DOI:10.12158/j.2096-3203.2020.02.023

基于多尺度特征提取的电力客户欠费风险预测

葛安同[1]，谢晓慧[2]，谭忠恒[2]，李铁香[2]，张云[1]，黄睿[1]
（1. 国网江苏省电力有限公司扬州供电分公司，江苏 扬州 225009；
2. 东南大学数学学院，江苏 南京 210098）

摘　要：针对用户拖欠电费频繁发生的现象，如何利用科学方法和技术手段来预测电力客户的欠费风险，降低自身的经营风险，是供电企业急需解决的问题。文中以某地高压用户为例，分析了导致用户欠费的影响因素，从欠费频度、违约时长等多个尺度提取用户欠费风险的特征，并基于逻辑回归算法建立模型用于预测用户欠费风险。模型的评价结果表明，在所获取的用户信息不够全面的情况下，模型的预测准确率、精确率和召回率等评价指标仍较为精准，特别是模型对风险用户的识别较为灵敏。该风险模型可用于指导供电企业制定欠费风险管理对策，提高管理水平。

关键词：电力客户；电费回收风险；多尺度特征提取；风险预警；逻辑回归
中图分类号：TM71　　**文献标志码**：A　　**文章编号**：2096-3203(2020)02-0159-07

0　引言

随着现代社会的发展，电力供应已经延伸到社会的各个角落。电力市场不断扩大，用户不断增加，供用电关系也愈加复杂多变，电费回收成为供电企业至关重要的基础工作。但实际工作中电力客户欠费行为时有发生，实现电费完全回收面临巨大挑战[1]。高压用户每月用电量较大，产生电费较多，结算方式也各不相同。当用电企业发生经营不善、人为拖欠等情况时，电费按时保量回收存在困难，最终会造成国有资产流失。因此对高压用户的欠费风险进行预警分析[2]一直是电力行业的一个重要课题。

关于风险预测指标的研究起步较晚，尚缺少具有针对性的系统深入研究成果和清晰的理论框架与实际应用指导。文献[3—4]从预测力指标的角度构建电费回收风险指标体系。文献[5]提出了电力客户画像的构建方法，指出了指标应覆盖的领域，但并没有提出具体的指标，且模型预测效果不够精确(模型的召回率较低)。此类研究的共同点是，指标覆盖领域比较分散导致特征提取困难，指标没有细化导致模型预测精度不够。随着大数据行业的迅猛发展，出现了以决策树算法[6]、改进LR-Bagging算法[7]建立的电力客户欠费风险预测模型。近年来，长短期记忆网络[8]和反向传播(back propagation，BP)神经网络[9]等算法的建模是众多专家、学者关注的重点。此类研究的共同特点是强调算法的改进和参数调整设置，但是对电力客户欠费风险模型预测指标特征研究不够完善，导致模型预测的适用性和稳定性不够。

文中以某地高压用户为例，创建基于逻辑回归算法的欠费风险模型。为克服因高压用户的全面信息涉及多个领域而获取困难的不足，文中对原始数据特别是用户用电行为数据展开多尺度、精细化的欠费风险特征提取，然后用逻辑回归算法进行欠费风险建模。最后通过案例实证可知，该风险模型可以达到较为精确的预测结果。

1　风险预测技术概要

电力客户风险预测的目的是从用户的属性和用电行为数据中提取有用信息，并按照一定的算法规则识别风险用户。为实现这一目的，首先需要从电力公司内部多种渠道收集用户属性及历史用电行为等多维度数据。但多源的数据结构可能不同，质量可能参差不齐，需要对原始数据进行质量评估、清洗和整合等预处理，建立以用户编号为索引的多维度数据集。其次，用户风险特征不能由数据直观地显示，需要分析或加工提取出能影响风险预测的特征，作为预测用户风险的依据。此外，如果数据规模很大，高维度特征会带来计算效率低下以及模型预测精度降低的问题，因此还需要借助算法来筛选特征集，降低数据规模。再次，风险用户的识别规则可以由机器学习算法习得。逻辑回归算法适用于本次的研究，模型的系数可以通过参数估计法求解并迭代计算得出。最后需要考察模型的优劣，一个好的模型必须对新用户也有比较好的预测能力，因此模型的训练误差和泛化误差都必须控

