基于BO-Stacking集成学习的客户流失预测

耿 宇

(安徽建筑大学数理学院，合肥 230601)

摘要： 为了提高客户流失预测的准确性，提出一种基于贝叶斯优化算法(BO)的改进Stacking集成学习方法。首先，依据模型的预测性能和相关性确定基学习器的种类；然后，针对传统的Stacking方法中忽略基学习器间差异性的缺陷，引入贝叶斯优化算法来精细地调整各基学习器的权重，以降低预测误差；最后，将各基学习器的预测结果进行加权组合，并选用Logistic回归作为元学习器进行最终预测。结果显示，相较于单一模型和传统的Stacking方法，所提出的BO-Stacking模型在召回率、F1-score和AUC(敏感度曲线下方的面积)上均表现最佳，验证了所提方法的有效性，可为企业制定有效的客户保留策略提供参考。

关键词： 贝叶斯优化算法(BO)；Stacking算法；集成学习；客户流失预测

中图分类号： F274 **文献标志码：** A **文章编号：** 1671－1807(2025)13－0241－05

在网络信息迅速发展的时代，世界经济正快速向全球化和市场化迈进，各行各业的竞争日益激烈，电信行业也不例外，正面临着较高的客户流失率所带来的严峻挑战，解决客户流失问题已成为当务之急。因此，利用客户的历史交易数据进行流失预测，对电信行业的持续发展至关重要。

近年来，国内外学者对客户流失的预测进行了广泛研究。周艳聪和郝园媛[1]使用BP神经网络构建客户流失预警模型，发现通过参数调优和结构调优后的模型的预测准确率更高。Zhang等[2]以中国三大电信运营商为样本，使用Fisher判别方程与Logistic回归建立流失预测模型，结果表明Logistic回归模型具有较高的预测精度。然而，这些研究主要依赖单一机器学习模型进行预测，可能因数据的随机性和复杂性而导致泛化能力不足。集成学习是一种基于机器学习的方法，它通过组合多个基本算法显著提高了预测精度和泛化能力。Swetha和Dayananda[3]提出了一种改进的XGBoost(极限梯度提升算法)模型用于预测流失客户。Senthan等[4]也证实了XGBoost模型在处理复杂数据集时的优越表现。杨洪岩[5]结合RF(随机森林算法)和XGBoost等算法，通过Voting融合方法提高了预测的准确性。王变霞[6]将RF和CatBoost(类别型特征梯度提升算法)等算法作为基学习器，将Logistic回归作为元学习器，构建Stacking模型，实现了对银行客户流失的预测。刘梅等[7]提出了一种基于主成分分析(principal component analysis，PCA)和改进灰狼优化算法(grey wolf optimization，GWO)的集成特征选择方法，并结合模型堆叠构建客户流失预测模型，结果显示该方法显著提升了预测性能。

尽管现有的集成学习模型在客户流失预测方面已取得显著成就，但它们往往未能充分考虑基学习器间的差异性对最终预测结果的影响。为此，本文提出一种改进的Stacking集成学习预测方法，旨在通过优化策略进一步提升客户流失的预测精度。具体而言，该方法通过相关性分析选择差异较大且预测效果较好的模型作为基学习器，并引入贝叶斯优化算法，实现基学习器权重的动态优化，以提高预测效果和泛化能力，从而更好地满足电信行业对客户流失精准预测的需求。

1 理论基础

1.1 贝叶斯优化算法

贝叶斯优化算法(Bayesian optimization，BO)[8]是一种基于序列模型的优化算法，通过构建目标函数的概率模型，可以快速搜索出给定参数空间内的最优超参数配置。贝叶斯优化框架主要由两个核心部分组成：概率代理模型(probabilistic surrogate model)和采集函数(acquisition function)，前者用来拟合目标函数，后者根据已知数据递推估算最优值。

收稿日期： 2024-12-28

基金项目： 安徽省高等学校科学研究重点项目(2022AH050247)

