DOI: 10.13718/j.cnki.xdzk.2022.06.020

基于营销大数据的电力客户多维度信用评价模型研究

刘翠玲，　胡聪，　王鹏，　洪德华，　张庭曾

国网安徽省电力有限公司 信息通信分公司，合肥 236000

摘要：随着电网客户量的迅速增长，如何对电力客户进行准确的信用评价成为了一个重要问题，构建一个可以准确预测客户信用的模型是电力营销部门需要解决的关键问题. 本文通过对已知信用评价模型的研究，结合集成学习思想，构建了一种基于 XGB 算法的客户多维信用评价模型，该模型通过采用多维度的营销数据，并基于特征重要度方法进行特征选择，采用极限梯度提升方法以及树模型对客户信用进行模型构建，计算不同节点上不同的增益值来获取最佳的预测效果，从而构建一个准确、稳定的客户信用评价模型. 在经过客户历史数据进行模拟分析后，得出了客户信用评价结果，并与目前主流的机器学习算法包括基于梯度下降算法与基于树的传统算法进行比较，结果表明，与 Logistic 回归和其他 3 种基于树的模型相比，XGB 模型不论是特征选择的准确性还是其分类性能都具有明显的优势.

关　键　词：客户信用评价；极限梯度提升；树模型；集成学习

中图分类号：TM734；F407.61　　**文献标志码**：A

文 章 编 号：1673－9868(2022)06－0198－11　　开放科学(资源服务)标识码(OSID)：

Research on Multi-Dimensional Credit Evaluation Model of Electric Power Customers Based on Big Data of Marketing

LIU Cuiling，　HU Cong，　WANG Peng，
HONG Dehua，　ZHANG Tingzeng

Information and Telecommunication Branch，State Grid Anhui Electric Power Co. Ltd.，Hefei 236000，China

Abstract：At present，with the rapid growth of electricity customers，how to make an accurate credit evaluation for customers has become an important issue. Building a model that can accurately predict the credit of customer is a key step for the power marketing department. In this paper，we studied the credit evaluation model proposed by previous authors and combined the integrated learning idea to build a multidimensional credit evaluation model for customers based on XGB algorithm. Based on feature importance method，feature selection was carried out with the model by using multidimensional marketing data. The model for customer credit was built using gradient boosting method and tree model. By calculating the different

收稿日期：2021－09－27
基金项目：五凌电力有限公司综合智慧能源业务及数字化科学技术研究项目(320115JX0120210002).
作者简介：刘翠玲，工程师，主要从事大数据应用分析、数据增值服务的研究.

gain values on different nodes to obtain the best prediction effect, an accurate and stable customer credit evaluation model was constructed. After simulation analysis of customer's historical data, the user credit evaluation results were obtained, which were compared with the current mainstream machine learning algorithms including gradient descent-based algorithms and traditional tree-based algorithms. The experimental study showed that the XGB model has significant advantages over Logistic regression and three other tree-based models, both in the accuracy of feature selection and classification performance.

Key words: customer credit evaluation; XGB; tree models; integrated learning

随着电力体制改革的深入，电力系统由于其复杂的结构受到了多方面的影响. 对于电力营销部门而言，电力市场交易类型更加多元化，随着电力客户的不断增长，如何对电力客户信用进行评价成为电力营销部门的一个关键问题. 因此，电力市场监管机构迫切需要一个准确的方法对各级市场客户进行信用评价，为有效防范市场风险提供决策依据，为保障市场规范运行和健康有序发展提供技术支持.

如何建立一个可靠的电力用户信用评价指标是许多研究者关注的重点. 文献[1]提出了客户信用评价指标体系，包括企业财务状况、支付状况等；文献[2]计算了不同信用类别之间的相关系数；文献[3]基于运筹学中的层次分析法对电力客户的信用风险进行了评价；文献[4]采用熵权方法构建了电力客户风险评价模型.

根据方法和类型不同，信用评价方法大致可以分为以下 3 种：专家系统、统计模型和人工智能方法[5]. 传统方法依赖于大量结构化的历史数据，基于大数据的信用评价方法则通过分析和挖掘海量、多样化的动态数据，然后利用机器学习算法设计信用评价模型，多维度刻画信用主体的“画像”，向信息使用者呈现信用主体信用状况[6]. 因此，由于其高效率、高性能和基于大数据样本的优秀处理能力，许多研究者将各种机器学习算法应用于信用风险预测领域.

基于大数据的信用评价最常用的算法是逻辑回归(Logistic Regression，LR)，因为它结构简单、可解释性、准确性高. 尽管深度学习模型显示出显著的信用评估准确性，但它缺乏可解释性，在处理相对较小的数据集时性能较差，这使得它无法广泛应用于信用评价系统. 除此之外，模糊数学在决策类型的算法中也常常能表现出较好的效果，如文献[7]中采用基于模糊的层次分析对风险进行评估，文献[8]采用模糊数学进行了节点选举的决策. 同时，以决策树(Decision Tree，DT)、随机森林(Random Forest，RF)、梯度提升决策树(Gradient Boosting Decision Tree，GBDT)和极限梯度提升(eXtreme Gradient Boosting，XGB)为代表的机器学习算法在小数据集上具有优越的预测评价效果，在电力客户信用评价方面有广泛应用[9]. 机器学习算法在分类和回归问题上具有良好的性能，并且可以在相对较短的训练时间内获得更好的预测效果. 其中 DT 是预测个人信用的代表性算法，因为它在效率、准确性和可解释性方面表现出显著的优势[10]. 然而，单一的机器学习方法往往会导致过拟合，并且难以处理实际问题中出现的数据集不平衡问题，为了弥补单一机器学习方法的不足，集成学习技术应运而生，并逐渐成为机器学习研究领域的主流方法.

