



(12)发明专利申请

(10)申请公布号 CN 107578277 A

(43)申请公布日 2018.01.12

(21)申请号 201710736117.X

(51)Int.Cl.

(22)申请日 2017.08.24

G06Q 30/02(2012.01)

G06Q 50/06(2012.01)

(71)申请人 国网浙江省电力公司电力科学研究院

地址 310014 浙江省杭州市下城区朝晖八区华电弄1号

申请人 国网浙江省电力公司
国网浙江省电力公司绍兴供电公司

(72)发明人 王庆娟 张维 吕诗宁 欧阳柳
丁麒 徐家宁 俞佳莉 陈齐瑞
沈然 骆云江 叶珺歆 赵融融
张一池 程清 吴越人 徐千
张梁 许海霄 李海峰 陈楚楚

(74)专利代理机构 浙江翔隆专利事务所(普通合伙) 33206

代理人 戴晓翔 王晓燕

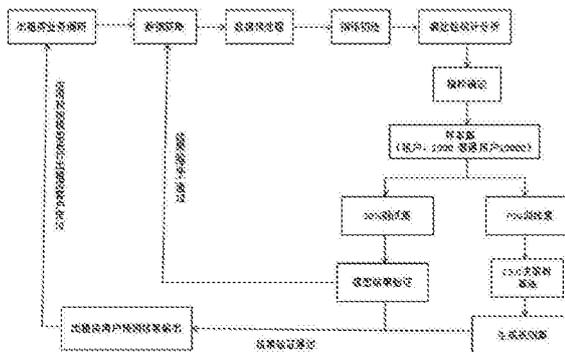
权利要求书2页 说明书10页 附图2页

(54)发明名称

用于电力营销的出租房客户定位方法

(57)摘要

用于电力营销的出租房客户定位方法,涉及出租房客户定位方法。目前,精准定位出租房客户,并配合服务策略规范其用电行为,实现精准营销,刻不容缓。本发明包括以下步骤:1)指标初选;2)指标分析,包括连续变量分析和离散变量分析;3)指标确定;4)出租房客户预测模型构建;5)根据确定的出租房客户预测模型,进行出租房预测结果输出,定位出租房客户。本技术方案首先对出租房客户进行特征分析,从基础信息、交费行为、用电特征三大维度出发,提炼出多个影响出租房客户分析的指标作为预测指标集,通过C5.0决策树算法构建出租房用户预测模型,准确定位出租房用户,实现精准营销,规范出租房客户用电行为,提高电费回收效率、降低安全隐患。



1. 用于电力营销的出租房客户定位方法,其特征包括以下步骤:

1) 指标初选,根据出租房业务调研结果,获取建模所需目标数据群,并对获取的数据进行数据的预处理,初步选取建模指标;

从基础信息、交费行为、用电特征三个维度提炼出8个指标进行模型构建,分别为城乡类别、年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比、设定时间内不同收款部门数及近一年支付宝交费次数;其中城乡类别为:城镇、农村;过年期间电量占比为:过年期间用电量/全年用电量*100%;清明节假期电量占比为:清明假期用电量/4月用电量*100%;端午节假期用电量占比:端午假期用电量/5月用电量*100%;

2) 指标分析,包括连续变量分析和离散变量分析;

201) 连续变量分析:将出租户和普通用户的年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比这5个连续变量的均值进行分析,得到出租户与普通用户对应指标的差别程度;

202) 离散变量分析:对出租户和普通用户的近一年不同收款部门数这一指标进行分析,其中,租户各收款部门变化次数客户数占比 = 各收款部门变化次数客户数/出租房总数*100%;普通用户各收款部门变化次数客户数占比 = 各收款部门变化次数客户数/普通用户总数*100%;得到出租户与普通用户对应指标的差别程度;

3) 指标确定

根据指标分析结果对初选指标进行调整,选择出租户与普通用户差别程度大于设定值的对应指标为确定指标,确定最终建模指标;

4) 出租房客户预测模型构建

401) 根据确定的建模指标,随机筛选样本集中70%作为训练集,30%作为测试集构建出租房客户预测模型;

402) 生成规则集,利用C5.0决策树算法,对训练集进行训练和学习生成出租房客户预测模型规则集并获得各指标对模型的影响程度及预测混淆矩阵;

