

一种城镇长租低压居民客户识别方法

李悦,程雅梦,马玉龙,陈欣

(国网江苏省电力有限公司 电力科学研究院,南京 210000)

The identification model of urban long-rent low voltage customer

LI Yue, CHENG Yameng, MA Yulong, CHEN Xin

(Electric Power Research Institute, State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210000, China)

摘要:出租房客户电费回收难等问题是长期困扰电力公司服务工作的痛点,预付费缴费模式能很好地解决这一问题。通过专家咨询法形成城镇长租低压居民客户分析指标,利用逻辑回归法构建城镇长租低压居民客户识别模型。结果表明,模型能准确定位出租房客户,较好地支撑预付费模式推广,辅助提高电费回收率。

关键词:出租房客户;专家咨询法;逻辑回归法;预付费模式

Abstract: The difficulty of recovering tariff of the rental housing customer puzzles the electric power company service for a long time. Prepaid model have certain help to solve the problem. Using expert consultation method to form urban long-rent low voltage customer analysis indicators. The logistic regression method is used to construct recognition model. Practice shows that the model can accurately locate the rental customers, support prepaid model promotion, auxiliary improve tariff recovery.

Key words: rental housing customer; expert consultation method; logistic regression method; prepaid model

中图分类号:TM727;F407.61 文献标志码:B

0 引言

我国出租房客户具有数量庞大、流动性高等特点,此类客户电费回收难是长期困扰电力公司服务工作的痛点,而租户欠缴电费也对房东个人征信带来了一定的压力^[1-2]。为减少此类问题给电力公司和房东带来的影响,国家电网公司倡导采用预付费的缴费模式进行管理^[3],而如何准确识别出租房客户,成为亟待解决问题。

出租房客户按照租期可分为长租(按月或季度租)和短租(按日租),一般情况下,短租户由房东缴纳电费,长租户由租客或房东缴纳电费,而农村客户出租户相对较少,非居民客户出租原因较为复杂,难以研究其规律。因此,本文的研究对象主要为城镇长租低压居民客户(以下简称“租户”)。

本文基于专家咨询法形成租户分析指标体系,利用营销业务应用系统客户基础数据、用电数据以及部分外部数据,基于逻辑回归法构建租户识别模型,对预测结果为租户的客户开展预付费模式推广,结果表明,该模型的准确率较高,能较好地支撑预付费模式推广,辅助提高电费回收率。

1 租户识别模型构建

1.1 租户分析指标体系构建

租户有一定的流动性,用电特征和缴费习惯与一

般居民用户存在差异。考虑到专家咨询法能充分利用专家的知识和经验,灵活度较高,本文基于专家咨询法建立租户识别指标体系。通过对5位客户服务领域的专家进行访谈,考虑数据的可获取性,初步筛选出表1所示的指标,作为判断客户是否为租户的指标。

表1 租户分析指标体系

Table 1 The index of analysing rental housing customer

指标编号	指标名称	选取原因
指标1	地市	不同城市的出租率不同,一、二线城市出租率高于三、四线城市
指标2	区县	市区出租率高于其他区县的出租率
指标3	是否高出租房率小区	出租率高的小区,房屋出租的可能性更高
指标4	户主房产数量	户主房产数量大于1套,房屋出租的可能性更高
指标5	年空置天数	长租房换租客期间,会出现空置期(无用电量)。而正常用电的客户一般只出现用电量低的情况
指标6	中介来电次数	拨打95598电话咨询电量电费且记录为中介来电的,可能存在房屋出租或买卖
指标7	户号绑定账户变更次数	若某一用户的户号被不同账号绑定/解绑,且变更较为频繁,可能为租客操作
指标8	是否代扣	租客往往不会办理代扣

1.2 数据预处理

数据预处理主要包括冗余变量和脏数据处理。冗余变量识别方法包括相关性分析、卡方检验等,本文采用相关性分析法筛选冗余变量^[4]。相关系数描述的是2个事物之间的相关程度,如表2所示,相关系数的计算方法及判别标准如下

收稿日期:2019-07-01;修回日期:2019-08-22

基金项目:国家自然科学基金(61772131)

This work is supported by National Natural Science Foundation of China(No.61772131)

$$r_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \quad (1)$$

式中: $r_{x,y}$ 为变量 X 和 Y 的相关系数; $Cov(X,Y)$ 为变量 X 和 Y 的协方差; σ_x 为变量 X 的方差; σ_y 为变量 Y 的方差。