集成学习理论起源于 Kearns 等[11]提出的强学习和弱学习的等价原则，这种原则规定为得到一个优秀的强学习模型，可以将几个简单的弱学习模型结合起来进行综合考虑. 考虑到 DT 算法在各种机器学习算法中性能较好，训练时间相对较短，提出了基于 DT 的集成学习方法，RF、GBDT 和 XGB 就是集成学习和 DT 相结合的典型集成算法. 文献[12]首次结合集成方法、DT 算法和随机子空间方法提出了 RF 算法；文献[13]提出了解决回归和分类问题的 GBDT 算法；文献[14]采用了一种改进的 GBDT 算法和 XGB 算法并与 RF 模型进行了实验对比，结果表明，XGB 算法能够取得更优越的效果. XGB 起源于高效 Boosting 集成学习模型，它利用树分类器获得了更好的预测结果和更高的运行效率，在数据竞赛平台 Kaggle 的竞赛中取得了优异的成绩. 近年来，文献[15]将 LR,DT,神经网络与 XGB 进行了比较，并通过混淆矩阵和蒙特卡罗模拟基准作为评价指标，验证了 XGB 算法在信用评价中的优越性能；文献[16]提出了一种新的集成模型，利用叠加法框架对 XGB、支持向量机(Support Vector Machine，SVM)和 RF 进行建模，实验表明，所提出的集成模型对信用评价方面具有更好的预测能力.

基于以上分析，集成学习理论适用于电力营销部门的多维数据中，采用 XGB 模型能够对客户数据进

行建模并对客户信用情况进行预测. 因此，本文首先针对分类问题对XGB模型进行了重组，并给出了使用XGB模型进行特征选择的理论依据. 随后，本文利用电网营销平台提供的电网客户评估数据进行建模，并采用基于XGB算法的非线性方法代替传统的线性方法对特征选择方法进行改进，结果表明，基于XGB模型的特征选择方法对5种模型都有显著的改进效果. 随后，对XGB,LR,DT,GBDT,RF 5个模型的信用评价效果进行了对比分析，其中XGB模型在电力营销数据中对于信用评价效果具有最优表现.

1 信用评价模型基本概念及关键技术

1.1 信用评价模型基本概念

客户信用是客户与电力营销机构之间的交易关系，是按照约定合同履行相应义务的情况[17]. 而信用是一个无法衡量的概念，为了对信用进行精细的等级划分，就需要一个具有可量化的指标便于管理. 对于电力营销机构而言，为对客户进行信用等级分类，建立一个能够准确分析用户信息，从而构建有效的区分客户信用等级的评估方法，同时对于新加入的客户根据历史客户信息进行划分也是十分重要的.

信用评估模型就是对客户信用水平、客户质量等级进行准确地评级和科学的评估[18]，并以此对不同信用等级客户进行管理的依据，这对于提高电力营销策略制定效果、优化营销效率、规避由于客户的不守约行为而导致的亏损等具有实践性的意义. 通过对相关文献的研究和电力客户历史数据信息，同时根据电力行业客户信用等级划分准则，采用集成学习的思想对电力客户信用级别进行了划分.

1.2 极限梯度提升模型介绍

符号及其含义如表1所示.

表1 符号参考表

符号	代表的含义	符号	代表的含义
m	特征总数	$\hat{y}_i^t$	到第 i 棵树的预测值
x_i	第 i 个样本的特征信息，$x_i \in R^m$	$l(y_i, \hat{y}_i)$	第 i 个样本的损失函数
y_i	第 i 个样本的真实值	$L(y, \hat{y})$	总样损失函数
$\hat{y}_i$	第 ii 个样本的预测结果	(f_k)	目标函数中防止过拟合的正则项，其中 f_k 代表第 k 个决策树的结果

一个数据集中包含 n 个样本，每个样本具有 m 维特征，那么对于信任评估模型可以抽象为 $D=\{(x_i, y_i) \mid x_i \in R^m, y_i \in R^m\}$ 代表样本空间，其中 $x_i=\{x_{i1}, x_{i2}, \cdots, x_{im} \mid i=1, 2, \cdots, n\}$ 是每个输入的特征表示. 对于XGB模型的主要任务是构建 t 棵子树作为弱学习模型，其中预测值 $\hat{y}_i^t$ 与决策树之间的关系要满足以下公式：

$$\begin{aligned}
\hat{y}_i^0 &= 0\\
\hat{y}_i^1 &= f_1(x_i) = \hat{y}_i^0 + f_1(x_i)\\
\hat{y}_i^2 &= f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i)\\
&\cdots\\
\hat{y}_i^t &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i)
\end{aligned} \tag{1}$$

在梯度增强算法的每次迭代中，生成弱分类器 $f_t(x_i)$(比如决策树生成的结果)，并且这次迭代的预测值 $\hat{y}_i^t$ 是前一次迭代的预测值 $\hat{y}_i^{t-1}$ 和这一次迭代的决策树结果 $f_t(x_i)$ 的和.

