文章编号：2096-1472(2023)-03-25-05
DOI:10.19644/j.cnki.issn2096-1472.2023.003.006

基于XGBoost的不平衡员工晋升预测

黄 静[1]，郑慧慧[2]

(1.浙江理工大学信息科学与工程学院，浙江 杭州 310018；
2.浙江理工大学计算机科学与技术学院，浙江 杭州 310018)
✉syhj_sy@163.com; 472596438@qq.com

摘 要：人力资源管理在决策方式上逐渐智能化。为了辅助人力资源进行员工晋升决策，提高管理者晋升决策的公平性和有效性，提出基于集成学习的极端梯度提升(XGBoost)模型对数据分布不平衡的员工晋升数据进行预测分析。以Kaggle(数据科学竞赛平台)的员工晋升数据集为对象进行预处理，建立基于XGBoost算法的员工晋升预测模型，结合准确率、*F*1值和*AUC*值这几个评价指标，与其他算法模型进行比较分析。实验结果表明，相比逻辑回归(LR)、支持向量机(SVM)、人工神经网络(ANN)、多层感知机(MLP)模型，XGBoost模型的三项评价指标更具优势，应用于员工晋升预测，效果更好。

关键词：人力资源；员工晋升预测；机器学习；XGBoost
中图分类号：TP319 **文献标识码**：A

Unbalanced Employee Promotion Prediction based on XGBoost

HUANG Jing[1], ZHENG Huihui[2]

(1.*School of Information Science and Engineering*, *Zhejiang Sci-Tech University*, *Hangzhou* 310018, *China*;
2.*School of Computer Science and Technology*, *Zhejiang Sci-Tech University*, *Hangzhou* 310018, *China*)
✉syhj_sy@163.com; 472596438@qq.com

Abstract: Human resource management is becoming more intelligent in decision-making. In order to assist human resources in making employee promotion decisions and improve the fairness and the effectiveness of managers' promotion decisions, this paper proposes an XGBoost model based on integrated learning to predict and analyze employee promotion data with unbalanced data distribution. Taking the employee promotion dataset of Kaggle data science competition platform as the object for preprocessing, employee promotion prediction model is established based on XGBoost algorithm. This model is then compared with other algorithm models on the evaluation indicators of accuracy, *F*1 value and *AUC* value. The experimental results show that XGBoost model has more advantages in three evaluation indicators than LR (Logistic Regression), SVM (Support Vector Machine), ANN (Artificial Neural Network) and MLP (Multilayer Perceptron) models, and it has better effect when applied to employee promotion prediction.

Keywords: human resources; employee promotion prediction; machine learning; XGBoost

1 引言(Introduction)

随着市场竞争越来越激烈，人才已经成为非常重要的竞争资源，也是企业发展的核心要素。晋升能够对员工进行有效的激励，促使员工发挥更大的潜力和价值，也能为企业留住更多有才华的员工，为其创造更多的收益[1]。互联网信息时代，人力资源数据类型和数量逐渐增多和增大，其数据化价值持续放大。员工信息表现出越来越多样和繁杂的特征，人力资源部门需要采用信息化、数据化的方式提升对员工晋升的分析、决策效率，以期改善传统人力资源管理的信息更新缓慢的缺陷和决策的单调性，促使人才晋升透明化，以此有效激励员工积极工作[2]。

目前，机器学习在人力资源管理领域的应用和研究有很

基金项目：浙江省重点研发计划项目(2022C01207)。

多[3]，研究的内容大多涉及人才招聘、人才离职流失、预防人才流失等方面。高超[4]分析了数据挖掘在人才招聘、人才管理和离职流失分析等人力资源管理中的具体应用。赖华强等[5]和张金艳[6]对数据挖掘在离职管理方面的应用进行了分析和实现。PUNNOOSE等[7]为了解决人员流失的问题，应用了极限梯度增强技术预测员工流动率。KUMAR等[8]实现了一个人力资源排名模型，可用于预测简历的排名和分类，有效地简化了人力资源招聘工作。KHERA等[9]建立了一个基于支持向量机的员工离职模型，主要用来预测企业的员工流失率。随着机器学习在人力资源领域的影响不断扩大，张敏等[10]对机器学习正在重塑人力资源管理者的管理理念和方式的探讨，为本文将XGBoost预测模型应用于人力资源的晋升场景带来了更深入的思考。