403) 根据训练集模型预测结果,将模型应用到测试集上进行模型测试,判断训练集和测试集的预测效果否达到了理想效果,若是,则确定该模型为出租房客户预测模型,否则,返回步骤1) 重新调整数据和指标并进行模型的构建;

5) 根据确定的出租房客户预测模型,进行出租房预测结果输出,定位出租房客户。

2. 根据权利要求1所述的用于电力营销的出租房客户定位方法,其特征包括:步骤401) 中C5.0决策树算法通过最大信息增益率来选择属性进行节点拆分;第一次拆分确定的样本子集随后再次拆分,通常根据另一个字段进行拆分,这一过程重复进行直到样本子集不能再被拆分为止;最后,重新检验最低层次的拆分,那些对模型值没有显著贡献的样本子集被剔除或者修剪。信息增益率计算规则如下:

设 T 为数据集,类别集合为 $\{C_1, C_2, \dots, C_k\}$,选择一个属性 V 把 T 分为多个子集。

设 V 有互不重合的 n 个取值 $\{v_1, v_2, \dots, v_n\}$,则 T 被分为 n 个子集 T_1, T_2, \dots, T_n ,这里 T_i 中的所有实例的取值均为 v_i 。

令: $|T|$ 为数据集的 T 例子数, $|T_i|$ 为 $v=v_i$ 的例子数, $|C_j| = \text{freq}(C_j, T)$ 为 C_j 的例子数, $|C_{jv}|$ 是 $V=v_i$ 例子中具有 C_j 类别的例子数。

则有：

(1) 类别 C_j 的发生率：

$$P(C_j) = |C_j| / |T| = \text{freq}(C_j, T) / |T|$$

(2) 属性 $V=v_i$ 的发生概率：

$$P(v_i) = |T_i| / |T|$$

(3) 属性 $V=v_i$ 的例子中，具有类别 C_j 的条件概率：

$$P(C_j | v_i) = |C_{jv}| / |T_i|$$

(4) 类别的信息熵

$$H(c) = - \sum_j p(C_j) \log_2(P(C_j)) = - \sum_j \frac{\text{freq}(C_j, T)}{|T|} \times \log_2 \left\{ \frac{\text{freq}(C_j, T)}{|T|} \right\} =$$

$\text{info}(T)$

(5) 类别的条件熵

按照属性 V 把集合 T 分割，分割后的类别条件熵为：

$$H(C|V) = - \sum P(v_i) \sum P(C_j | v_i) \log_2 P(C_j | v_i) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times \text{info}(T_i) =$$

$\text{info}_v(T)$

(6) 信息增益，即互信息

$$I(C, V) = H(C) - H(C|V) = \text{info}(T) - \text{info}_v(T) = \text{gain}(V)$$

(7) 属性 V 的信息熵

$$H(V) = - \sum_i p(v_i) \log_2(P(v_i)) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|} \right) = \text{split_info}(V)$$

(8) 信息增益率

$$\text{gain}_{\text{ratio}} = \frac{I(C, V)}{H(V)} = \text{gain}(V) / \text{split_info}(V)。$$

3. 根据权利要求2所述的用于电力营销的出租房客户定位方法，其特征在于：在步骤401)中，生成的规则集包括：第一规则、第二规则、第三规则、第四规则；

第一规则：判断是否近一年总用电量 >0.61 万kw.h、近一年总用电量 ≤ 6.8 万kw.h、近一年不同收款部门数 >1 、过年期间电量占比 $>0.1\%$ 、过年期间电量占比 $\leq 0.4\%$ 、城乡类别是城镇，若均是，则认为是出租房客户；

第二规则：判断是否近一年总用电量 ≤ 0.03 万kw.h、过年期间电量占比 $\leq 0.1\%$ 、城乡类别=城镇，若均是，则认为是出租房客户；

第三规则：判断是否近一年总用电量 >6.88 万kw.h、近一年不同收款部门数 >1 、城乡类别=城镇，若均是，则认为是出租房客户；

第四规则为：判断是否近一年总用电量 >0.61 万kw.h、近一年不同收款部门数 >1 、过年期间电量占比 $>0.8\%$ 、清明假期电量占比 $\leq 0.1\%$ 、城乡类别=城镇，若均是，则认为是出租房客户。