因此，构建XGB模型需要解决3个关键问题.

1) 每轮迭代如何建立决策树，也就是叶节点如何进行拆分；

2) 如何确定每个DT上叶节点的预测值；

3) 每个DT与前一个DT有什么关系.

以上这3个问题由目标函数决定如式(2)，该算法的目标函数可以表示为

$$\min L^t(y, \hat{y}_i^t) = \min\Big(\sum_{i=1}^{n} l(y_i, \hat{y}_i^t)\Big) + \sum_{k=1}^{t} \Omega(f_k) \tag{2}$$

其中 $l(y_i, \hat{y}_i^t)$ 代表不同类型弱学习算法的损失函数，用于度量叶节点上真实值 y_i 和预测值 $\hat{y}_i^t$ 之间的差距. 而 $\sum_{k=1}^{t}\Omega(f_k)$ 是一个模型的正则项，也被称为惩罚项，用于度量整个模型的复杂程度，常用的正则项包括 $L1$ 和 $L2$ 正则，本文用到的正则项可以表示为

$$\Omega(f_k)=\gamma T_k+\frac{1}{2}\lambda\sum_{j=1}^{T_k}\omega_{kj}^2 \tag{3}$$

其中 T_k 是第 k 棵树的叶子节点数，γ 是叶结点 T 的收缩系数，ω_{kj} 是第 k 棵树在 j 个叶子节点上的分数，λ 是叶节点得分 ω 的惩罚系数，$\Omega(f_k)$ 的值需要通过交叉验证进行优化.

根据公式(1)，将 t 轮迭代中第 i 个样本的预测值 $\hat{y}_i^t$ 带入目标函数式(2)中，并通过式(3)求出的正则化项进行限制. 为了方便求解，将式(2)采用二阶泰勒展开，那么等式结果为

$$\min L^t=\min\left(\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)\right) \tag{4}$$

其中 g_i 和 h_i 分别是损失函数 $l(y_i, \hat{y}_i^t)$ 的一阶和二阶偏导数.

定义 $I_j=\{i\mid q(x_i)=j\}$ 作为 DT 中第一个叶节点 j 上的样本点集合，其中构造函数 $q(x)$ 将样本点 x 映射到叶节点 j 的位置，v 表示每个叶节点的分数，因此 DT 的结果可以用 $\omega_{q(x)}$ 表示. 考虑每个 DT 中 $f(x)$ 包含一个独立的树结构 $q(x)$ 和这棵树结果为 $\omega_{q(x)}$，因此每个 DT 可以表示为

$$f(x)=\omega_{q(x)},\ \omega\in R^T,\ q: R^d\to\{1, 2, \cdots, T\} \tag{5}$$

将公式(5)代入公式(4)，可推导出以下方程：

$$\min L^{(t)}=\min\left(\sum_{j=1}^{T_t}\left[G_j w_{t,j}+\frac{1}{2}(H_j+\lambda)\omega_{t,j}^2\right]+\gamma T_t\right) \tag{6}$$

其中 $G_j=\sum_{i\in I_j}g_i$，$H_j=\sum_{i\in I_j}h_i$.

在本文中将信用评价问题看为多分类问题，样本实际标签 y_i 的值是一个概率，因此本文选择对数损失函数作为损失函数：

$$l(y_i, \hat{y}_i^{(t)})=-(y_i\log(p_i)+(1-y_i)\log(1-p_i)) \tag{7}$$

其中 $p_i=\dfrac{1}{1+e^{-\hat{y}_i^t}}$.

对数损失函数的推导如下，在样本空间 $(x_1, y_1), \cdots, (x_n, y_n)$ 中，对于多分类问题，标号列 y_i 的值是 0 或 1. n 个样本的概率 $Y_i=y_i(i=1, 2, \cdots, n)$，可以根据多项式分布的概率公式得到：

$$p\{Y_1=y_1, Y_2=y_2, \cdots Y_n=y_n\}=\frac{N!}{y_1!\ y_2!\ \cdots y_n!}p_1^{y_1}p_2^{y_2}\cdots p_n^{y_n} \tag{8}$$

在这种情况下，需要满足最大似然函数 $\ln P$ 最大值. 如果将损失函数设置为 $-\ln P$，它可以等效于损失函数的最小值. 然后，损失函数可以表示为公式(7)，g_i 和 h_i 的值可以推导为

$$g_i=p_i-y_i \tag{9}$$

$$h_i=p_i(1-p_i) \tag{10}$$

同样，对于多分类问题，也可以选择 Softmax 函数作为损失函数：

$$l(y, p)=-\sum_{m=1}^{M}y_m\log p_m$$

$$p_m=\frac{e^{\hat{y}_i^t}}{\sum_{m=1}^{M}e^{\hat{y}_i^t}} \tag{11}$$

其中 M 代表标签的类别.