表2 相关系数判别标准表
Table 2 The standard of coefficient of association

相关系数	关联强度
0.8~1.0	极强相关
0.6~0.8	强相关
0.4~0.6	中等程度相关
0.2~0.4	弱相关
0~0.2	极弱相关或无相关

由于过多的变量可能会带来干扰和过度拟合问题,影响模型的稳定性和计算效率,因此,在建立模型前需剔除冗余变量^[5]。本文对强相关和极强相关变量进行剔除。

输入数据中的脏数据主要包括缺失值、异常值等。处理缺失值的方法一般有删除或者填充,本文采用均值填充方法补足所需指标数据。用于填充缺失值的均值计算方法如下

$$x_i = \frac{x_1 + \dots + x_{i-1} + x_{i+1} + \dots + x_k}{k-1} \quad (2)$$

式中: x_i 为变量 X 中的缺失值; k 为变量 X 包含的数据个数。

将与平均值的偏差超过3倍标准差的测定值视为异常值。常用的异常值处理方法主要有:删除、转换变量、替换、分开处理^[6],本文采用删除法对异常值进行处理。

1.3 基于逻辑回归法的租户识别模型构建

租户识别可视为一个二分类问题,本文选择逻辑回归法进行租户识别。逻辑回归法^[7-8]是一种对2群样本的分界面进行拟合的机器学习方法,它将输入变量经过线性回归和sigmoid函数映射到0~1区间,预测值大于等于0.5,该样本分类至类别1;预测值小于0.5,该样本分类至类别0。假设有 m 个观测样本,每个样本有 n 个特征。样本 x 分类至类别1和类别0的概率分别为

$$p(y=1|x; w) = h_w(x) = \frac{e^{-(wx+w_0)}}{1+e^{-(wx+w_0)}} \quad (3)$$

$$p(y=0|x; w) = 1 - h_w(x) = 1 - p(y=1|x; w) = \frac{1}{1+e^{-(wx+w_0)}} \quad (4)$$

式中: $p(y=1|x; w)$ 为该样本分类至类别1的概率; $p(y=0|x; w)$ 为该样本分类至类别0的概率; y 为类别值,取0或1; $x \in R^n$ 为含有 n 个特征的输入向量; $w \in R^n$ 为权重向量参数; w_0 为偏置参数。

对每个样本

$$p(y|x; w) = h_w(x)^y \times (1 - h_w(x))^{1-y} \quad (5)$$

取观测样本的联合概率作为似然函数

$$L(w) = \prod_{i=1}^m (h_w(x^i))^{y^i} \times (1 - h_w(x^i))^{1-y^i} \quad (6)$$

$$\log L(w) = \sum_{i=1}^m (y^i \times \log(h_w(x^i)) + (1 - y^i) \times \log(1 - h_w(x^i))) \quad (7)$$

通过对 $\log L(w)$ 求极大值,得到参数 w 的极大似然估计。利用梯度下降法,代价函数为

$$J(w) = -\frac{\log L(w)}{m} = -\frac{1}{m} \sum_{i=1}^m (y^i \times \log(h_w(x^i)) + (1 - y^i) \times \log(1 - h_w(x^i))) \quad (8)$$

$$\begin{aligned} \frac{\partial J(w)}{\partial w} &= -\frac{1}{m} \sum_{i=1}^m \left(\frac{y^i}{h_w(x^i)} \times \frac{\partial h_w(x^i)}{\partial w} + \frac{1-y^i}{1-h_w(x^i)} \times \left(-\frac{\partial h_w(x^i)}{\partial w} \right) \right) = -\frac{1}{m} \sum_{i=1}^m \left(\frac{y^i - h_w(x^i)}{h_w(x^i) \times (1 - h_w(x^i))} \times \frac{\partial h_w(x^i)}{\partial w} \right) \\ &= -\frac{1}{m} \sum_{i=1}^m \left(\frac{y^i - h_w(x^i)}{h_w(x^i) \times (1 - h_w(x^i))} \times h_w(x^i) \times (1 - h_w(x^i)) \times x^i \right) = -\frac{1}{m} \sum_{i=1}^m ((y^i - h_w(x^i)) \times x^i) \end{aligned} \quad (9)$$

根据梯度逐步调整参数 w ,直至 w 不再减小或者减小的幅度低于预设值,则 w 为最优解

$$w_{j+1} = w_j - \eta \times \frac{\partial J(w)}{\partial w_j} \quad (10)$$

式中: η 为调整步长, $\eta \in (0, 1)$; w_j 为第 j 个特征的权重。

1.4 模型效果评估

本文用准确率、精准率、召回率和 F_1 值评估模型分类效果,准确率即预测结果正确的户数与输入模型总户数的比值,精准率即预测为租户的数据中预测正确的户数,召回率即实际为租户的数据中预测正确的户数, F_1 值是精准率和召回率的调和平均值。准确率、精准率、召回率、 F_1 值越大,表明模型分类效果越好。表3为模型分类结果表。