收稿日期：2019-09-27；修回日期：2019-11-04

制在较低水平。通过模型的评价结果反馈调节算法的超参数，不断重复训练生成一个比较好的模型。该模型可用于新用户的风险预测，为供电企业决策提供支持，整个风险预测技术的流程见图1。

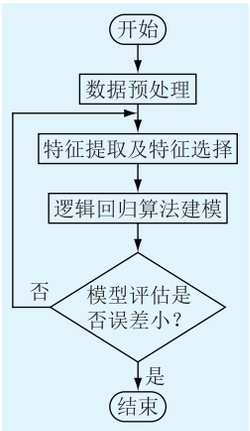

图1　风险预测流程

Fig.1　Flow chart of risk forecasting

2　数据预处理及特征工程

2.1　数据预处理

根据业务逻辑从多个渠道收集的原始高压用户数据主要有2类：包含了用户基本信息的静态数据和包含了用电行为的动态数据。其中静态数据记录了高压用户的供电单位、合同容量、运行容量、行业分类、用电类别、定价策略类型、基本电费计算方式、需量核定值和功率因数考核方式等信息。动态数据记录了用户的缴费信息，其字段包括缴费方式、用电月份、应收电费、实收电费、实收违约金、违约金起算日期和实收日期等。

这些从数据库中导出的原始数据有一定程度的冗余和污染，在对数据进行挖掘、加工前，必须对数据进行预处理。数据清洗的过程主要包括去冗余、处理缺失值和异常值、数据转换、数据合并等。例如原始数据中对于同一笔数据有冗余记录，此时可以根据特定字段展开去重复操作，删除冗余信息。

2.2　特征提取

特征提取是根据原始数据加工构建出一些具有实际意义的特征，目的是最大限度地从原始数据中挖掘有用信息[10]，以供算法和模型使用。电力客户欠费风险模型的预测效果主要取决于提取的特征的好坏[11]。数据的高质量特征提取有时甚至比算法更重要。与电力客户欠费风险相关的特征指标通常包括客户基本信息、用电行为、高压用户经营状况和银行信用状况等。因部分第三方数据获取困难，文中面对的数据仅包含了用户基本信息和用电行为信息。其中，用户基本信息包含用户的身份信息、用电属性等，这些数据一般不会经常变动。用电行为数据包含缴费欠费行为、用电量等信息，这部分数据经常变动，需要及时更新维护。该原始数据不适宜直接进行计算，所以，如何充分利用有限的数据进行特征提取是文中的研究重点。

对于高压用户来说，电费逾期次数和电费逾期天数等动态信息是衡量欠费风险的重要的因素，文中将考虑对其进行多尺度、有目的地提取特征。首先，用户的用电行为数据在动态生成中，不同时期的用电行为数据对用户风险有不同程度的影响。例如，低风险用户可能因近期的不良用电行为导致欠费风险升高，而高风险用户也可能因近期持续良好的用电行为而欠费风险降低。因此把握风险组成因素在时间上的波动特征及规律，是研究电力客气欠费风险特征提取的一个重要尺度。考虑以短期（近1个月），中期（近3个月），远期（近6个月）分时段展开用电行为特征提取。其次，通常长期电费逾期比短期电费逾期行为风险更高，因此逾期时长也是研究欠费风险特征提取的一个重要尺度。结合原始数据的逾期时长分布情况，考虑按照电费长期逾期（20 d以上），电费中期逾期（10~20 d），电费短期逾期（1~10 d）分别展开特征提取。最终电费回收风险特征如表1所示。