1.3 XGB模型需要解决的问题

叶节点预测值的确定是在XGB模型中需要解决的问题之一，假设树结构 $q(x)$ 是固定的并且保持不变. 根据公式(6)，由于 $G_j w_{t,j}+\frac{1}{2}(H_j+\lambda)\omega_{t,j}^2$ 是一个关于叶结点预测值 $w_{t,j}$ 的凸函数，通过对预测值取偏导数，可以找到最佳优化点 $w_{t,j}^*$ 使得目标函数 L^t 取得最小值.

$$w_{t,j}^*=\frac{G_j}{H_j+\lambda} \tag{12}$$

在这种情况下，目标函数的最小值为

$$\widetilde{L}^t=-\frac{1}{2}\sum_{j=1}^{T_t}\frac{G_j^2}{H_j+\lambda}+\gamma T_t \tag{13}$$

其中 $\widetilde{L}^t$ 可视为衡量树结构 $q(x)$ 质量的评分函数，因为 $\widetilde{L}^t$ 值越小，树的结构越好. 换句话说，只要可以确定树结构 $q(x)$，就可以获得最佳值.

XGB模型另一个需要确定的就是叶结点分裂的方式，对于第 t 棵节点的构建由这个树上的预测值 $\hat{y}_i^{t-1}$ 和真实值 y 共同决定. 对于叶结点而言，首先需要确定选择全部特征或者部分特征作为候选特征，然后使用贪心算法，尝试在每次迭代中加入一个分裂点到叶结点中，通过计算高斯分数可以找到最优的分裂点，该过程会一直进行迭代直到达到停止条件，决策树的整个构建也可以称为树结构的确定.

假设在叶节点分裂后，在左节点 I_L 和右节点 I_R 上重新发送样本集，$I=I_L\cup I_R$. 对于每个特征，在每次迭代中不分裂的叶节点其得分函数为 $\widetilde{L}_{\text{No-Split}}^t$，分裂后的得分函数值为 $\widetilde{L}_{\text{Split}}^t$，这两个得分函数的计算可以由以下公式得出：

$$\begin{aligned}\widetilde{L}_{\text{Split}}^t&=-\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}\right]+\gamma T_{\text{Split}}\\ \widetilde{L}_{\text{No-Split}}^t&=-\frac{1}{2}\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}+\gamma T_{\text{No-Split}}\end{aligned} \tag{14}$$

其中 G 代表分裂后一阶导数之和，H 代表分裂后二阶导数之和，L 和 R 分别代表左节点和右节点. 在公式(14)从 $\widetilde{L}_{\text{No-Split}}^t$ 减去 $\widetilde{L}_{\text{Split}}^t$，就能得到第 t 棵树叶节点的损失增益：

$$G_{\text{ain}}=-\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{15}$$

对每个叶节点选择增益值最大的特征节点作为分裂点，直到达到分裂停止条件才能完成第 t 棵树的生成. 在生成每棵树之后，将其添加到原始模型中，计算更新后的预测值 $\hat{y}_i^t$，然后根据上面的规则继续迭代，最终完成整个模型的构建.

XGB通常将公式(15)中的增益值作为评价特征重要性的指标，可以用来衡量某个特征对于预测结果的影响能力，因此可以根据增益值对特征进行过滤.

考虑到在一个叶节点的分裂过程中，总是选择 G_{ain} 值最大的特征进行分裂(这意味着这个特征具有最强的区分能力)，而XGB是一个通过不断累加分支结果进行预测的模型，因此必须同时考虑多棵树的影响. 因此，通过计算所有树上第一个特征的增益值 $\sum_{k=1}^{t}G_{\text{ain}\,r}^{\ k}$ 与所有树上的所有特征增益值之和 $\sum_{r=1}^{m}\left(\sum_{k=1}^{t}G_{\text{ain}\,r}^{\ k}\right)$ 的比值作为第 r 个特征变量的重要性指数：

$$I_{\text{mpr}}=\frac{\sum_{k=1}^{t}G_{\text{ain}\,r}^{\ k}}{\sum_{r=1}^{m}\left(\sum_{k=1}^{t}G_{\text{ain}\,r}^{\ k}\right)} \tag{16}$$

其中，$G_{\text{ain}\,r}^{\ k}$ 表示第 k 次迭代中第 r 个特征变量的增益值，t 是算法的迭代总数，m 是特征变量的总数. 重要性指数越高，区分样本的能力越强.

1.4　XGB 模型算法预测过程

1.4.1　*初始化*

首先根据公式(9)和公式(10)计算出损失函数的一阶导数 g_i 和二阶导数 h_i，然后计算每棵树上的 $\hat{y}_i^{t-1}$，根据第 $t-1$ 棵树上采用的预测算法以及样本值 x_i 获取预测值结果. 将样本值 x_i 对应的真实值 y_i 也作为 XGB 的输入. 对于 XGB 模型而言，根据公式(1)第 0 棵子树其预测值设为 0，即 $\hat{y}_i^0=0$.

1.4.2　*根节点设定*

为确定当前根节点，需要首先遍历并计算所有特征的增益值，以找到增益得分最大的特征节点作为当前根节点. 迭代过程的具体流程如图 1 所示.