作者简介： 耿宇(2000—)，女，安徽滁州人，硕士研究生，研究方向为数据分析中的统计方法及应用。

最常见的概率代理模型有高斯过程(Gaussian process, GP)、随机森林等,本文所使用的代理模型为高斯过程,表达式为

$$f(x) \sim \mathrm{GP}[m(x),k(x,x')] \tag{1}$$

式中:$m(x)$为 $f(x)$的均值函数 $E[f(x)]$;$k(x,x')$为 x 的协方差函数。高斯过程可以看作是对函数的一种先验分布。通过已知的数据点对其进行推断,可以得到一个对函数的后验分布的估计。

最常见的采集函数有概率提升函数 PI、期望提升函数 EI 和置信上界函数 UCB,本文选用的是期望提升函数 EI,表达式为

$$\mathrm{EI}(x)=\begin{cases}[\mu(x)-f(x^+)]\Phi(z)+\sigma(x)\Phi(z), & x<0\\ 0, & x\geqslant 0\end{cases} \tag{2}$$

式中:$f(x^+)$为当前的最大值;$\Phi(z)$为正态累计分布函数;$\mu(x)$和 $\sigma(x)$分别为代理模型在点 x 处预测的均值和标准差。

综上所述,贝叶斯优化在超参数优化中的基本流程如下。

步骤 1:在给定的超参数空间内,随机选取一组初始的超参数向量 $\boldsymbol{X}_{init}$。

$$\boldsymbol{X}_{init}=[x_0,x_1,\cdots,x_n] \tag{3}$$

步骤 2:将超参数向量代入模型中,计算并获取该参数组合下的标签与初始数据集 D_0。

$$D_0=\{X_{init},f(X_{init})\} \tag{4}$$

步骤 3:基于代理模型 $g(x)$,最大化采集函数,从而确定下一个需要评估的超参数点 x_t。

$$x_t=\mathrm{argmax}\alpha(x|D_{t-1}) \tag{5}$$

式中:$\alpha(\cdot)$为采集函数。

步骤 4:获取评估点 x_t 的函数值 $f(x_t)$,并将其加入现有的评估点集合中。

$$D_t=D_{t-1}\cup\{x_t,f(x_t)\} \tag{6}$$

步骤 5:设定一个最大迭代次数,若当前迭代次数达到了预设的最大值,则停止算法迭代,并从所有评估过的超参数组合中选取最优的一组:$\{x^*,f(x^*)\}$。若未达到最大迭代次数,则返回步骤 4 继续进行下一轮迭代。

1.2　基学习器的选择

Stacking 集成学习算法[9]是一种异质集成学习算法,其核心思想在于利用多种不同类型的基学习器对同一数据集进行训练,然后将各基学习器的预测结果按列组合成一个新的数据集,该数据集随后作为元学习器的输入,用于产生最终的预测结果。在 Stacking 方法中,基学习器的选择将直接影响到整体的预测效果。因此,各基学习器在有较高的预测精度的同时还应具有差异性[10]。基于上述考虑,选择梯度提升决策树(GBDT)、极限梯度提升算法(XGBoost)、类别型特征梯度提升算法(CatBoost)和随机森林算法(RF)作为基学习器的预选模型。各算法的优缺点见表 1。

为了选出最优的基学习器组合,首先分析不同模型的预测能力,并通过计算各模型预测结果之间的 Pearson 相关系数来评估它们的相关性,以此量化模型间的差异程度。

表 1　各算法优缺点

算法	优点	缺点
GBDT	高鲁棒性和适应性;通过弱分类器的级联提升整体性能	稳定性较低;难以并行化
XGBoost	支持自动并行计算;通过引入正则化项减少过拟合风险;支持稀疏数据、自定义损失函数等功能	仅接受特定格式的数据作为输入;对噪声敏感
CatBoost	类别特征处理能力优秀;通过排序提升策略减少偏差	解释性较差;训练时间较长
RF	抗过拟合能力强;算法简单易于实现	解释性较差;对噪声敏感;训练速度较慢

1.3　BO-Stacking 模型

传统 Stacking 集成算法将第 1 层基学习器的输出直接输入到第 2 层元学习器中,且每个基学习器被赋予相同的权重。然而,这种做法忽略了基学习器之间的差异性,可能影响模型整体的预测性能。为了解决这一问题,提出一种改进的 Stacking 方法,即 BO-Stacking,该方法根据每个基学习器的性能动态地赋予不同的权值,性能越优的基学习器获得越大的权值。这样做不仅提高了训练元学习器数据的可信度,还增强了模型的分类性能。