表1　电费回收风险特征

Table 1　Feature of tariff recovery risk

数据类型	原始字段	特征
静态数据	供电单位	供电单位
	运行容量	运行容量
	用电类别	用电类别
	定价策略类型	定价策略类型
	基本电费计算方式	基本电费计算方式
	需量核定值	需量核定值
	功率因数考核方式	功率因数考核方式
动态数据	缴费方式	缴费方式
	用电月份	最近6/3/1个月电费逾期次数
	应收电费	最近6/3/1个月电费逾期总天数
	实收电费	最近6/3个月电费逾期≥20 d次数
	实收违约金	最近6/3个月电费逾期[10,20) d次数
	违约金起算日期	最近6/3个月电费逾期[1,10) d次数
	实收日期	最近6/5/4/3/2/1个月的电费

2.3 特征选取

针对初始提取的特征可能存在信息冗余的问题[12],文中将通过特征选择来选择合适的入模型的特征,剔除不相关特征或者冗余的特征,减少有效特征的个数,有效提高模型的精确度和稳健性。

对于定性特征的选择,可以采用信息值 I 衡量该特征重要性。从而决定是否入模型。信息值表达式为:

$$I=\sum_{i=1}^{r}\left(\frac{n_{i1}}{n_1}-\frac{n_{i2}}{n_2}\right)\ln\left(\frac{n_{i1}/n_1}{n_{i2}/n_2}\right) \tag{1}$$

式中:n_{i1} 为该特征变量第 i 个属性对应的正常用户数;n_1 为样本中总的正常用户数;n_{i2} 为该特征变量第 i 个属性对应的风险用户数;n_2 为样本中总的风险用户数。一般当 $I>0.3$ 时,说明该特征有良好的预测能力,应当选入模型;当 $0.1<I<0.3$ 时,说明该特征有中等预测能力,可以选入模型[13]。

对于定量特征,可以采用相关性分析方法,以皮尔逊相关系数为指标来选择合适的特征,将相关性高的特征选入模型。

变量 X 和变量 Y 之间的皮尔逊相关系数定义为 2 个变量之间的协方差和标准差的商,估算样本的协方差和标准差,可得到皮尔逊相关系数如下:

$$\rho=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2\right]^{\frac{1}{2}}} \tag{2}$$

式中:(x_i,y_i) 为样本数据点取值;$\bar{x}$ 和 $\bar{y}$ 为样本均值;N 为样本量。

还可以采用递归消减的方法来选择特征,通过使用一个逻辑回归作为基模型来进行多轮训练,逐步递归地减少候选特征并对比预测误差是否显著减少,来考察每个候选特征的实际预测能力,消除预测力不足的特征,最终得到一个最优的特征集。该过程可以通过 Python 的机器学习包来自动实现。

综合以上各种特征选择方法,可以得到最终的定量特征集。

2.4 数据编码及归一化处理

为了将定性特征引入算法参与计算,需要将定性特征转换为定量特征,通常采用独热编码[14]方式进行实现。也就是说,若某定性特征有 N 种取值,则将这一个特征扩展为 N 种特征,当原始特征值为第 i 种定性值时,第 i 个扩展特征赋值为 1,其他扩展特征赋值为 0。

另外,不同定量特征的取值属于不同量纲,不能够直接放在一起直接进行计算,必须对其进行数据归一化处理。数据归一化方法为:

$$z=\frac{x-\mu}{\sigma} \tag{3}$$

式中:μ 为训练数据的均值;σ 为训练数据的标准差。

3 风险预测模型的构建

文中从电力用户的用电数据出发,分析欠费风险特征,运用逻辑回归算法模型对用户特征数据进行训练、学习,形成电力客户欠费风险预警模型,分析并预测电力客户不能及时缴纳电费的可能性,形成电力客户风险评价。

逻辑回归是一种非线性概率模型,被广泛地运用于风险建模分析[15]。其表达式如下:

$$P(Y=1\mid\boldsymbol{X}=\boldsymbol{x})=\frac{e^{(\boldsymbol{w}\cdot\boldsymbol{x}+b)}}{1+e^{(\boldsymbol{w}\cdot\boldsymbol{x}+b)}} \tag{4}$$