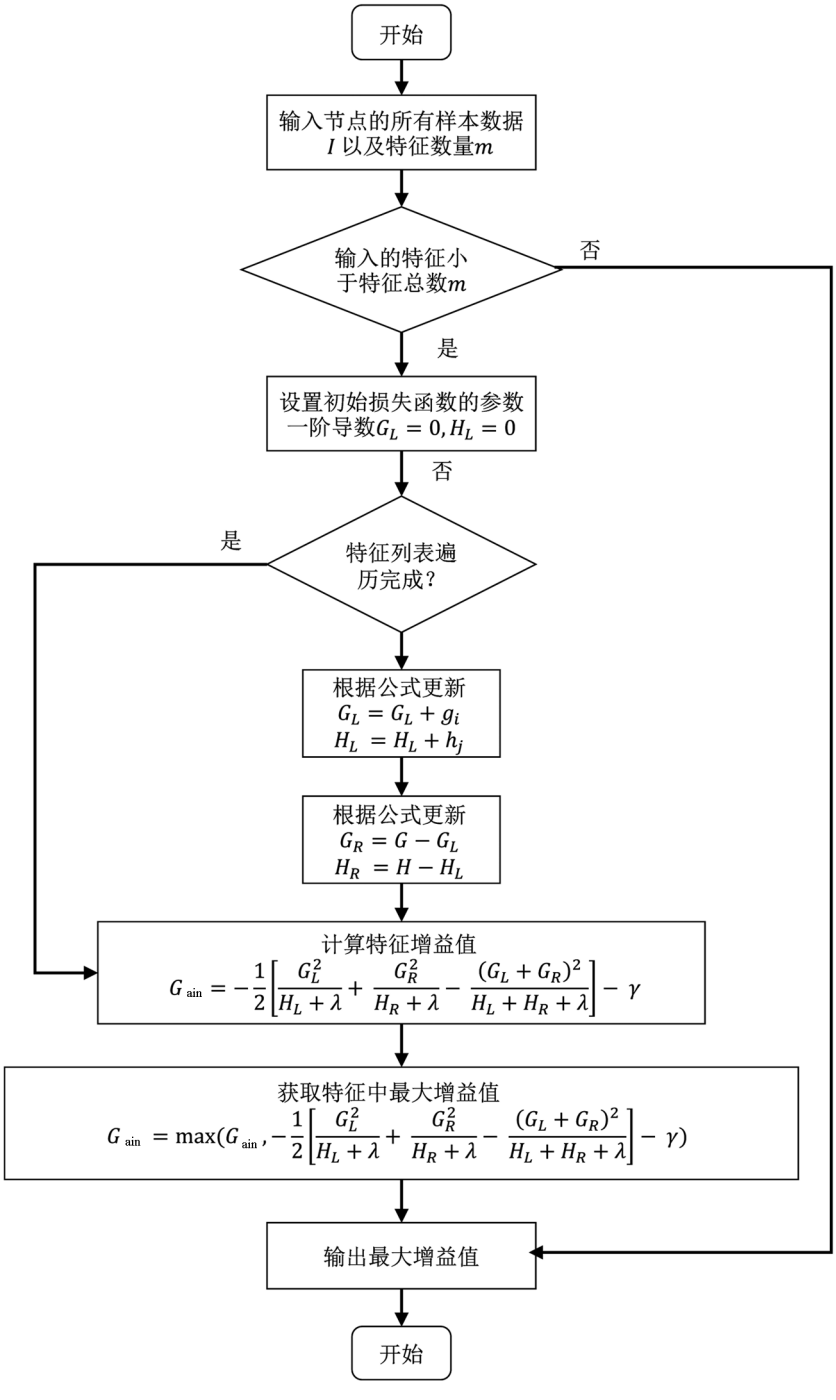


图 1　迭代寻找最大增益流程

1.4.3 建立叶节点集

根据第二步找到增益值最大的特征，将样本集分为两部分，得到两个叶节点样本集，对两个叶节点的集合分别重复上述第二步与样本集划分的过程，不断构建分支节点，直到增益分数为负或满足其停止条件，以此建立整棵树.

1.4.4 计算所有节点的预测值

根据公式(12)，叶节点的值 ω_j 可以计算出来，而第二棵树上的预测结果可以通过 $\hat{y}_i^2=\hat{y}_i^1+f_2(x_i)$ 根据公式(1)和公式(6)依次递推计算出整个树上的所有节点的预测值.

重复执行步骤1.4.1到1.4.4，直到建立足够数量的树来保证预测结果能够使评估函数获得最好的效果.

1.4.5 输出分类的结果

据公式(7)获取不同结果的概率，对于多分类问题采用一对多拆分的思想，对每个类别都将其他类别设置为反例，那么就有 N 个分类器. 每个分类器都能够识别一个固定类别，如果有一个分类器为正类，则就为该类别，若有多个分类器为正类，则选择置信度最高的分类器用于分类. 为使每轮预测值更接近真实值，每棵树都是基于前一棵树的预测结果构建的，从而提高模型的预测效果.

2 实验结果与分析

本节将XGB算法应用于一个营销客户数据集中，并与其他4种机器学习算法进行比较，验证了XGB算法在特征选择和分类方面的能力. 首先对数据集进行预处理，因为数据集中存在各种质量问题可能会干扰分类结果，例如数据不平衡、空值、数据异构等问题；然后，将根据特征选择算法提取数据中特征；最后对模型进行训练、测试和评估.

2.1 数据描述

本文采用的数据是基于电力营销部门提供的数据集，每季度更新一次. 原始数据集包括2018年1月1日至2018年12月31日期间的样本1 272个，于2019年8月5日下载，包含143个特征变量和1个标签列，标签是根据历史行为信息人工标注的信用等级评价. 考虑到标签栏中包含的7种状态以及电力营销部门的实际状况，将“全额支付”“支付中”“宽限期”“违约”“违约(16～30 d)”“违约(31～120 d)”“指控中”通过编码的形式进行转换. 数据集中不同标签的分布如表2所示.