用于电力营销的出租房客户定位方法

技术领域

[0001] 本发明涉及一种出租房客户定位方法,尤其涉及用于电力营销的出租房客户定位方法。

背景技术

[0002] 据国家有关部门统计数据显示,我国现有出租房已超亿户。如何对庞大的出租房进行管理,一直是社会关注的热点与难点,而出租房客户用电安全隐患大、电费回收难等问题是长期困扰公司营销工作的痛点。近期,部分地区政府已实施“租售同权”,出租房客户数量将持续攀升,对社会治安、企业服务成本、房东个人征信等带来更大压力。因此,精准定位出租房客户,并配合服务策略规范其用电行为,实现精准营销,刻不容缓。

发明内容

[0003] 本发明要解决的技术问题和提出的技术任务是对现有技术进行完善与改进,提供用于电力营销的出租房客户定位方法,以达到准确判别出租房客户的目的。为此,本发明采取以下技术方案。

[0004] 用于电力营销的出租房客户定位方法,包括以下步骤:

[0005] 1) 指标初选,根据出租房业务调研结果,获取建模所需目标数据群,并对获取的数据进行数据的预处理,初步选取建模指标;

[0006] 从基础信息、交费行为、用电特征三个维度提炼出8个指标进行模型构建,分别为城乡类别、年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比、设定时间内不同收款部门数及近一年支付宝交费次数;其中城乡类别为:城镇、农村;过年期间电量占比为:过年期间用电量/全年用电量*100%;清明节假期电量占比为:清明假期用电量/4月用电量*100%;端午节假期用电量占比:端午假期用电量/5月用电量*100%;

[0007] 2) 指标分析,包括连续变量分析和离散变量分析;

[0008] 201) 连续变量分析:将出租户和普通用户的年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比这5个连续变量的均值进行分析,得到出租户与普通用户对应指标的差别程度;

[0009] 202) 离散变量分析:对出租户和普通用户的近一年不同收款部门数这一指标进行分析,其中,租户各收款部门变化次数客户数占比=各收款部门变化次数客户数/出租房总数*100%,普通用户各收款部门变化次数客户数占比=各收款部门变化次数客户数/普通用户总数*100%;得到出租户与普通用户对应指标的差别程度;

[0010] 3) 指标确定

[0011] 根据指标分析结果对初选指标进行调整,选择出租户与普通用户差别程度大于设定值的对应指标为确定指标,确定最终建模指标;

[0012] 4) 出租房客户预测模型构建

[0013] 401) 根据确定的建模指标, 随机筛选样本集中70%作为训练集, 30%作为测试集构建出租房客户预测模型;

[0014] 402) 生成规则集, 利用C5.0决策树算法, 对训练集进行训练和学习生成出租房客户预测模型规则集并获得各指标对模型的影响程度及预测混淆矩阵;

[0015] 403) 根据训练集模型预测结果, 将模型应用到测试集上进行模型测试, 判断训练集和测试集的预测效果否达到了理想效果, 若是, 则确定该模型为出租房客户预测模型, 否则, 返回步骤1) 重新调整数据和指标并进行模型的构建;

[0016] 5) 根据确定的出租房客户预测模型, 进行出租房预测结果输出, 定位出租房客户。

[0017] 本项目基于电力公司营销业务系统、用电信息采集系统中的明细数据, 结合95598工单, 一体化缴费平台数据, 首先对出租房客户进行特征分析, 从基础信息、交费行为、用电特征三大维度出发, 提炼出多个影响出租房客户分析的指标作为预测指标集, 通过C5.0决策树算法构建出租房用户预测模型, 准确定位出租房用户, 实现精准营销, 规范出租房客户用电行为, 提高电费回收效率、降低安全隐患。

[0018] 作为对上述技术方案的进一步完善和补充, 本发明还包括以下附加技术特征。

[0019] 进一步的, 步骤401) 中C5.0决策树算法通过最大信息增益率来选择属性进行节点拆分; 第一次拆分确定的样本子集随后再次拆分, 通常根据另一个字段进行拆分, 这一过程重复进行直到样本子集不能再被拆分为止; 最后, 重新检验最低层次的拆分, 那些对模型值没有显著贡献的样本子集被剔除或者修剪; 信息增益率计算规则如下:

[0020] 设T为数据集, 类别集合为 $\{C_1, C_2, \dots, C_k\}$, 选择一个属性V把T分为多个子集。

[0021] 设V有互不重合的n个取值 $\{v_1, v_2, \dots, v_n\}$, 则T被分为n个子集 T_1, T_2, \dots, T_n , 这里 T_i 中的所有实例的取值均为 v_i 。

[0022] 令: $|T|$ 为数据集的T例子数, $|T_i|$ 为 $v = v_i$ 的例子数, $|C_j| = \text{freq}(C_j, T)$ 为 C_j 的例子数, $|C_{jv}|$ 是 $V = v_i$ 例子中具有 C_j 类别的例子数。

[0023] 则有:

[0024] (1) 类别 C_j 的发生率:

[0025] $P(C_j) = |C_j| / |T| = \text{freq}(C_j, T) / |T|$

[0026] (2) 属性 $V = v_i$ 的发生概率:

[0027] $P(v_i) = |T_i| / |T|$

[0028] (3) 属性 $V = v_i$ 的例子中, 具有类别 C_j 的条件概率:

[0029] $P(C_j | v_i) = |C_{jv}| / |T_i|$

[0030] (4) 类别的信息熵

[0031]
$$H(c) = - \sum_j p(C_j) \log_2(P(C_j)) = - \sum_j \frac{\text{freq}(C_j, T)}{|T|} \times \log_2 \left\{ \frac{\text{freq}(C_j, T)}{|T|} \right\} =$$

$\text{info}(T)$

[0032] (5) 类别的条件熵

[0033] 按照属性V把集合T分割, 分割后的类别条件熵为:

$$[0034] \quad H(C|V) = -\sum P(v_i) \sum P(C_j|v_i) \log_2 P(C_j|v_i) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times \text{info}(T_i) = \text{info}(T)$$

[0035] (6) 信息增益,即互信息

$$[0036] \quad I(C,V) = H(C) - H(C|V) = \text{info}(T) - \text{info}(V) = \text{gain}(V)$$

[0037] (7) 属性V的信息熵

$$[0038] \quad H(V) = -\sum_i p(v_i) \log_2(P(v_i)) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|} \right) = \text{split_info}(V)$$

[0039] (8) 信息增益率

$$[0040] \quad \text{gain_ratio} = \frac{I(C,V)}{H(V)} = \text{gain}(V) / \text{split_info}(V)。$$

[0041] 进一步的,在步骤401)中,生成的规则集包括:第一规则、第二规则、第三规则、第四规则;

[0042] 第一规则:判断是否近一年总用电量>0.61万kw.h、近一年总用电量≤6.8万kw.h、近一年不同收款部门数>1、过年期间电量占比>0.1%、过年期间电量占比≤0.4%、城乡类别是城镇,若均是,则认为是出租房客户;

[0043] 第二规则:判断是否近一年总用电量≤0.03万kw.h、过年期间电量占比≤0.1%、城乡类别=城镇,若均是,则认为是出租房客户;

[0044] 第三规则:判断是否近一年总用电量>6.88万kw.h、近一年不同收款部门数>1、城乡类别=城镇,若均是,则认为是出租房客户;

[0045] 第四规则:判断是否近一年总用电量>0.61万kw.h、近一年不同收款部门数>1、过年期间电量占比>0.8%、清明假期电量占比≤0.1%、城乡类别=城镇,若均是,则认为是出租房客户。

[0046] 有益效果:本技术方案基于电力公司营销业务系统、用电信息采集系统中的明细数据,结合95598工单,一体化缴费平台数据,首先对出租房客户进行特征分析,从基础信息、交费行为、用电特征三大维度出发,提炼出多个影响出租房客户分析的指标作为预测指标集,通过C5.0决策树算法构建出租房用户预测模型,准确定位出租房用户,实现精准营销,规范出租房客户用电行为,提高电费回收效率、降低安全隐患。

附图说明

[0047] 图1是本发明流程图。

[0048] 图2是不同收款部门个数客户占比对比图。

[0049] 图3是变量重要性结果图。

具体实施方式

[0050] 以下结合说明书附图对本发明的技术方案做进一步的详细说明。如图1所示,本发明包括以下步骤:

[0051] 1) 指标初选

[0052] 从基础信息、交费行为、用电特征三个维度提炼出8个指标进行模型构建,分别为

城乡类别、年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比、设定时间内不同收款部门数及近一年支付宝交费次数；其中城乡类别为：城镇、农村；过年期间电量占比为：过年期间用电量/全年用电量*100%；清明节假期电量占比为：清明假期用电量/4月用电量*100%；端午节假期用电量占比：端午假期用电量/5月用电量*100%；

[0053] 2) 指标分析,包括连续变量分析和离散变量分析

[0054] 201) 连续变量分析:将出租户和普通用户的年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比这5个连续变量的均值进行分析,得到出租户与普通用户对应指标的差别程度;

[0055] 202) 离散变量分析:对出租户和普通用户的近一年不同收款部门数这一指标进行分析,其中,租户各收款部门变化次数客户数占比=各收款部门变化次数客户数/出租房总数*100%,普通用户各收款部门变化次数客户数占比=各收款部门变化次数客户数/普通用户总数*100%;得到出租户与普通用户对应指标的差别程度;

[0056] 3) 指标确定

[0057] 根据指标分析结果,选择出租户与普通用户差别程度大的对应指标为确定指标;

[0058] 4) 出租房客户预测模型构建

[0059] 401) 根据确定指标,利用C5.0决策树算法,随机筛选样本集中70%作为训练集,30%作为测试集构建出租房客户预测模型,生成规则集,并获得指标对模型的影响程度;

[0060] 402) 根据样本集结果分析,判断训练集和测试集的预测正确率是否都达到了90%以上,若是,则确定该模型为出租房客户预测模型,否则,返回步骤401重新在规则集中选择新的模型。

[0061] 5) 根据确定的出租房客户预测模型,进行出租房预测结果输出,定位出租房客户。

[0062] 具体实施方式如下:

[0063] 1模型影响因素分析及变量确定

[0064] 基于实地考察、业务专家访谈以及资料查询,结合电力公司现有用户用电数据信息情况^[2],以浙江省绍兴袍江地区12000户用电客户为研究对象,包括2000户租户和10000户普通用户,分析出租户与普通用户的差异,其中,普通用户包括租户和非租户。结合实际情况,考虑到出租房客户可能在用电行为、交费方式等方面与非出租房客户会存在一定的差异性,如由于出租房客户的群租性,其用电量较非出租房客户可能会偏高;出租房客户在过年期间及节假日,用电量较平时用电量可能会出现偏少现象;出租房客户晚上用电量可能比白天用电量多,即谷电量占比可能会偏高。经过分析,最终从基础信息、交费行为、用电特征三个维度提炼出租房用户特征,模型影响变量如表1所示:

[0065] 表1模型影响变量表

| 维度 | 指标 |
|-------------|-------------|
| 基础信息 | 城乡类别 |
| [0066] 用电特征 | 年用电量 |
| | 过年期间电量占比 |
| | 清明节电量占比 |
| | 端午节电量占比 |
| 交费行为 | 4-5 月份谷电量占比 |
| | 近一年不同收款部门数 |
| | 近一年支付宝交费次数 |

[0067] 1.1 指标解释

[0068] 经过多次调整,最终从基础信息、交费行为、用电特征三个维度提炼出8个指标进行模型构建,分别为城乡类别、年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比、近一年不同收款部门数、近一年支付宝交费次数。

[0069] 城乡类别:城镇、农村;

[0070] 年用电量:2016年8月-2017年7月近一年的用电量(单位:万kw.h);

[0071] 过年期间电量占比:过年期间用电量/全年用电量*100%;

[0072] 清明节假期电量占比:清明假期用电量/4月用电量*100%;

[0073] 端午节假期用电量占比:端午假期用电量/5月用电量*100%;

[0074] 4-5月份谷电量占比:4-5月份谷电量/4-5月份用电总量*100%,考虑到天气影响因素,所以选择了4月份和5月份非空调使用季节进行谷电量分析;

[0075] 近一年不同收款部门数:2016年8月-2017年7月近一年不同收款部门数,由于租房客户的流动性,交费方式会具有多样性,所以收款单位较普通用户可能也会偏多;