为了更好地比较传统 Stacking 模型与 BO-Stacking 模型的预测效果,两种模型均采用相同的结构。构建 BO-Stacking 模型的主要步骤如下。

步骤 1:将数据集按 7∶3 划分为训练集和测试集,在第一层的基学习器中,对选择的 3 个模型进行五折交叉验证训练,得到基学习器的结果。

步骤 2:结合各个基学习器的权重,构建元学习器的加权数据集。具体计算方式如下:

$$\boldsymbol{M}_i=[w_iA_{i1},w_iA_{i2},\cdots,w_iA_{ij}]^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{N}_i=[w_iB_{i1},w_iB_{i2},\cdots,w_iB_{ij}]^{\mathrm{T}} \tag{8}$$

式中:$\boldsymbol{M}_i$ 为第 i 个基学习器在训练集上加权后的预测结果;$A_{i1},A_{i2},\cdots,A_{ij}$ 为第 i 个基学习器的第 j 折交叉验证的训练集输出;w_i 为第 i 个基学习器对应

的权重；$\boldsymbol{N}_i$ 为第 i 个基学习器在测试集上加权后的预测结果；B_{i1}，B_{i2}，…，B_{ij} 为第 i 个基学习器的第 j 折交叉验证的测试集输出。完成五折交叉验证后，获得经加权处理的基学习器的测试数据集 M 和 N。

步骤 3：在第 2 层元学习器中，使用 Logistic 模型作为预测模型，将第 1 层加权的预测结果作为第二层的数据集进行训练，得到最终的预测结果。

采用贝叶斯优化算法计算最优权重，将目标函数设定为 1-AUC（敏感度曲线下方的面积），并设置 500 次迭代，同时采用十折交叉验证的方法。通过不断迭代优化，每个模型的权重均得以动态调整与更新，最终得到使模型在测试集上目标函数最小化，即 AUC 值最大化的权重系数。

1.4　评价指标

1.4.1　混淆矩阵

本文主要研究关于客户流失的二分类问题，对于这类问题，一般使用混淆矩阵作为评价准则，其基本形式见表 2。表 2 中，TP 和 TN 分别表示正确识别的流失客户数和正常客户数，FP 和 FN 分别表示错误识别的正常客户数和流失客户数。本文主要采用准确率、精确率、召回率和 F1-score 作为评价指标，其计算公式如下：

$$\text{准确率(accuracy)}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{9}$$

$$\text{精确率(precision)}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{召回率(recall)}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$\text{F1-score}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{12}$$

精确率（precision）和召回率（recall）呈反向变动关系，因此合理地平衡两者可以实现模型性能的优化。指标 F1-score 综合考量了精确率和召回率，克服了两者反向变动的缺陷。F1-score 的取值为 0～1，值越大，表示模型的分类效果越佳。

表 2　混淆矩阵

变量	预测流失	预测正常	合计
实际流失	True Positive (TP)	False Negative (FN)	TP+FN
实际正常	False Positive (FP)	True Negative (TN)	FP+TN
合计	TP+FP	FN+TN	TP+FN+FP+TN

1.4.2　ROC 曲线与 AUC

ROC 曲线又称敏感度曲线，该曲线通过将假阳性率（false positive rate，FPR）和真阳性率（true positive rate，TPR）作为横纵坐标来描绘分类器在不同阈值下的性能，计算方式如下：

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}},\ \text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{13}$$

AUC 表示 ROC 曲线下方的面积，通常其范围为 0.5～1。AUC 越大，表明模型的预测性能越佳。

2　模型建立与结果分析

2.1　数据来源与预处理

使用 IBM 公司提供的 Telecom 企业的开源数据集，该数据集包含 7 043 个样本，共有 3 个连续型变量和 18 个离散型变量，共有 1 869 位流失客户和 5 174 位未流失客户，涵盖了客户个人信息、电信服务使用情况以及其他重要因素。

