上均表现突出。因此，通过模型融合以整合各模型的优势，特别是在提高准确率和 AUC 方面，可能会显著提升整体分类性能。

3.3. Stacking 集成学习模型

我们利用以上六个模型的输出概率(为 1 的概率)作为第二层的输出，其中生成的新指标，如表 4 所示。

Table 4. The new metrics generated by individual models
表 4. 单一模型产生的新指标

随机森林	GBDT	Adaboost	XGBoost	SVM	逻辑回归
0.991475412	0.927301714	0.506394948	0.940501451	0.976957867	0.999697396
0.004	0.004610408	0.483397018	0.005830545	0.063165375	0.015630914
0.945629784	0.955006218	0.50693327	0.97356385	0.845409893	0.877307074
0.90569697	0.558104939	0.499092915	0.739449859	0.216911425	0.380050599
0.839789955	0.841635837	0.503271716	0.823562562	0.921073626	0.996355955

将上述指标带入第二层模型中，我们得到表 5 的结果。

Table 5. Evaluation and comparison of Stacking model fusion
表 5. Stacking 模型融合评价比较

模型	accuracy	precision	recall	f1	auc
SVM	0.864516129	0.839224629	0.891041162	0.864357017	0.927457186
随机森林	0.863343109	0.854241338	0.865617433	0.859891762	0.927845587
GBDT	0.861583578	0.858880779	0.85472155	0.856796117	0.918905481
Naive Bayes	0.860997067	0.839677047	0.881355932	0.860011813	0.935658367
XGB	0.859237537	0.860837438	0.846246973	0.853479853	0.924416779
LDA	0.854545455	0.822544643	0.892251816	0.855981417	0.923063574
逻辑回归	0.854545455	0.822544643	0.892251816	0.855981417	0.939420484

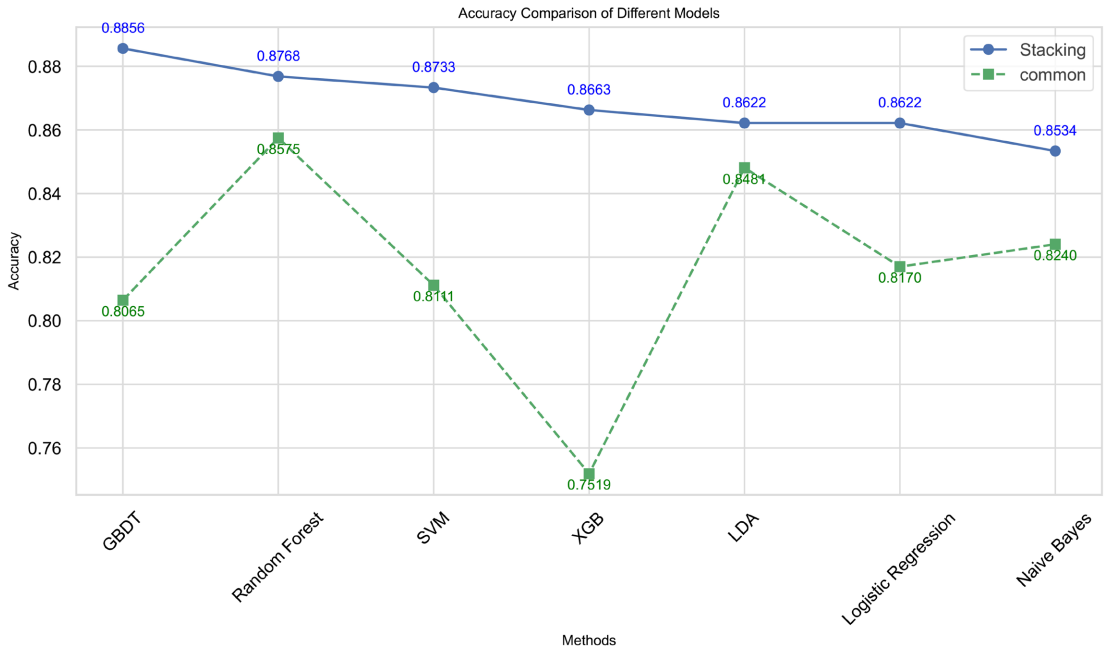

Figure 4. Accuracy comparison between Stacking and individual models
图 4. Stacking 模型与单一模型的准确率比较