[0076] 近一年支付宝交费次数:2016年8月-2017年7月近一年支付宝交费次数。

[0077] 1.2 指标分析

[0078] 连续变量分析

[0079] 对于出租户和普通用户年用电量、过年期间电量占比、清明节假期电量占比、端午节假期电量占比、4-5月份谷电量占比这5个连续变量的均值进行分析。

[0080] 表2租户与普通用户连续变量均值对比表

| | 指标 | 租户 | 普通用户 |
|--------|-----------------|--------|--------|
| [0081] | 近一年平均电量(万 Kw.h) | 5.553 | 3.3685 |
| | 过年平均用电量占比 | 5% | 9.70% |
| | 清明节平均用电量占比 | 0.90% | 11.10% |
| | 端午节平均用电量占比 | 3.80% | 10.60% |
| | 4-5 月份平均谷电量占比 | 33.90% | 36.20% |

[0082] 通过对以上指标进行分析发现,(1)出租户的近一年平均用电量较高,是普通用户的1.65倍;(2)出租房客户在过年期间、清明节、端午节假期用电量占比均远低于普通用户的用电占比;(3)租户与普通用户在4-5月份谷电用电量占相差不大,在建模的时候可能也并没有重要影响。

[0083] 离散变量分析

[0084] 对出租户和普通用户的近一年不同收款部门数这一指标进行分析,租户和普通用户不同收款部门个数客户数占比如下表:

[0085] 表3租户与普通用户不同收款部门个数客户占比对比表

[0086]

| 不同收款部门个数 | 租户各收款部门变化次数客户数占比 | 普通用户各收款部门变化次数客户数占比 |
|----------|------------------|--------------------|
| 1 | 30.47% | 67.47% |
| 2 | 31.84% | 20.7% |
| 3 | 23.69% | 8.16% |
| 4 | 9.98% | 2.86% |
| 5类及以上 | 4.02% | 0.81% |

[0087] 其中,租户各收款部门变化次数客户数占比=各收款部门变化次数客户数/出租房总数*100%,普通用户各收款部门变化次数客户数占比=各收款部门变化次数客户数/普通用户总数*100%。

[0088] 通过对近一年不同收款部门数这一指标进行分析发现,67.47%的普通客户一年内收款部门没有发生,而租户收款部门一年内没有发生变化的比例为30.47%,租户和普通用户各收款部门种类客户数占比如图2所示。

[0089] 2出租房客户预测模型构建

[0090] 2.1模型技术原理说明

[0091] 在有监督学习的二分类模型中,决策树模型可读性好,效率高,特别是在数据量不大的情况下,往往也能获得较高的准确度,且利用C5.0决策树算法、Logistic逻辑回归算法和神经网络算法分别对样本进行分类预测,通过对比发现利用Logistic逻辑回归算法和神经网络算法构建的出租房预测模型准确率和命中率均低于C5.0决策树模型的预测准确率和命中率,因此本项目采用C5.0决策树算法构建出租房客户预测模型。

[0092] C5.0决策树算法通过最大信息增益率来选择属性进行节点拆分。第一次拆分确定的样本子集随后再次拆分,通常根据另一个字段进行拆分,这一过程重复进行直到样本子集不能再被拆分为止。最后,重新检验最低层次的拆分,那些对模型值没有显著贡献的样本子集被剔除或者修剪。信息增益率计算规则如下:

[0093] 设T为数据集,类别集合为 $\{C_1, C_2, \dots, C_k\}$,选择一个属性V把T分为多个子集。

[0094] 设V有互不重合的n个取值 $\{v_1, v_2, \dots, v_n\}$,则T被分为n个子集 T_1, T_2, \dots, T_n ,这里 T_i 中的所有实例的取值均为 v_i 。

[0095] 令: $|T|$ 为数据集的T例子数, $|T_i|$ 为 $v=v_i$ 的例子数, $|C_j| = \text{freq}(C_j, T)$ 为 C_j 的例子数, $|C_{jv}|$ 是 $V=v_i$ 例子中具有 C_j 类别的例子数。

[0096] 则有:

[0097] (1) 类别 C_j 的发生率:

[0098] $P(C_j) = |C_j| / |T| = \text{freq}(C_j, T) / |T|$ 式(1)

[0099] (2) 属性 $V=v_i$ 的发生概率:

[0100] $P(v_i) = |T_i| / |T|$ 式(2)

[0101] (3) 属性 $V=v_i$ 的例子中,具有类别 C_j 的条件概率:

[0102] $P(C_j | v_i) = |C_{jv}| / |T_i|$ 式(3)