数据预处理主要分为以下 3 步。

步骤 1：缺失值处理。本文所使用的数据集中，TotalCharges（总花费）变量中有 11 个缺失值，因其对应的客户留存期为 0，表明服务尚未开通，故直接删除这些样本。

步骤 2：连续型变量处理。对连续型变量进行标准化处理，以消除量纲的影响和数值差异所带来的误差。

步骤 3：离散型变量处理。将二分类变量进行字符串编码，将“Yes”转为 1、“No”转为 0；对多类别的无序变量，如 InternetService、OnlineSecurity 等，应用 One-hot 独热编码。

最终将经过缺失值处理、标准化和 One-hot 独热编码后的原始数据集按照 7∶3的比例划分为训练集和测试集，其中训练集有 4 922 条数据，测试集有 2 110 条数据。

2.2　Lasso 回归筛选变量

使用 Lasso 方法筛选出对客户流失状态有显著影响的变量。Lasso 方法实质是构造惩罚项来压缩模型的回归系数，使对模型贡献较小的系数被逐渐压缩到 0，从而筛选出对因变量有重要影响的协变量，最终达到精简模型的效果。假设样本数据为 (x_i,y_i)，$i=1,2,\cdots,n$，Lasso 方法的定义如下：

$$\hat{\beta}=\text{argmin}\left\{\sum_{i=1}^{n}\left[\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right]\right\} \tag{14}$$

式中：y_i 为因变量；$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$，$x_{ij}$ 为第 i 个样本的第 j 个协变量；$\hat{\beta}$ 为需要计算的参数；β_0 为模型的截距项；参数 λ 被用来控制 Lasso 的复杂程度；$\lambda\sum_{j=1}^{p}|\beta_j|$ 为正则化项，也称为惩罚项。

图 1 展示了通过 10 倍交叉验证来确定 λ 最优

取值的过程，纵轴表示二项式偏差，横轴表示 λ 的对数，左侧虚线表示在模型误差最小时取到的调整参数 λ 的对数值 lnλ，右侧虚线表示模型误差最小值的一倍标准误对应的 lnλ。本着用较少变量得到较高准确率的原则，选取右侧虚线对应的 lnλ 值，此时 λ=0.011 3，筛选出的变量个数为 14，对应的系数见表 3。

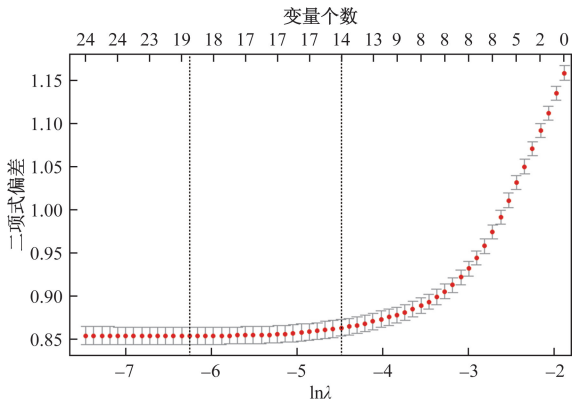

图 1　交叉验证过程

表 3　最优 λ 下的系数

变量	变量含义	系数
SeniorCitizen	是否是老年人	0.099 876 30
Partner	是否使用电子账单	−0.067 173 64
Dependents	有无合作伙伴	−0.020 995 43
TotalCharges	总费用	−1.847 723 47
InternetService_Fiber. optic	互联网服务_光纤线路	0.899 771 20
OnlineSecurity_No	网络安全服务_无	0.495 815 27
OnlineBackup_No	在线备份服务_无	0.126 459 09
DeviceProtection_No	设备保护功能_无	0.030 051 51
TechSupport_No	技术支持功能_无	0.378 167 63
StreamingTV_Yes	流媒体电视功能_有	0.121 028 82
StreamingMovies_Yes	流媒体电影功能_有	0.131 366 03
Contract_One. year	合同期限_1 年	−0.723 544 21
Contract_Two. year	合同期限_2 年	−1.228 023 50
PaymentMethod_Electronic. check	支付方式_电子支票	0.334 607 56