式中:概率 $P(Y=1\mid\boldsymbol{X}=\boldsymbol{x})$ 为用户的电费欠费概率;$\boldsymbol{X}\in\mathbf{R}^n$ 为含有 n 个特征的输入向量;$\boldsymbol{w}\in\mathbf{R}^n$ 为权重向量参数;b 为偏置参数。

通过数据编码和归一化处理后的数据,可以作为特征输入逻辑回归算法进行训练,迭代计算出模型参数。

3.1 模型训练

将某年度 1—6 月份的数据集按照随机方式分割为训练集和验证集,其中训练集是用来训练模型和确定模型参数的,而验证集则用于评价模型的性能,从而选择出最优模型。文中训练集和验证集的分割比例为 7:3。

模型训练、学习的目的是用给定的训练数据计算并求解逻辑回归模型参数的数值解。逻辑回归的参数估计方法采用最大似然法,数值求解算法可以采用梯度下降法、拟牛顿法等优化算法迭代求解。运用最大似然法求解模型参数过程如下。

记用户欠费概率 $P(Y=1\mid\boldsymbol{X}=\boldsymbol{x})=\pi(x)$,不欠费概率 $P(Y=0\mid\boldsymbol{X}=\boldsymbol{x})=1-\pi(x)$,则似然函数和对数似然函数分别为:

$$L=\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{5}$$

$$\ln(L(\boldsymbol{w}))=\sum_{i=1}^{N}[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-\ln(1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i+b})] \tag{6}$$

只要对对数似然函数 $\ln(L(\boldsymbol{w}))$ 求极大值,就可得到关于参数 $\boldsymbol{w}$ 的极大似然估计值。参数求解问题就转变为以对数似然函数为目标函数的最优化问题:

$$\boldsymbol{w}_{\mathrm{MLE}}=\arg\max_{\boldsymbol{w}}\ln(L(\boldsymbol{w})) \tag{7}$$

数值求解算法可以采用梯度下降法、拟牛顿法等优化算法，通过迭代得到上述问题的最优解[16]。

3.2 模型评价

风险模型参数确定后，模型表达式也随之被确定。将用户的特征数据输入风险模型，根据式(4)可以得到用户的风险概率。在逻辑回归的预测中，将上述风险概率大于0.5的用户输出为正类(风险用户)，否则输出为负类(正常用户)。将测试数据集输入模型，可以得到模型预测的类别，然后与原始类别标签进行对比，评价模型的性能。

模型的评价指标一般有混淆矩阵、准确率、精确率、召回率和F_1值等。混淆矩阵如表2所示。

表2 混淆矩阵

Table 2 Confusion matrix

实际	预测	
	负类	正类
负类	T_N	F_P
正类	F_N	T_P

其中，T_N为负类判定为负类；T_P为正类判定为正类；F_N为正类判定为负类；F_P为负类判定为正类。准确率体现了模型对样本的预测精度，表达式为：

$$A=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{8}$$

但在非平衡数据中，因为负类样本占大多数，正类样本仅占少数，仅仅使用准确率评价模型有时会失真，不能全面评价模型的性能。因此还需要引入其他评价指标。精确率表示在模型预测为正类的样本中，实际是正类的比率，体现了模型识别正类样本的准确度。其表达式如下：

$$P=\frac{T_P}{T_P+F_P} \tag{9}$$

召回率表示实际为正类的样本中，被模型预测为正类的比率，体现了模型识别正类样本的敏感度。其表达式如下：

$$R=\frac{T_P}{T_P+F_N} \tag{10}$$

F_1值是精确率和召回率的调和平均，综合评价了模型的性能。其表达式如下：

$$F_1=\frac{2T_P}{2T_P+F_P+F_N} \tag{11}$$

通过准确率、精确率、召回率和F_1值等指标，可以综合选出预测精度，识别正类样本的准确度和敏感度均优的模型。

4 实例分析

为验证特征提取的有效性和算法的合理性，以某地区高压用户为例，进行建模实例分析。样本训练集有研究对象9 427户，其中60户为风险用户，验证集有4 041户，其中26户为风险用户。