[0103] (4) 类别的信息熵

[0104] $H(c) = -\sum_j p(C_j) \log_2(P(C_j)) = -\sum_j \frac{\text{freq}(C_j, T)}{|T|} \times \log_2 \left\{ \frac{\text{freq}(C_j, T)}{|T|} \right\} =$
 [0105] **info(T)**

式(4)

[0106] (5) 类别的条件熵

[0107] 按照属性V把集合T分割,分割后的类别条件熵为:

$$H(C|V) = -\sum P(v_i) \sum P(C_j|v_i) \log_2 P(C_j|v_i) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times \text{info}(T_i) =$$

[0108] **infov(T)**

式(5)

[0109] (6) 信息增益,即互信息

[0110] $I(C, V) = H(C) - H(C|V) = \text{info}(T) - \text{infov}(T) = \text{gain}(V)$

[0111] 式(6)

[0112] (7) 属性V的信息熵

[0113] $H(V) = -\sum_i p(v_i) \log_2(P(v_i)) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|} \right) = \text{split_info}(V)$

式(7)

[0114] (8) 信息增益率

[0115] $\text{gain_ratio} = \frac{I(C, V)}{H(V)} = \text{gain}(V) / \text{split_info}(V)$ 式(8)

[0116] 最大信息增益率是属性选择及样本分区的准则,通过最大信息增益率来选择属性进行节点拆分,过程重复进行直到样本子集不能再被拆分为止。最后,重新检验最低层次的拆分,那些对模型值没有显著贡献的样本子集被剔除或者修剪。

[0117] 2.2模型建立及验证

[0118] 本次建模利用C5.0决策树算法,将绍兴袍江的12000户用户作为样本集,随机筛选样本集中70%作为训练集,30%作为测试集构建出租房客户预测模型。

[0119] 2.2.1模型规则输出结果

Rule 1 for 出租房

If 近一年总用电量>0.61 万 kw.h

And 近一年总用电量<=6.8 万 kw.h

And 近一年不同收款部门数>1

And 过年期间电量占比>0.1%

And 过年期间电量占比<=0.4%

And 城乡类别=城镇

Then 出租房客户

Rule2 for 出租房

If 近一年总用电量<=0.03 万 kw.h

And 过年期间电量占比<=0.1%

[0120]

And 城乡类别=城镇

Then 出租房客户

Rule 3 for 出租房

If 近一年总用电量>6.88 万 kw.h

And 近一年不同收款部门数>1

And 城乡类别=城镇

Then 出租房客户

Rule 4 for 出租房

If 近一年总用电量>0.61 万 kw.h

And 近一年不同收款部门数>1

And 过年期间电量占比>0.8%

And 清明假期电量占比 $\leq 0.1\%$

[0121] And 城乡类别=城镇

Then 出租房客户

[0122] 2.2.2变量重要性输出结果

[0123] 如图3所示,城乡类别、近一年不同收款部门数、过年期间电量占比、年总电量4个指标对模型影响较大,结合规则集可知,年总用电量较高、过年期间用电量占比较小且一年内收款部门变化较多的城镇用户为出租房客户的可能性较大。

[0124] 2.2.3样本集结果分析

[0125] 由样本集输出结果可知,对训练集和测试集的预测正确率都达到了90%以上,预测准确率已经比较理想。

[0126] 训练集预测结果

[0127] 表4训练集预测混淆矩阵表

| | 预测 实际 | 出租户 | 普通用户 | 总计 |
|--------|----------|-----|------|------|
| [0128] | 出租户 | 594 | 413 | 1007 |
| | 普通用户 | 374 | 6984 | 7358 |
| | 总计 | 968 | 7397 | 8365 |

[0129] 其中,行值为实际值,列值为预测值,由混淆矩阵可以得出,训练集实际为出租房客户的数量为1007户,其中正确预测为出租房客户的户数为594户,错误预测为普通用户的户数为413户,训练集具体正确预测率、命中率和覆盖率如下表:

[0130] 表5训练集预测准确率、命中率和覆盖率表

| | 准确率 | 覆盖率 | 命中率 |
|--------|--------|--------|--------|
| [0131] | 90.59% | 58.99% | 61.36% |