表2 数据集标签分布

分类	样本数	占比/%	分类	样本数	占比/%
全额支付	932	65.41	违约(16～30 d)	15	2.75
支付中	123	21.15	违约(31～120 d)	12	0.94
宽限期	32	2.52	追究责任	7	0.55
普通违约	55	6.68	总数	1 272	100

2.2 数据预处理

由于样本之间的比例不均衡，出现了长尾现象，本文采用了分层随机采样的方式，将大样本量与小样本量的分布控制在4∶1. 根据实际经验，手动删除了29个不符合进入模型标准的特征，选择了114个特征，考虑到某个特征在有大量空值时不再具有代表性，因此删除了缺失比例超过50%的35个特征，还剩余79个特征. 对于字符数据通过独热编码的方式进行转换，即将字符类别变量转化为易于模型学习的数字形式，并采用所有非空值的平均值填充数据列中空值，针对离散数据列中的空值采用0进行填充.

2.3　特征选择

为了验证 XGB 特征选择的有效性，首先使用传统基于线性的皮尔逊相关图选择特征方法进行特征提取，传统线性方式的流程如下：首先输入数据预处理后的 79 个特征，对于每个自变量计算其预测能力，用皮尔逊相关图找出任意两个特征之间的相关系数. 为了提高模型效率和减少数据冗余，对于两个相关系数大于 0.6 的特征，只能保留预测能力较高的特征. 经过基于线性的皮尔逊相关图选择特征方法，一共选择了 17 个特征，这意味着处理后的最终数据集将只包含这 17 个特征.

基于 XGB 的特征选择模型采用经过数据预处理后的 79 个特征作为特征提取模型的数据集，将所有的特征放入 XGB 特征提取模型中，计算每个特征的增益值与重要性指数，获取不同特征对于预测值的重要度的排名，可以采用 XGB 库中提供的 Get_Fscore 函数返回由高到低排序的特征重要性索引表，然后选取重要度指标最高的 17 个特征，与之前方法选择的特征相比，有 9 个不同的特征.

2.4　结果分析

在本节中，将 XGB 的预测结果与 LR,DT,RF 和 GBDT 进行比较，从而评估模型的性能.

原始数据集在每个部分的样本比例保持不变的前提下，随机分为 5 个部分，模型的预测效果将通过五重交叉验证来验证，这意味着将随机选择 4 个部分作为训练集，剩余一个部分作为测试集. 对于原始数据将被随机分成 5 个集合 s_1，s_2，…，s_5，使 5 个集合的大小和分布相等. 然后将 5 个集合中的一个用作测试集合，其余 4 个集合被用作训练集合. 最后，将这 5 个测试集的平均值作为模型的最终预测结果.

在模型训练过程中，采用 Python 提供的相关函数进行处理，通过 GridSearch(网格搜索)函数返回每个模型的最优参数. 为便于处理，对没有违约的客户作为正例，而违约的客户作为反例. 根据抽样后分组的数据即每组 254 个样本数据，在对比两种特征选择方法后构建混淆矩阵如表 3、表 4，可以看出采用 XGB 作为特征选择的方式有一定优势，对于区分客户的信用能力更强.

表 3　传统特征选择方法的混淆矩阵

	传统特征选择方法的混淆矩阵结果	
	预测：好	预测：坏
真实：好	201	4
真实：坏	15	34

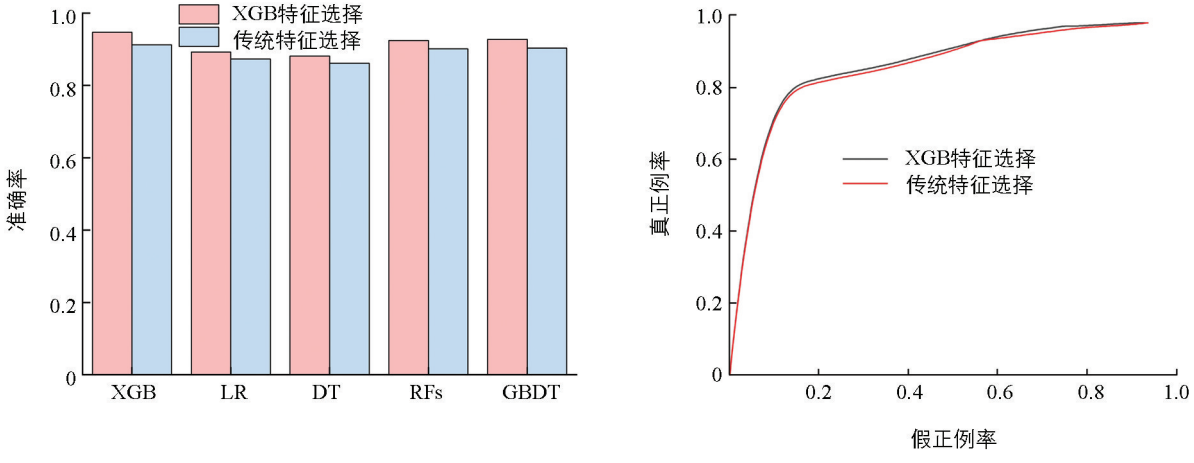
表 4　XGB 特征选择方法的混淆矩阵

	XGB 特征选择方法的混淆矩阵结果	
	预测：好	预测：坏
真实：好	204	2
真实：坏	13	35

分类问题中最常见的评价指标是准确率，如 K_{appa}，A_{UC} 等，它们通常是一起考虑的. 其中 K_{appa} 通常用于衡量不平衡数据集中的分类准确率，K_{appa} 越高，分类准确率越高. 通过计算分类问题标准评价指标 R_{OC} 和 A_{UC}，绘制 R_{OC} 曲线. R_{OC} 曲线通常用于衡量模型的预测性能，A_{UC}(R_{OC} 曲线下的面积)用于评估分类系统的性能. A_{UC} 的取值范围为[0，1]，A_{UC} 越大，模型的分类效果越好.