2.3　单一预测模型构建与基学习器选择

对数据建立基于集成学习算法的单一预测模型，利用贝叶斯优化算法(BO)对模型进行超参数调优，各模型的超参数取值见表 4。模型在测试集上的预测性能见表 5。

从表 5 中可以看出，GBDT 模型表现最佳，其 F1-score 和 AUC 分别为 60.06%和 0.821 5，其余模型均具有较好的预测性能，可作为备选的基学习器。

通过相关系数计算进一步筛选基学习器，获得各模型预测结果之间的 Pearson 相关系数，如图 2 所示。从图 2 中可以看出，CatBoost 与 XGBoost、RF 模型间的相关系数均超过了 0.7。鉴于 Stacking 集成方法的效果很大程度上依赖于基学习器的多样性，当基学习器间高度相关时，集成的效果可能会受到限制，因此在选择基学习器时去除一个高度相关的模型有助于提高整体的集成性能。考虑到 XGBoost、RF 和 CatBoost 的 AUC 分别为 0.814 1、0.815 6 和 0.796 0，综合评估相关性和预测性能后，选择移除与其他模型相关性较高且 AUC 较低的 CatBoost 模型。最终，选定了 XGBoost、GBDT 以及 RF 作为 Stacking 框架下的基学习器，以期达到更好的综合预测效果。

表 4　不同优化算法下各模型主要超参数取值

模型	最佳超参数	模型	最佳超参数
XGBoost	n_estimators=100	CatBoost	max_depth=7
	max_depth=9		learning_rate=0.16
	min_child_weight=3		l2_leaf_reg=3
	Subsample=0.8		Iterations=500
	learning_rate=0.54		Subsample=0.6
GBDT	n_estimators=50	RF	n_estimators=50
	max_depth=5		max_depth=9
	Subsample=1		min_samples_split=1
	learning_rate=0.05		min_samples_leaf=2

表 5　各模型的性能指标

模型	准确率/%	精确率/%	召回率/%	F1-score/%	AUC
XGBoost	72.32	48.63	72.55	58.23	0.814 1
CatBoost	72.46	48.73	68.63	56.99	0.796 0
GBDT	73.46	50.06	75.04	60.06	0.821 5
RF	72.46	48.77	70.94	57.81	0.815 6

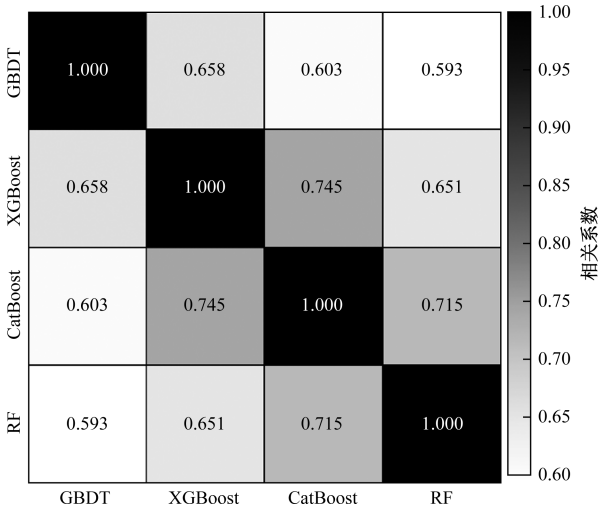

图 2　各模型相关性分析

2.4　BO-Stacking 模型构建与预测结果对比

在选择 XGBoost、GBDT 以及 RF 模型的基础上，构建基于贝叶斯优化算法改进的 Stacking 模型，即 BO-Stacking 模型，各基学习器的权重见表 6。

为了验证本文提出的 BO-Stacking 模型预测性

能的优越性，将该模型分别与单一模型及传统 Stacking 模型进行比较。由表 5 和表 7 可以看出，相比于单一模型，两种融合模型的各项指标均有不同程度的提升。与传统 Stacking 模型相比，尽管 BO-Stacking 模型在准确率和精确率上稍有下降，但在召回率、F1-score 和 AUC 值这 3 个关键指标上都有显著提升。特别是在客户流失预测中，高召回率意味着能够更早、更准确地发现有流失风险的客户，从而帮助企业实施有效的挽留策略。因此，从整体上看，BO-Stacking 模型的预测效果更好，不仅提高了对潜在流失客户的识别能力，还确保了整体预测的准确性和稳定性，为解决客户流失问题提供了更为有效的方法。