为帮助企业决策者调整人才晋升管理策略、提升员工晋升公正性，本文通过分析预处理Kaggle提供的员工分析数据集，并利用XGBoost算法构建员工晋升预测模型，与其他机器学习模型进行相应模型评价指标比较，验证XGBoost模型的效果和有效性，从而进一步分析影响员工晋升的因素。

2 XGBoost模型介绍(Introduction to XGBoost model)

在门店销售、客户行为、广告点击率等营销方面和灾害风险等方面，可利用XGBoost[11]进行相关预测；在高能物理事件、Web文本、恶意软件、产品等领域，可利用XGBoost进行相应的分类判断。在各领域的广泛问题上，XGBoost都给出了相对较好的效果。

XGBoost[12]是一种基于boosting思想的并行回归树模型，其中boosting思想是指在已有的若干弱分类器进行加权求和得到最终的分类器。XGBoost模型是由CHEN等[11]在梯度下降决策树(Gradient Boosted Decisiom Tree，GBDT)的基础上改进而来。与GBDT[13]模型比较，XGBoost极大地提升了模型训练计算的速度和预测及分类的精度，是GBDT算法的升级版。XGBoost[14]是由多棵决策树(即CART回归树)[15]组建构成的，每一棵决策树学习的是目标值与预测值的残差，其中预测值是之前所有决策树的预测值之和。所有决策树训练完成后进行共同决策，样本在每一棵树上得到相应的预测值之后进行累加作为其最终预测结果，在训练阶段，每一棵新的树都是在已训练完成建成的树的基础上进行训练的。其中，每一棵决策树都是弱学习器。通过boosting技术将所有弱学习器提升成为一个强学习器。为了避免模型过拟合，同时增强泛化能力，XGBoost在GBDT模型的损失函数上增加正则化项。传统GBDT计算损失函数采用一阶泰勒展开，利用负梯度值代替残差进行拟合，XGBoost则对损失函数增加二阶泰勒展开，使用二阶导数收集梯度方向信息，以此提高模型的精确性。此外，XGBoost对每一个特征实行分块并排序，因此在寻找最佳分裂点时可以实现并行化计算，从而提高了计算速度。

对于给定包含n 个样本和m 个特征的数据集，该数据集表示为$D=\{(x_i,y_i)\}(|D|=n,x_i\in\boldsymbol{R}^m,y_i\in\boldsymbol{R})$，树集成模型使用$K$ 个可加函数预测输出。

$$\hat{y}_i=\phi(x_i)=\sum_{k=1}^{K}f_k(x_i),\quad f_k\in F \tag{1}$$

式(1)中，$\hat{y}_i$表示n 棵决策树$f_k(x_i)$串联叠加的XGBoost模型。$F=\left\{f(\boldsymbol{x})=\omega_{q(x)}\right\}\left(q:\boldsymbol{R}^m\to T,\omega\in\boldsymbol{R}^T\right)$，表示一个表征决策树函数空间，其中$q$表示将示例映射到对应叶索引的每棵树的结构，$T$ 是树中的叶节点数，每一个f_k 对应一个独立的树结构q和叶节点权重ω。不同于单棵决策树，每棵决策树在每个叶子节点上都带有权值，用ω_i表示第i 个叶子上的权值。针对某个给定的数据样本，XGBoost将通过树中的决策机制将其分类到叶子中，并利用相应的叶子上的权值进行求和计算得到最终预测，为了学习模型中的函数集合，最小化目标函数如下：

$$L(\phi)=\sum_i l(\hat{y}_i,y_i)+\sum_k\Omega(f_k) \tag{2}$$

式(2)中，$l(\hat{y}_i,y_i)$是一个可微凸损失函数，表示预测值$\hat{y}_i$和目标值y_i之间的误差，迭代过程中将不断拟合减少误差。正则项$\Omega(f_k)$用来约束决策树的叶子节点数T 和叶子权值ω。

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2 \tag{3}$$

式(3)中，γ和λ分别表示为叶子节点数T 和叶子权值ω的L_2平方模系数，正则化项有助于平滑最终权重，避免过拟合。将式(2)进行二阶泰勒展开，将二阶形式作为近似目标函数。

$$L^{(t)}\approx\sum_{i=1}^{n}\left[l\left(y_i,\hat{y}_i^{(t-1)}\right)+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t)+C \tag{4}$$