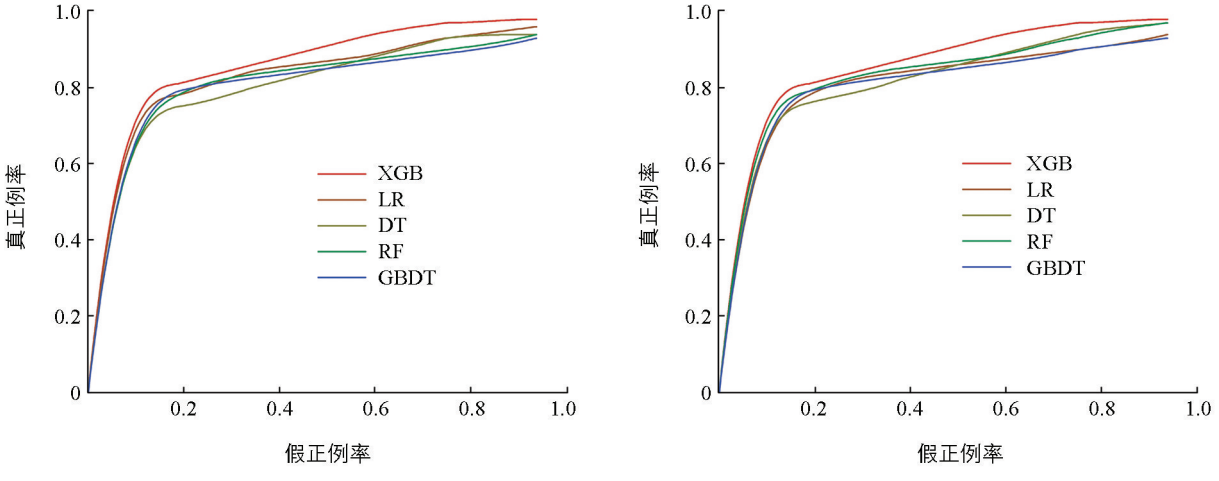
基于传统特征选择和 XGB 特征选择方法的 5 种模型的性能分别如表 5 和表 6 所示. 可以看出本文采用的基于 XGB 特征选择方法比传统特征选择方法在各个模型上都能取得明显的优势，同时 XGB 在分类效果上能取得最好的效果. 两种特征选择方法的 R_{OC} 曲线与准确率如图 2 所示，也能明显看出 XGB 特征选择方法由于采用了特征重要度指标与增益值两个指标，对于特征选择的效果优于传统基于皮尔逊系数的方

法. 两种特征选择方法的5个对比模型的 R_{OC} 曲线如图3所示，可以看出基于树的模型如XGB，RF，GBDT等采用集成学习的思想能够对客户信用做出一个较为准确的预测，而传统的方法只是采用单一模型去拟合电力客户数据，可能会欠缺某些关键特征的获取. 而XGB模型由于能够计算不同叶节点上的增益值，汇聚不同节点的预测结果从而达到了最好的预测效果.



(a) 两种特征选择算法的准确率对比　　(b) 两种特征选择方法的R_{OC}对比

图2　不同特征选择方法对比



(a) 基于XGB特征选择的R_{OC}曲线　　(b) 基于传统特征选择的R_{OC}曲线

图3　两种特征选择方法的5个对比模型的 R_{OC} 曲线

表5　基于传统特征选择的5种模型表现

	准确率	K_{appa} 系数	A_{UC}	K_S
XGB	0.926 1	0.738 2	0.938 3	0.733 2
LR	0.864 4	0.469 0	0.849 8	0.539 0
DT	0.866 5	0.584 9	0.804 2	0.608 3
RF	0.909 5	0.664 4	0.882 3	0.637 0
GBDT	0.920 4	0.707 4	0.922 6	0.697 0

表 6　基于 XGB 特征选择的 5 种模型表现

	准确率	K_{appa} 系数	A_{UC}	K_S
XGB	0.937 0	0.776 3	0.948 1	0.770 0
LR	0.892 1	0.595 6	0.873 0	0.587 8
DT	0.880 1	0.622 2	0.819 4	0.638 7
RF	0.924 9	0.726 4	0.907 7	0.699 5
GBDT	0.927 9	0.737 0	0.934 3	0.720 3

从图表分析看，XGB 特征选择方法可以有效提取电力营销大数据中的关键特征，这种特征选择方法的优势在不同模型中都可以体现出来. 此外，在表 5 和表 6 中，XGB 的准确率、K_{appa} 值、A_{UC} 值和 K_S 值最高，其中准确率说明了 XGB 模型在预测效果上远超其他模型，这可能是由于多维客户数据基于集成思想的 XGB 模型能够更好地处理样本数据之间的关系. 而对于 K_{appa} 值，它代表处理数据不平衡的能力，而本文采用的数据是呈现明显的长尾分布，而 XGB 模型通过修改部分参数能够处理样本不平衡的情况. K_S 值能够反映处理好坏样本的能力，而 K_S 值通常与 A_{UC} 一起来判断模型预测的效果，K_S 值越大，体现模型对好坏样本的区分能力越强. 而 XGB 模型通过集成学习的思想对数据中存在的差异性能够更好地捕捉，实验结果也表明了 XGB 模型的优越性，效果明显优于其他 4 个模型，在客户信用评价中能起到更好的表现.