定性特征依据其信息值进行特征选择。以“供电单位”这个定性特征为例，考察某地区的5所供电单位的用户构成并计算信息值I来进行特征选择，该特征的具体评价结果如表3所示。

表3 信息值

Table 3 Information value

供电单位	正常用户(1)/%	风险用户/%	信息值 I	该单位用户(2)/%	该单位风险用户(3)/%
单位1	32.4	11.6	0.212	32.2	0.2
单位2	18.4	64	0.567	18.7	2.2
单位3	18.9	11.6	0.035	18.8	0.4
单位4	11.2	4.7	0.057	11.1	0.3
单位5	19.2	8.1	0.095	19.1	0.3
合计			0.966		0.6

注：(1) 正常用户、风险用户分别为各个供电单位正常、风险用户数占全体样本正常、风险用户总数的百分比；(2) 该单位用户为各个供电单位用户数占全体样本用户总数的百分比；(3) 该单位风险用户为各个供电单位内部的风险用户数占该单位用户总数的百分比。

可知，全体样本的风险用户占比为0.6%。但不同供电单位内部的风险用户占比有差异，例如供电单位2内部风险用户占比达到2.2%，而供电单位1内部仅为0.2%，说明风险用户的分布与供电单位相关，该特征能为风险预测提供价值，因此直观上分析“供电单位”这个特征应当选入模型。同时从数值上该特征的信息值I为0.966，显著大于入模型的标准0.3，这也说明了信息值I可以作为衡量定性特征预测能力的依据。经过计算，最终选入模型的特征及其信息值见表4。

表4 特征选择

Table 3 Feature selection

特征	信息值 I
供电单位	0.966 145 660
用电类别	0.963 609 310
缴费方式	0.414 282 791
基本电费计算方式	0.126 757 932
功率因数考核方式	0.123 368 080
定价策略类型	0.117 094 040

定量特征与风险类别的相关系数如图2所示。从图中可以看出，除了运行容量及每月电费和用户风险类别的的相关系数绝对值较小外，其他特征在

相关程度上比较明显。结合使用 Python 库的递归消减法的结果,选择了 19 个定量特征参与逻辑回归模型训练。

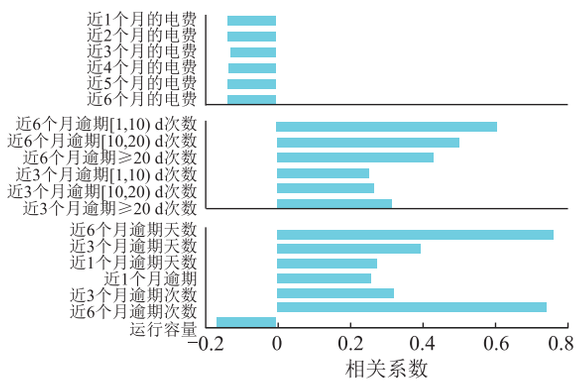

图 2 定量特征与风险类别的相关系数

Fig.2 Pearson correlation between quantitative characteristics and risk categories

逻辑回归的超参数设置如表 5 所示。

表 5 逻辑回归参数设置

Table 5 Logistic Regression parameters setting

参数	说明	取值
penalty	损失函数的正则化项	l1
solver	模型优化方法	liblinear
class_weight	类别权重	balanced
C	正则化项强度的倒数	0.1
max_iter	最大迭代次数	1 000

将经过编码和归一化后的特征输入逻辑回归进行训练,风险模型参数计算结果如表 6 所示。

用占总样本数据集 30%的同期验证数据集输入逻辑回归模型,可以得到预测的类别,然后与原始类别进行对比,评价模型的性能。混淆矩阵的结果如表 7 所示。准确率、精确率、召回率和 F_1 值的结果如表 8 所示。