式(4)中，$g_i=\partial\hat{y}_i^{(t-1)}l(y_i,y_i^{(t-1)})$和$h_i=\partial^2\hat{y}_i^{(t-1)}l(y_i,y_i^{(t-1)})$为目标函数的一阶导数和二阶导数。将式(3)代入式(4)，并约去常数项可化简为式(5)：

$$L^{(t)}\approx\sum_{i=1}^{n}\left[g_i\omega_{q(x_i)}+\frac{1}{2}h_i\omega_{q(x_i)}^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{5}$$

式(5)中，q 为输入映射至叶子的索引，即$q:\boldsymbol{R}^m\to T$，定义每个叶子的样本集合为$I_j=\{i\,|\,q(x_i)=j\}$，将式(5)进行如下改写：

$$\begin{aligned}L^{(t)}&\approx\sum_{i=1}^{n}\left[g_i\omega_{q(x_i)}+\frac{1}{2}h_i\omega_{q(x_i)}^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2=\\&\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)\omega_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)\omega_j^2\right]+\gamma T=\\&\sum_{j=1}^{T}\left[G_j\omega_j+\frac{1}{2}\left(H_j+\lambda\right)\omega_j^2\right]+\gamma T\end{aligned} \tag{6}$$

式(6)中，$G_j=\sum_{i\in I_j}g_i$，$H_j=\sum_{i\in I_j}h_i$，问题即可转化为求解一元二次方程最优解问题，解得最优解ω_j^*和目标函数最优解L^*。

$$\omega_j^* = -\frac{G_j}{H_j+\lambda} \tag{7}$$

$$L^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\lambda T \tag{8}$$

式(7)和式(8)中，构成目标函数的G_j和H_j在取值上是由第j个树叶上数据样本所决定的，而第j个树叶具有的数据样本是由树结构函数q决定的，则推导可知决策树结构q，易求得目标函数值，L^*代表当指定一个树的结构时，目标函数上最多减少多少，故把L^*作为评价一棵树模型的评分函数，评分越小，表明该树的结构模型越优。训练的目的在于寻求最佳决策树结构q^*，使得目标函数取得最优解。

3　基于XGBoost 的预测方法(Prediction method based on XGBoost)

3.1　数据集描述

本文采用Kaggle平台HR Analytics: Employee Promotion Data(人力资源分析：员工晋升数据)提供的公开员工数据集作为数据源。Kaggle作为目前最大的机器学习数据及数据分析竞赛平台，能确保其数据的真实性和适用性。根据企业的实际情况，只有少数员工能获得晋升机会，该数据集存在不平衡问题，数据集中的训练集共有54,808 个样本，测试集有23,490 个样本。训练集样本中有未晋升员工50,140 个，晋升员工4,668 个。数据集包括12 个特征变量列，1 个标签列。其中，标签列“晋升状况”，0=未晋升，1=已晋升。特征变量列有5 个数值型变量和7 个类别型变量。数值型变量包括“上一年完成其他软技能、技术技能等培训次数”“年龄”“上一年员工的评级”“工龄”“当前培训评估的平均分”，类别型属性变量如表1所示。

表1 类别型特征变量含义描述

Tab.1 Meaning description of category characteristic variables

特征变量名称	特征变量描述	值说明(举例)
department	员工所在部门	Analytics，Finance，HR，Legal，Operations…
region	就业地区	region 1，region 2，region 3，region 4，region 5…
education	教育程度	Master’s & above，Bachelor’s，Below Secondary
gender	员工性别	F，M
recruitment_channel	招聘渠道	Sourcing，Other，Referred
KPIs_met ＞ 80%?	KPI是否达到80%	Yes，No
Awards_won?	上一年中是否获奖	Yes，No

3.2　XGBoost模型预测流程

基于XGBoost的员工晋升预测流程如图1所示，主要包括以下步骤：针对员工数据集进行预处理；采用训练集构建XGBoost模型并确认最终模型参数；预测测试集的员工晋升结果，查看模型的预测效果。