表 6　基学习器的权重

基学习器	权重
XGBoost	0.363 7
GBDT	0.501 9
RF	0.134 4

表 7　各 Stacking 模型的评价指标

模型	准确率/%	精确率/%	召回率/%	F1-score/%	AUC
传统 Stacking 模型	73.79	50.48	74.87	60.30	0.825 6
BO_Stacking 模型	73.36	49.94	78.43	61.03	0.834 2

3　结论

基于 IBM 公司提供的电信企业客户流失数据集，针对如何提高客户流失的预测性能，提出了一种改进的 Stacking 集成学习预测模型，即 BO-Stacking 模型。该模型综合考虑了不同单一模型的预测性能和相关性，优选出 XGBoost、GBDT 和 RF 算法作为基学习器。为解决传统的 Stacking 集成方法中因忽略基学习器间差异性而导致预测能力不足的缺陷，采用基于贝叶斯优化算法的动态权重分配策略，实现基学习器权重系数的自适应调整与全局最优配置。研究表明，BO-Stacking 模型在预测效果上具有显著优势，能够为企业提供更精准的客户流失预测，并助力企业制定更为有效的客户挽留策略。

参考文献

[1] 周艳聪，郝园媛. 基于机器学习的运营商客户行为分析[J]. 科学技术与工程，2022，22(2)：585-592.

[2] ZHANG T，MORO S，RAMOS R F. Adata-driven approach to improve customer churn prediction based on telecom customer segmentation[J]. Future Internet，2022，14(3)：1-17.

[3] SWETHA P，DAYANANDA R B. Improvised_XGBoost machine learning algorithm for customer churn prediction[J]. EAI Endorsed Transactions on Energy Web，2020，7(30)：e14.

[4] SENTHAN P，RATHNAYAKA R，KUHANESWARAN B，et al. Development of churn prediction model using XGboost-telecommunication industry in sri lanka[C]//IEEE International IOT，Electronics and Mechatronics Conference. Toronto：IEEE，2021：1-7.

[5] 杨洪岩. 数据挖掘技术在通信用户流失预警中的研究[D]. 沈阳：辽宁大学，2021.

[6] 王变霞. 基于 Stacking 模型融合的银行信用卡客户流失预测[D]. 兰州：兰州大学，2022.

[7] 刘梅，郑立君，段永良，等. PCA+GWO 集成特征选择和模型堆叠的客户流失预测[J/OL]. 计算机工程与应用，1-16[2024-11-05]. http：//kns. cnki. net/kcms/detail/11.2127. TP. 20240823.1756.009. html.

[8] SHAHRIARI B，SWERSKY K，WANG Z，et al. Taking the human out of the loop：a review of Bayesian optimization[J]. Proceedings of the IEEE，2015，104(1)：148-175.

[9] KHOSHKROUD A，SANI H P，AAJAMI M. Stacking ensemble-based machine learning model for predicting deterioration components of steel w-section beams[J]. Buildings，2024，14(1)：1-21.

[10] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

Customer Churn Prediction Based on BO-Stacking Ensemble Learning

GENG Yu

(School of Mathematics and Physics，Anhui Jianzhu University，Hefei 230601，China)

Abstract：To enhance the accuracy of customer churn prediction，an improved Stacking ensemble learning method with Bayesian optimization (BO) incorporated was introduced. First，base learners were selected based on their predictive performance and inter-model correlations. Noticing the fact that the performance variation among base learners was neglected in the traditional Stacking methods，the Bayesian optimization was introduced to fine-tune the weights of each base learner for minimizing prediction errors. Finally，the weighted predictions from the base learners were combined，and the Logistic Regression serves as the meta-learner for the final prediction. The results demonstrate that the proposed BO-Stacking model outperforms both the single models and the traditional Stacking methods in terms of recall rate，F1-score，and AUC(area under the curve) value，which validates the effectiveness of the proposed approach. This provides a reliable reference for enterprises to develop effective customer retention strategies.

Keywords：Bayesian optimization(BO)；Stacking algorithm；ensemble learning；customer churn prediction