通过本文提出的 XGB 模型，分析历史的客户信息以及交易信息，从而建立一套可以用于区分客户信用级别的多维度信用评价模型. 该模型采用集成学习的思想，通过树模型的结构对特征计算信息增益，同时采用多个弱学习模型的结果组成一套强学习模型，建立了客户画像，能够有效地对新客户根据其信息进行信用等级预测，能够帮助电力营销部门进行精准营销、规避风险.

3　结论

我国作为一个用电大国，对于电力客户信用的研究相对起步较晚，而客户信用问题一直都是困扰电网部门实现精准营销的关键问题. 本文首先研究了近年来客户信用评价模型的研究现状，对电力营销客户信用数据进行了分析，提出了基于 XGB 算法的信用等级分类问题的理论建模，然后基于电力营销数据集与客户信息数据，将 XGB 应用于电力营销系统中的信用等级预测，取得了较好的成果，结果表明：

1）采用 XGB 模型进行客户信用等级预测，改变了之前采用层次分析法与熵值法无法处理大量数据的情况.

2）通过采用增益值计算方法与特征重要度指标作为特征选择方法，比传统的特征选择方法其实验表现效果更好.

3）通过与 LR，DT，RF 和 GBDT 的性能对比，验证了 XGB 在特征选择和分类性能上具有明显优势.

参考文献：

[1] 余培，刘其辉，石城. 新电改背景下电力用户信用风险预警模型与评价方法 [J]. 电力需求侧管理，2020，22(2)：34-38.

[2] 夏兴平. 基于模糊层次分析法的电信业务网项目后评价研究 [J]. 重庆邮电大学学报(自然科学版)，2012，24(6)：792-797.

[3] 王恺，卢帮明，于新钰，等. 基于熵值法和加权聚类的电力零售市场风险评估 [J]. 科技创新与应用，2021(2)：83-85，88.

[4] 柯行思，吴梦昭，李博，等. 基于改进熵权法的电力信用数据敏感度监控算法 [J]. 电子设计工程，2020，28(24)：66-69，75.

[5] LI F L，XU E F，SUN K，et al. Credit Risk Evaluation of Large Power Consumers Considering Power Market Transaction [J]. Iop Conference Series：Earth and Environmental Science，2018，127：012017.

[6] 胡亚红，方豪达，江大川，等. 基于 AHP 和 k-means 算法的电力用户信用度评价 [J]. 浙江工业大学学报，2018，46(5)：515-521.

[7] 邓君，王菲，张长志，等. 基于模糊层次分析法的荆江航道整治水下抛石施工风险评估 [J]. 西南大学学报(自然科学版)，2018，40(6)：183-188.

[8] XU G X，LIU Y，KHAN P W. Improvement of the DPoS Consensus Mechanism in Blockchain Based on Vague Sets [J]. IEEE Transactions on Industrial Informatics，2020，16(6)：4252-4259.

[9] 余培. 基于 XGBoost 算法的电力用户信用风险预警及分级服务策略 [D]. 北京：华北电力大学，2020.

[10] 王明胜，石会燕，马万里，等. 基于电力大数据的企业信用评价方法研究 [J]. 电子技术与软件工程，2021(11)：163-164.

[11] KEARNS M J，VALIANT L G. Cryptographic Limitations on Learning Boolean Formulae and Finite Automata [J]. Journal of the Association for Computing Machinery，1993：433--444.

[12] BREIMAN. Random Forests [J]. Mach Learn，2001，45(1)：5-32.

[13] FRIEDMAN J H. Greedy Function Approximation：A Gradient Boosting Machine [J]. The Annals of Statistics，2001，29(5)：1189-1232.

[14] 姜玉苹，余诚，林燕榕，等. 基于集成学习算法的慢性肾病早期筛查方法 [J]. 西南大学学报(自然科学版)，2020，42(10)：17-24.

[15] MANDALA I G N N，NAWANGPALUPI C B，PRAKTIKTO F R. Assessing Credit Risk：An Application of Data Mining in a Rural Bank [J]. Procedia Economics and Finance，2012，4：406-412.

[16] LI G I，SHI Y L，ZHANG Z H. P2P Default Risk Prediction Based on XGBoost，SVM and RF Fusion Model [C] // Proceedings of the 1st International Conference on Business，Economics，Management Science (BEMS 2019). April 20-21，2019. Hangzhou City，China. Paris，France：Atlantis Press，2019.

[17] 程超鹏，彭显刚，曾勇斌，等. 相异模型下 Stacking 集成结构的异常用电用户识别方法 [J]. 电网技术，2021，45(12)：4828-4836.

[18] 牛小梅，张银玲. 层次分析法在电力客户信用风险中的评价 [J]. 计算机仿真，2011，28(5)：333-336.

责任编辑 汤振金