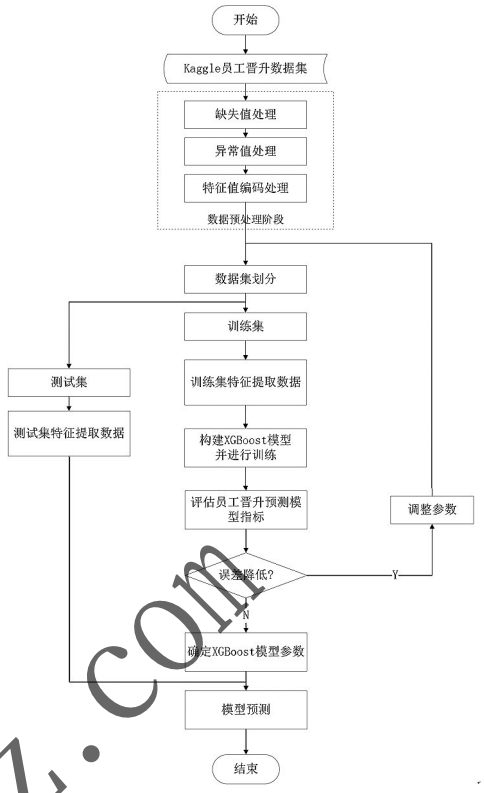

图1 基于XGBoost的员工晋升预测基本流程

Fig.1 Basic process of employee promotion prediction based on XGBoost

(1)数据预处理。员工数据中部分特征存在缺失值，重要特征值的缺失将会影响模型训练效果。本文将对缺失特征值的样本进行适当剔除或填充处理[16]。特征分为类别型特征和数值型特征，需要对类别型特征进行编码处理。在类别型特征中，对性别、教育程度等特征进行序号编码(OrdinalEncoder)[17]，对员工所在部门、就业地区、招聘渠道等特征进行独热编码(One-HotEncoder)[18]。因为实际情况是只有少数人员才能获得晋升机会，所以在数据分布上会存在数据不平衡问题[19]。本文采用SMOTE方法对数据集进行重采样，处理数据集不平衡问题。

(2)学习和确定模型。采用交叉验证的思想，将数据预处理之后得到的数据集以7:3的比例随机分为训练数据集和测试数据集。训练数据集将输入XGBoost模型进行学习训练，不断调整模型参数提升预测精度，最终确定模型参数。

(3)预测晋升结果。预测测试数据集的员工晋升结果，计算预测评估指标，分析XGBoost模型的准确性，并与其他预测模型相比较，查看模型的预测效果。

4　实验分析(Experimental analysis)

4.1　数据预处理

本文主要针对Kaggle平台发布的源数据集中的部分重要特征存在的缺失值问题、类别型特征编码问题及数据不平衡问题进行数据预处理，防止影响模型训练结果。首先针对重要特征存在的缺失值问题，采用过滤删除样本或填充特征值

方法处理数据；其次采用序号编码、独热编码和二进制编码对类别型特征进行编码处理，使其数值化；最后采用SMOTE过采样技术解决数据不平衡问题。

检查数据是否存在缺失值、重复值和无关变量，发现在教育程度(education)和上一年员工评级(previous_year_rating)存在缺失值，如图2所示。

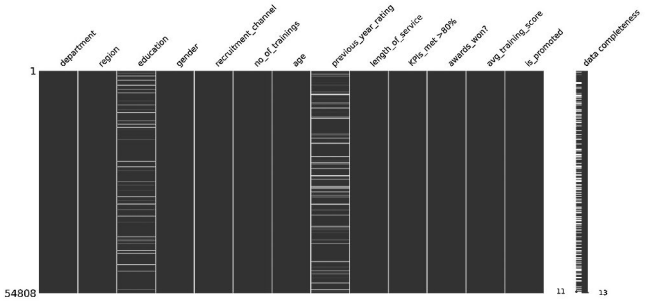

图2 数据集缺失值分布情况图

Fig.2 Distribution of missing values in dataset

由于“教育程度”是一个类别型特征，表示一个人是否达到了特定教育水平，它是一个较为重要的特征，不可随意指定，这是因为该员工可能还未达到指定水平，这将导致分析不准确，训练效果不好。在数据集的54,808 个样本中，存在2,409 个样本的“教育程度”为空值，占全部样本的4.39%，由于占比较小，因此过滤剔除这部分样本不会给模型训练带来重大变化。“上一年的员工评级”是一个数字型特征，表示员工在上一年的评级，该特征值为空，表示该员工加入公司的时间少于1 年，尚未存在上一年的评级记录，因此可用“0”填充该特征值。

针对数据集中的类别型特征，本文将通过序号编码和独热编码对这些类别型特征进行编码处理。序号编码一般用来处理类别值间具有大小、上下关系的数据。其中，“教育程度”的类别值Master’s & above，Bachelor’s，Below Secondary之间具有大小关系，故采用序号编码映射编码成[0,2]的整数。“所在部门”“就业地区”“招聘渠道”这几个特征的类别值之间不具有大小关联，因此使用独热编码进行编码处理。剩余类别型特征的类别值仅有两种，因此使用二进制编码方式用0和1进行编码。