在使用 Stacking 集成方法后，各模型的表现都有所提升。尤其是 SVM 和随机森林的表现显著提高，SVM 的准确率从 0.806 提升到 0.864，而随机森林的准确率也从 0.857 增加到 0.863。总体来看，Stacking 模型的各项指标都优于单一模型，说明通过模型融合，集成方法能够有效地结合各个模型的优点，提升整体的分类效果。在 Stacking 模型中，SVM 表现出了最高的准确率(0.864)、较高的精确率(0.839)以及 AUC (0.927)，说明它在集成方法中的表现优于其他模型，也符合我们要求第二层模型尽量选择简单模型避免过拟合。图 4 也可以看出我们 Stacking 的集成模型的准确率均优于单一模型。

3.4. 影响因素

在机器学习中，特征重要性(Feature Importance)是一种评估各个特征对模型预测结果影响力的技术。这通常是通过分析决策树在使用某个特征进行分裂时，分裂前后信息增益的大小来确定的，信息增益较大的特征被视为更重要，因为它们在模型决策过程中起到了关键作用。特征重要性的计算为我们提供了一个有力的工具，可以从定量的角度解释模型的预测结果，明确指出哪些特征在预测过程中具有较大的影响力。这不仅有助于优化模型，减少无关特征的干扰，还可以增强模型的解释性和透明度。图 5 给出特征重要性，这将帮助我们理解哪些特征对预测结果具有决定性的作用，进而优化模型结构和参数设置，以提高预测准确性和模型的泛化能力。

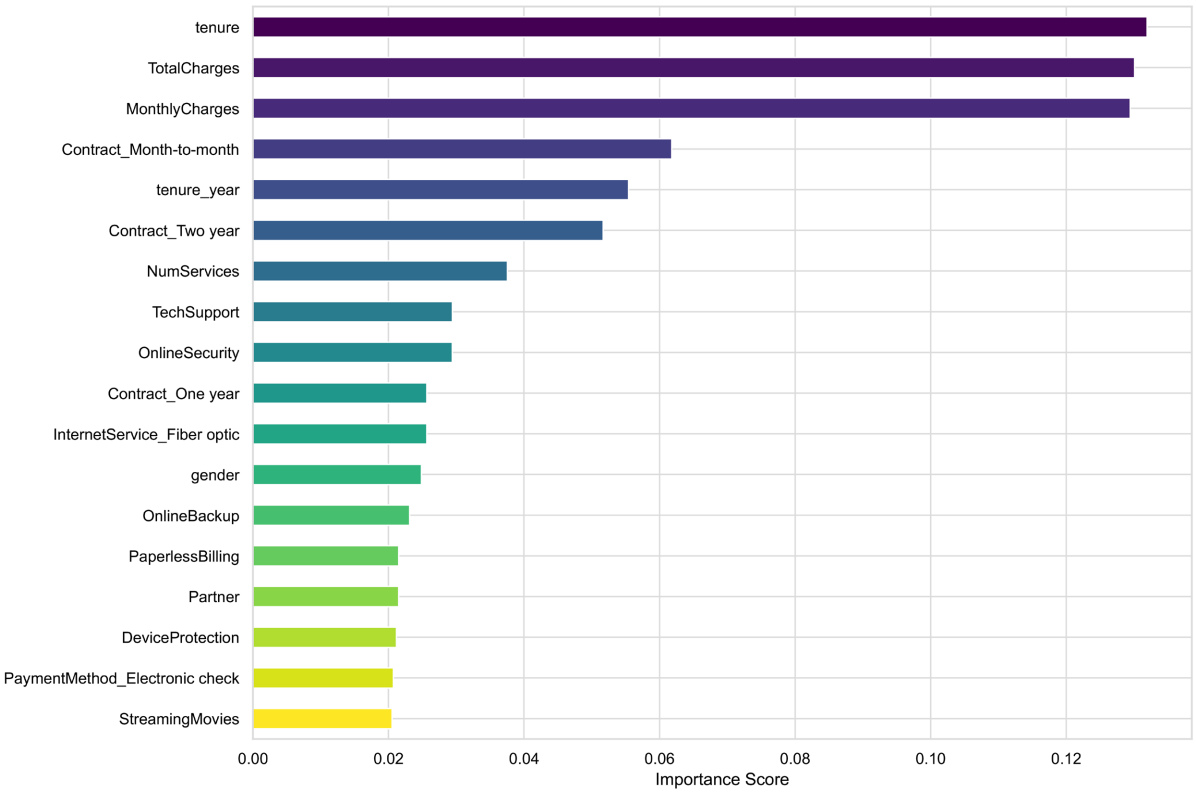

Figure 5. Feature importance ranking
图 5. 特征重要性排序

从特征重要性排序图可以看出，服务时长(tenure)是客户流失预测中最重要的特征，表明客户使用服务的时间越长，流失的风险越低。同时，总费用(TotalCharges)和月费用(MonthlyCharges)也是关键特征，显示出消费金额对客户流失的显著影响。关于合同类型，月度合同(Contract_Month-to-month)客户的流失