表3 模型分类结果
Table 3 The classification of the model

分类	实际为非租户	实际为租户
预测为非租户	TN	FN
预测为租户	FP	TP

$$P_{acc} = \frac{TN + TP}{TN + FN + FP + TP} \quad (11)$$

$$P_{pre} = \frac{TP}{TP + FP} \quad (12)$$

$$P_{rec} = \frac{TP}{TP + FN} \quad (13)$$

$$F_1 = \frac{2P_{pre} \times P_{rec}}{P_{pre} + P_{rec}} \quad (14)$$

式中: P_{acc} 为准确率; P_{pre} 为精准率; P_{rec} 为召回率; F_1 为精准率和召回率的加权调和平均值。

2 模型计算及研究应用

本文基于专家咨询法得到城镇长租低压居

民客户分析指标体系,利用逻辑回归法,结合某省公司1 500户城镇低压居民客户(含500户租户和1 000户非租户)具体数据,构建城镇长租低压居民客户识别模型,具体过程如下。

2.1 数据采集

梳理模型所需数据并从营销业务应用系统、用电信息采集系统、95598系统中提取所需的数据,如表4所示。

表4 某省电力公司租户分析指标取值

Table 4 The index value of analysing rental housing customer of an electric power company

指标编号	数据选取
指标1	取2017年该省各地市地区生产总值,数据来源于2018某省统计年鉴
指标2	取2017年该省各区县地区生产总值,数据来源于2018某省统计年鉴
指标3	利用网络爬虫获取链家网、58同城网中各地区的租房信息,统计各地区一年内挂出的出租房数量排名前20的小区,计为高出租率小区,取值为1,其他小区取值为0
指标4	基于电力营销业务应用系统中的身份证信息,统计与该户户主身份证号相同的户号数
指标5	基于用电信息采集系统中的日用电量数据,统计2018年用电量为0的天数
指标6	统计2018年该户产生的95598工单中记录为中介来电的工单数
指标7	统计2018年内绑定该户号的微信号和支付宝号数和解除该户号绑定关系的微信号和支付宝号数之和
指标8	取营销业务应用系统中代扣状态数据

2.2 数据预处理

经分析,1 500户相关的数据中脏数据主要包括10个缺失值和1个异常值,缺失值通过均值填充,异常值直接删除。

对处理脏数据后的数据进行相关性分析,结果显示,区县与地市的相关性系数为0.764,属于强相关性,因此,在计算时需剔除区县指标,指标数量由8个变为7个。

2.3 模型计算

将1 499户数据按70%训练集和30%测试集进行分割,其中,训练集含700户非租户和349户租户,共1 049户,该数据集用于训练模型和确定模型参数;测试集含300户非租户和150户租户,共450户,用于评价模型性能。

利用SPSS软件进行二元逻辑回归计算,“是否高出租率小区”和“户号绑定账户变更频率”的显著性水平(P 值)分别为0.06和0.681,大于0.05,说明这2个指标对结果没有显著性影响,再对剩余5个变量进行二元逻辑回归计算,得到预测模型参数如表5所示。

表5 模型参数计算结果

Table 5 The result of model parameters

变量名称	地市	户主房产数量	年空置天数	中介来电次数	是否办理电费代扣业务	偏置参数
参数	0.088	1.317	0.608	3.201	1.097	-6.685

训练数据预测结果如表6所示,由该表可知,模型准确率为90.75%,精准率为89.62%,召回率为81.66%, F_1 值为85.46%。

经分析,实际为非租户,但是模型预测为租户的原因可能包括:①房主房产较多;②该户可能存在出差、探亲、休假情况,且外出期间关闭空开,导致年空置天数较多,其用电特征与租户类似。实际为租户,但是模型预测为非租户,可能是因为该租户租期较长,其用电、交费特征与非租户接近。

表6 训练数据预测结果

Table 6 The predict result of training data

	户		
分类	实际为非租户	实际为租户	总计
预测为非租户	667	64	731
预测为租户	33	285	318
总计	700	349	1 049

2.4 模型验证

利用测试数据对模型进行验证,预测结果如表7所示,模型准确率为88.44%,精准率为78%,召回率为86.03%, F_1 值为81.82%。

表7 测试数据预测结果

Table 7 The predict result of test data

	户		
分类	实际为非租户	实际为租户	总计
预测为非租户	281	19	300
预测为租户	33	117	150
总计	314	136	450

3 研究应用

本部分基于文中提出的租户识别方法开展预付费缴费模式推广,通过推广结果,验证本文提出的租户识别方法的有效性。

在某经济发达城市随机挑选1 000户城镇低压居民客户,从营销业务应用系统、用电信息采集系统、95598系统等系统中获取模型相关数据。经统计,无缺失和异常数据,其中,405户客户已绑定微信或支付宝。利用本文建立的识别模型进行分析,预测结果为租户的客户共358户。

通过上门服务和电子服务渠道向1 000户客户推广预付费模式,其中,对405户已绑定微信或支付宝的客户,发送点对点消息推广, (下转第92页)