按照实际晋升情况，晋升员工样本在全部样本中占比很小，不利于模型训练学习，模型会倾向于学习比例较高的数据特征，对于比例低的数据只学习很少的特征。为克服在现实情况下因为数据不平衡问题导致训练效果不佳的问题，本文将采用SMOTE-Synthetic Minority Oversampling Technique(合成少数过采样技术)[20]通过复制少数实例随机增加少数类实例平衡类分布，解决数据不平衡的问题，提高模型的训练效果。利用SMOTE重采样之后，数据样本数量达到95,704 个，其中正负样本各47,852 个。

4.2 模型验证与评估

本文选用准确率(*Accuracy*)、*F*1 值和*AUC*值这三项分类算法评价指标衡量判断模型的效果。计算*AUC*值需求得描述分类器的混淆矩阵。把是否晋升的分类观测值放入矩阵中，得到混淆矩阵如表2所示。

表2 晋升分类结果混淆矩阵

Tab.2 Confusion matrix of promotion classification result

晋升状态	预测晋升	预测未晋升
实际晋升	*TP*	*FN*
实际未晋升	*FP*	*TN*

准确率是指对于给定的测试数据集，分类器进行正确分类的样本数与总样本数之比；*F*1 值是精确率和召回率的综合衡量指标，*F*1 值越接近1，则说明模型预测更准确。准确率和*F*1 值是由混淆矩阵计算得到。可利用混淆矩阵绘制出受试者工作特征(ROC)曲线，*AUC*值是由该曲线求得。*AUC*值越大，模型精度越高。准确率和*F*1 值的计算公式如式(9)和式(10)所示：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{9}$$

$$F1=\frac{2\times Precision\times Recall}{Precision\times Recall}\times 100\% \tag{10}$$

本文数据集经过预处理之后，样本总量达95,704 个，编码后特征列为55 列，是否晋升作为预测的结果标签。将特征变量与目标变量输入XGBoost模型，按照7:3的比例划分训练集数据与测试集数据，构建模型进行训练预测。

通过不断调整参数，得到的XGBoost模型最优超参数组合为n_estimators=100、learning_rate=0.3、max_depth=6、colsample_bynode=0.7、colsample_bytree=0.7、min_child_weight=2、subsample=0.8，其余参数则设为默认值。将建立之后不断调优得到的XGBoost模型与LR、SVM、ANN、MLP模型进行相应评价指标的交叉验证实验对比，对比结果如表3所示。

表3 模型对比结果

Tab.3 Comparison results of models

算法模型	准确率/%	*F*1值/%	*AUC*值/%
LR	80.61	81.19	80.63
SVM	87.36	88.76	87.36
ANN	89.98	90.28	88.49
MLP	89.93	90.28	89.96
XGBoost	96.71	96.61	96.56

ROC曲线下的面积称为*AUC*值。ROC曲线采用真阳性率(True Positive Rate，TPR)为纵轴，假阳性率(False Positive Rate，FPR)为横轴，其中真阳性率是指预测结果为晋升且实际结果也为晋升的实例，是混淆矩阵中的*TP*，又称灵敏度；假阳性率是指预测结果为晋升但是实际结果为未晋升的实例，是混淆矩阵中的*FP*。ROC曲线能直观地反映模型的性能。上述模型算法的ROC曲线如图3所示。

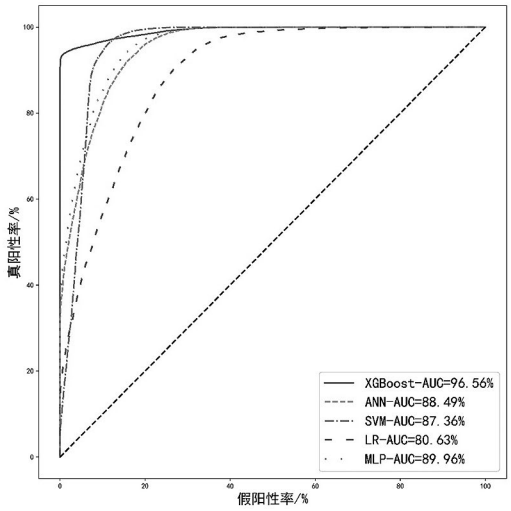

图3 模型ROC曲线对比图

Fig.3 Model ROC curve comparison diagram

分析模型对比的实验结果发现，本文建立的XGBoost模型在预测员工晋升时的准确率达到96.71%，*F*1值为96.61%，*AUC*值为96.56%，相较于LR、SVM、ANN、MLP四种模型，其三项指标都具有最佳表现，其中*AUC*值通过ROC曲线直观地表明XGBoost算法模型的预测效果最好。员工是否晋升与其相对的教育程度、工龄、年龄、上一年评级等特征之间存在较为复杂的影响关系。XGBoost模型基于集成方法，在模型的复杂度和精确性之间得到一个较好的平衡效果，并基于贪心算法思想，在建立决策树的过程中寻找最佳分裂点，较之上述其他算法具有一定的优越性。