率较高，而签订长期合同的客户(如一年或两年合同)流失风险相对较低。此外，技术支持(TechSupport)和在线安全(OnlineSecurity)等增值服务在减少流失方面也起到了重要作用，而性别、流媒体服务(StreamingMovies)等特征对预测的影响较小。

基于上述分析，我们给电信公司提出如下建议：首先重点关注高流失率的客户群体，特别是老年人、无伙伴和无家属的客户，他们的流失率较高。通过定期电话回访、感恩回馈、免费抽奖和赠送礼品卡等关怀活动，增强客户与产品的联系。其次，适当下调光纤网络的定价并提升服务质量，优化流媒体电影和电视服务，增加片源和改善画质，以减少客户流失。同时，调整套餐设计，增加 3 个月和半年等灵活期限的选择，并对长期套餐给予折扣，提供多元化选择。此外，针对新用户，建议降低首月开通费用并绑定自动续费，以提高客户留存率。

4. 结论

本文基于电信用户相关信息，使用机器学习模型，对电信用户流失趋势进行了预测与分析。用户流失将直接影响电信企业的收入，增加营销成本，同时也给市场竞争力带来挑战。本文了构建了包括 AdaBoost、随机森林、XGBoost 和 GBDT 等 6 种常见的机器学习模型。它们在用户流失预测任务中均表现出较好的效果，但是这些模型不够稳健，因此我们提出了使用 Stacking 集成学习的方式来构建模型，这样会比简单模型更加稳健，同时，能够避免选择很差的模型应用于预测，它们提高了预测能力较差的模型的预测能力，可以避免在简单的模型选择中会出现的遗失有用信息、目标偏离等缺点。Stacking 集成模型的各项指标明显优于传统单一模型，表明它在处理用户流失数据时的有效性和准确性。此模型适用于长期预测用户流失趋势，并能有效辅助电信运营商制定针对性的用户留存策略。

通过模型分析，得出影响电信用户流失的关键因素，包括合同类型、使用的增值服务(如技术支持、在线安全)、月度和总消费金额等。月度合同用户流失率显著高于长期合同用户，技术支持和在线安全服务的缺失也显著增加了流失风险。针对这些，给运营商提出了相应的建议，可以进一步优化用户体验、调整合同设计和提供更多增值服务等来减少客户流失，进而提升运营商的收益和利润。

致　谢

感谢所有在研究过程中曾经帮助过我们的良师益友，以及在设计中被我们引用或参考的论著的作者。正是有了他们的悉心帮助和支持，我们的研究工作才能顺利完成。

参考文献

[1] Ganesh, J., Arnold, M.J. and Reynolds, K.E. (2000) Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers. *Journal of Marketing*, **64**, 65-87. https://doi.org/10.1509/jmkg.64.3.65.18028

[2] 周支立, 刘斌. 基于客户信息的电信企业客户流失问题分析[J]. 情报杂志, 2003, 22(12): 97-99.

[3] 黄宝凤, 祁婷婷. 基于特征工程的个人信用风险评估组合模型[J]. 中国统计, 2021(6): 37-39.

[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] Tomek, I. (1976) Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, **6**, 769-772.

[6] Joachims, T. (1998) Making Large-Scale SVM Learning Practical. Technical Report.

[7] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[8] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[9] Freund, Y. and Schapire, R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to

Boosting. *Journal of Computer and System Sciences*, **55**, 119-139. https://doi.org/10.1006/jcss.1997.1504

[10] 王贝伦. 机器学习[M]. 南京: 东南大学出版社, 2021: 187-244.

[11] Wolpert, D.H. (1992) Stacked Generalization. *Neural Networks*, **5**, 241-259. https://doi.org/10.1016/s0893-6080(05)80023-1