表 7 和表 8 的模型评价结果表明:验证数据集共 4 041 组用户,在逻辑回归的模型预测结果中,除了 1 组正常用户被错误预测为风险用户外,其他所有用户预测均准确,准确率达到 99.97%。召回率达到 100%,说明模型对风险用户的预测灵敏度高,可以有效的识别有欠费风险倾向的用户。

用后移 6 个月的未来数据集(同年度 7—12 月份数据)输入风险模型进行预测,此数据集有研究对象 13 473 户,其中 96 户为风险用户。验证结果以混淆矩阵的方式在表 9 呈现。

未来数据集验证的准确率、精确率、召回率和 F_1 值结果如表 10 所示。

以上模型验证结果显示,在验证数据集和未来数据集下的模型预测准确率、精确度和召回率均比

表 6 模型参数计算结果

Table 6 Model parameter calculation result

特征	取值	特征	取值
供电单位 1	−0.554 1	运行容量	0.041 2
供电单位 2	−0.314 2	最近 6 个月逾期次数	0.190 3
供电单位 3	−0.591 4	最近 3 个月逾期次数	0.058 1
供电单位 4	−0.404 9	最近 1 个月是否逾期	0.045 9
供电单位 5	−0.443 1	最近 1 个月逾期天数	0.046 4
用电类别 1	−0.851 5	最近 3 个月逾期总天数	0.058 2
用电类别 2	−0.004 6	最近 6 个月逾期总天数	1.844 1
用电类别 3	−0.197 3	最近 3 个月逾期≥20 d 次数	0.058 2
用电类别 4	−0.051 6	最近 3 个月逾期[10,20) d 次数	0.046 9
用电类别 5	−0.067 9	最近 3 个月逾期[1,10) d 次数	0.191 4
用电类别 6	−0.016 3	最近 6 个月逾期≥20 d 次数	0.079 8
用电类别 7	−0.782 5	最近 6 个月逾期[10,20) d 次数	0.107 0
用电类别 8	−0.453 8	最近 6 个月逾期[1,10) d 次数	0.151 4
用电类别 9	0.361 5	最近 6 个月的电费	0.079 1
用电类别 10	−0.243 7	最近 5 个月的电费	0.064 7
定价策略类型 1	−0.776 9	最近 4 个月的电费	0.146 1
定价策略类型 2	−1.530 9	最近 3 个月的电费	0.004 9
基本电费计算方式 1	−1.530 9	最近 2 个月的电费	0.038 3
基本电费计算方式 2	−0.595 3	最近 1 个月的电费	0.063 8
基本电费计算方式 3	−0.181 5	缴费方式 1	1.228 6
功率因数考核方式 1	−0.753 8	缴费方式 2	1.079 1
功率因数考核方式 2	−1.554 0		

表 7 验证数据预测结果

Table 7 The prediction result of testing data

实际	预测	
	正常用户	风险用户
正常用户	4 014	1
风险用户	0	26

表 8 模型评价结果

Table 8 Model evaluation result

指标	数值/%	说明
准确率	99.97	预测准确率
精确率	96.30	预测为风险用户中真正的风险用户比例
召回率	100.00	所有风险用户被正确预测的比例
F_1 值	98.11	精确率和召回率的调和平均

表 9 未来数据验证结果

Table 9 The prediction result of future data

实际	预测	
	正常用户	风险用户
正常用户	13 364	13
风险用户	10	86

表 10 未来数据模型评价结果

Table 10 Model evaluation result of future data

指标	数值/%	指标	数值/%
准确率	99.83	召回率	89.58
精确率	86.87	F_1值	88.21

较高,表明该电力客户欠费风险预警模型性能稳定,具有较好的推广应用价值。加强电力客户欠费的风险管理,应该充分运用电力客户欠费风险评价模型的预测结果制定科学的风险控制策略[17]。