5 结论(Conclusion)

当下环境，人力资源在决策策略方法、管理手段上数据化程度不断深化，基于大量数据和算法的员工晋升预测为企业的人才选拔和储备发展提供了新的视角和信息。本文对Kaggle平台提供的员工数据集采用XGBoost模型建立晋升预测模型，与LR、SVM、ANN、MLP模型进行相应的评价指标的实验对比，分析影响员工晋升的影响因素，XGBoost模型在晋升预测上优于其他模型，其*AUC*值达96.56%。下一步将考虑企业员工实际情况，增加新特征，进一步提高预测模型对于员工晋升问题的应用意义。

参考文献(References)

[1] 张昶.多因素作用下中国国有企业晋升机制研究[D].北京:北京邮电大学,2021.

[2] 郭宝.信息技术对人力资源管理模式的影响与作用研究[J].国际公关,2019(11):202.

[3] 韦家铎.机器学习在人力资源管理系统中的应用[D].武汉:华中科技大学,2019.

[4] 高超.数据挖掘技术对企业人力资源管理的影响分析[J].电子制作,2013(17):288.

[5] 赖华强,王三银,仲崇高.人力资源管理领域的数据挖掘应用展望——以基于灰色关联模型的离职管理实证分析为例[J].江苏商论,2018(08):42-47.

[6] 张金艳.数据挖掘在人力资源离职管理中的应用[D].北京:首都经济贸易大学,2016.

[7] PUNNOOSE R, AJIT P. Prediction of employee turnover in organizations using machine learning algorithms[J]. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, 2016, 5(9):22-26.

[8] KUMAR A, PANDEY A, KAUSHIK S. Machine learning methods for solving complex ranking and sorting issues in human resourcing[C]// SAI Y P, GRAG D. 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India: IEEE, 2017:43-47.

[9] KHERA S N, DIVYA. Predictive modelling of employee turnover in Indian IT industry using machine learning techniques[J]. Vision, 2019, 23(1):12-21.

[10] 张敏,赵宜萱.机器学习在人力资源管理领域中的应用研究[J].中国人力资源开发,2022,39(01):71-83.

[11] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// KRISHNAPURAM B, SHAH M. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016:785-794.

[12] 王志宁.基于XGBoost的员工离职预测及特征分析模型[J].数字技术与应用,2021,39(03):193-196.

[13] 陈启伟,王伟,马迪,等.基于Ext-GBDT集成的类别不平衡信用评分模型[J].计算机应用研究,2018,35(02):421-427.

[14] 叶景,李丽娟,唐臻旭.基于CNN-XGBoost的短时交通流预测[J].计算机工程与设计,2020,41(04):1080-1086.

[15] 唐容.基于特征选择的CART算法研究[D].成都:电子科技大学,2020.

[16] 袁建裕,闫春艳,叶志伟,等.离散型缺失数据填补法综合比较[J].湖北工业大学学报,2021,36(05):59-63.

[17] KHONGORZUL D, LEE S M, KIM M H. Ordinal Encoder based DNN for natural gas leak prediction[J]. Journal of the Korea Convergence Society, 2019, 10(10):7-13.

[18] 梁杰,陈嘉豪,张雪芹,等.基于独热编码和卷积神经.网络的异常检测[J]. 清华大学学报:自然科学版,2019(7):523-529.

[19] 李艳霞,柴毅,胡友强,等.不平衡数据分类方法综述[J].控制与决策,2019,34(04):673-688.

[20] 王忠震,黄勃,方志军,等.改进SMOTE的不平衡数据集成分类算法[J].计算机应用,2019,39(09):2591-2596.

作者简介：

黄　静(1965-)，女，博士，教授.研究领域：通信工程，大数据，深度学习.

郑慧慧(1997-)，女，硕士生.研究领域：智能计算与数据挖掘.