5 结语

文中着眼于电力客户欠费风险预测这一课题,利用高压客户历史用电数据,运用多尺度的欠费风险特征提取技术克服该案例客户信息不足的困难,采用相关性分析法和递归消减法来筛选特征,并基于逻辑回归算法构建了电力客户欠费风险评价模型,取得了显著的成效。该模型有效地预测了用户欠费风险,筛选出潜在的电力公司风险客户。下一步可以考虑在获取更全面的客户信息后,结合文中的细化特征提取方法建立指标体系,针对电力客户欠费风险预测问题展开进一步探讨。

参考文献:

[1] 王宗伟,金鹏,卜晓阳,等. 总部级电费回收营销稽查监控研究与应用[J]. 自动化与仪器仪表,2019(4):184-187.
WANG Zongwei, JIN peng, BU Xiaoyang, et al. Research and application of supervision and control of electricity recycling marketing inspection at ministry level[J]. Automation & Instrumentation,2019(4):184-187.

[2] 汤克艰,王树全,孙强. 新型营销电费回收控制研究及其应用[J]. 中国电力,2016,49(11):70-74.
TANG Kejian, WANG Shuquan, SUN Qiang. Research and application of new type of electricity tariff recovery control measures[J]. Electric Power,2016,49(11):70-74.

[3] 赵永良,秦萱,吴尚远,等. 基于数据挖掘的高压用户电费回收风险预测[J]. 电力信息与通信技术,2015,13(9):57-61.
ZHAO Yonghang, QIN Xuan, WU Shangyuan, et al. Electricity recovery risk prediction of high-voltage customers based on data mining[J]. Electric Power Information and Communication Technology,2015,13(9):57-61.

[4] 邹云峰,邓君华,徐超,等. 高压企业客户电力信用综合评价体系及应用[J]. 电力需求侧管理,2019,21(3):37-41.
ZOU Yunfeng, DENG Junhua, XU Chao, et al. Comprehensive evaluation system and application of high voltage power customers' credit[J]. Power Demand Side Management, 2019, 21(3):37-41.

[5] 裘华东,涂莹,丁麒. 基于标签库系统的电力企业客户画像构建与信用评估及电费风险防控应用[J]. 电信科学,2017,33(S1):206-213.
QIU Huadong, TU Ying, DING Qi. Construction of power customer portrait and its credit evaluation and electricity fee risk control based on tag library system[J]. Telecommunications Science,2017,33(S1):206-213.

[6] 黄文思,郝悍勇,李金湖,等. 基于决策树算法的电力客户欠费风险预测[J]. 电力信息与通信技术,2016,14(1):19-22.
HUANG Wensi, HAO Hanyong, LI Jinhu, et al. Prediction of power customer arrear risk based on decision tree algorithm[J]. Electric Power Information and Communication Technology, 2016,14(1):19-22.

[7] 吴漾,朱州. 基于特征选择改进 LR-Bagging 算法的电力欠费风险居民客户预测[J]. 电子产品世界,2017,24(4):70-75.
WU Yang, ZHU Zhou. The arrears risk prediction of power residential customers based on LR-Bagging algorithm improved by feature selection[J]. Power Management,2017,24(4):70-75.

[8] 谢林枫,钱立军,季聪,等. 基于长短期记忆网络算法的电费回收风险预警[J]. 电力工程技术,2018,37(5):98-103.
XIE Linfeng, QIAN Lijun, JI Cong, et al. Application of long short-term memory network algorithm in tariff recovery risk early warning for large power customers[J]. Electric Power Engineering Technology,2018,37(5):98-103.

[9] 王宇哲,雷霞,陈晓盛,等. 基于 BP 神经网络电力大客户信用等级评价研究[J]. 电力需求侧管理,2015,17(5):49-53.
WANG Yuzhe, LEI Xia, CHEN Xiaosheng, et al. Large power customer credit rating based on BP neural network[J]. Power Demand Side Management,2015,17(5):49-53.

[10] 李广明,诸唯君,周欢. P2P 网络融资中贷款者欠款特征提取实证研究[J]. 商业时代,2011(1):41-42,58.
LI Guangming, ZHU Weijun, ZHOU Huan. Lender key characteristics extraction of lagging payment in P2P network financing[J]. Commercial Times,2011(1):41-42,58.

[11] 赵红,丁茹. 互联网金融企业用户流失预测特征提取方式对比研究[J]. 预测,2018,37(6):61-66.
ZHAO Hong, DING Ru. Comparing feature constructing methods for customer churn prediction: a research in the field of internet finance[J]. Forecasting,2018,37(6):61-66.

[12] 赖学方,贺兴时. 最小冗余最大分离准则特征选择方法[J]. 计算机工程与应用,2017,53(12):70-75.
LAI Xuefang, HE Xingshi. Method based on minimum redundancy and maximum separability for feature selection[J]. Computer Engineering and Applications,2017,53(12):70-75.

[13] 马姆杜·雷法特. 信用风险评分卡研究基于 SAS 的开发与实施[M]. 王松奇,林治乾译. 北京:科学文献出版社,2013.
MAMDOUH Refaat. Credit risk scorecards: development and implementation using SAS[M]. WANG Songqi, LIN Zhiqian Transalted. Beijing: Sciences Academic Press,2013.

[14] 李河,麦劲壮,肖敏,等. 哑变量在 Logistic 回归模型中的应用[J]. 循证医学,2008,8(1):42-45.
LI He, MAI Jingzhuang, XIAO Min, et al. Application of dummy variable in logistic regression models[J]. The Journal of Evidence-Based Medicine,2008,8(1):42-45.

[15] 郭小波,王婉婷,周欣. 我国中小企业信贷风险识别因子的有效性分析——基于北京地区中小企业的信贷数据[J]. 国际金融研究,2011(4):62-67.

GUO Xiaobo,WANG Wanting,ZHOU Xin. A study in the efficiency of credit risk indicators for SMEs lending——Evidence from SMEs lending data of Beijing[J]. Studies of International Finance,2011(4):62-67.

[16] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012.
LI Hang. Statistical learning method[M]. Beijing:Tsinghua University Press,2012.

[17] 雷君召,禹冰. 电费回收及风险防范对策[J]. 煤炭技术,2013(6):280-282.
LEI Junzhao,YU Bing. Electricity recovery and risk prevention countermeasures[J]. Coal Technology,2013(6):280-282.

作者简介:


葛安同

葛安同(1990),男,硕士,工程师,从事电力营销管理、电力大数据分析应用工作(E-mail:504305924@qq.com);

谢晓慧(1988),女,硕士在读,研究方向为应用统计;

谭忠恒(1995),男,硕士在读,研究方向为统计学。

Arrears risk prediction of large power customers based on multi-scale feature extraction

GE Antong[1], XIE Xiaohui[2], TAN Zhongheng[2], LI Tiexiang[2], ZHANG Yun[1], HUANG Rui[1]

(1. State Grid Yangzhou Power Supply Company of Jiangsu Electric Power Co.,Ltd.,Yangzhou 225009,China;
2. School of Mathematics,Southeast University,Nanjing 210098,China)

Abstract:In view of the frequent occurrence of electricity arrears, how to use scientific methods and technical means to predict the arrears risk of power customers and reduce business risk is an urgent problem for the Power Grid Corp. Taking high-voltage customers from a certain areaas an example, this paper analyzes the factors affecting the recovery of electricity tariff, and extracts therisk features from multiple scales such as arrears frequency and duration, and then establishes a model based on logistic regression algorithm to predict the risk of user arrears. The model's evaluation results show that the indicators such as the prediction accuracy, precision and recall rate are still relatively accurate when user information is not comprehensive enough. The model issensitive to the identification of risk users, and can be used to guide the power grid corp to formulate arrears risk management policiesand improving their management capabilities.

Keywords:power customer;arrears risk;multi-scale feature extraction;risk early warning;logistic regression

(编辑 方晶)