[0132] 测试集预测结果

[0133] 表6测试集预测混淆矩阵

| | 预测 实际 | 出租户 | 普通用户 | 总计 |
|--------|----------|-----|------|------|
| [0134] | 出租户 | 274 | 228 | 502 |
| | 普通用户 | 132 | 3001 | 3133 |
| [0135] | 总计 | 406 | 3229 | 3635 |

[0136] 其中,行值为实际值,列值为预测值,由混淆矩阵可以得出,测试集实际为出租房客户的数量为502户,其中正确预测为出租户的户数为274户,错误预测为出租户的户数为228户,测试集具体正确预测率、命中率和覆盖率如下表:

[0137] 表7测试集预测准确率、命中率和覆盖率表

| | 准确率 | 覆盖率 | 命中率 |
|--------|--------|-------|--------|
| [0138] | 90.10% | 54.6% | 67.48% |

[0139] 总结

[0140] 研究表明,出租房客户预测模型的准确率达到90%以上,预测效果较好,所选取的指标城乡类别、近一年不同收款部门数、过年期间电量占比以及年总用电量对出租房客户预测模型影响较大。下一步,计划对模型做进一步的优化工作,根据模型结果进一步完善变量指标及模型参数,保留城乡类别、近一年不同收款部门数、过年期间电量占比以及年总用电量4个变量,同时考虑增加用电量波动、过年期间是否有空窗期以及房屋类型(如回迁房、酒店式公寓、学区房等)等变量,以提高模型的准确率和命中率,在模型优化的基础上,适时扩大活动运营的范围,采用多种营销方式,提高应用成效。同时,结合出租户用户实际情况生成特征标签,利用衍生标签信息,为其他主题场景的精准营销活动做支撑。

[0141] 以上图1所示的用于电力营销的出租房客户定位方法是本发明的具体实施例,已经体现出本发明实质性特点和进步,可根据实际的使用需要,在本发明的启示下,对其进行形状、结构等方面的等同修改,均在本方案的保护范围之列。

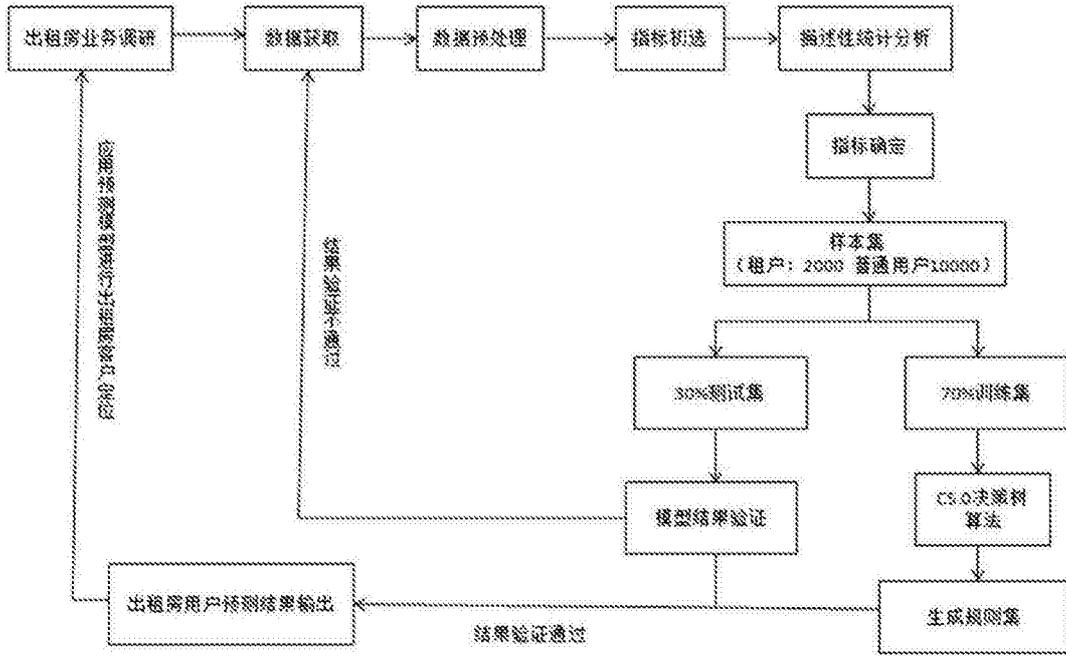


图1

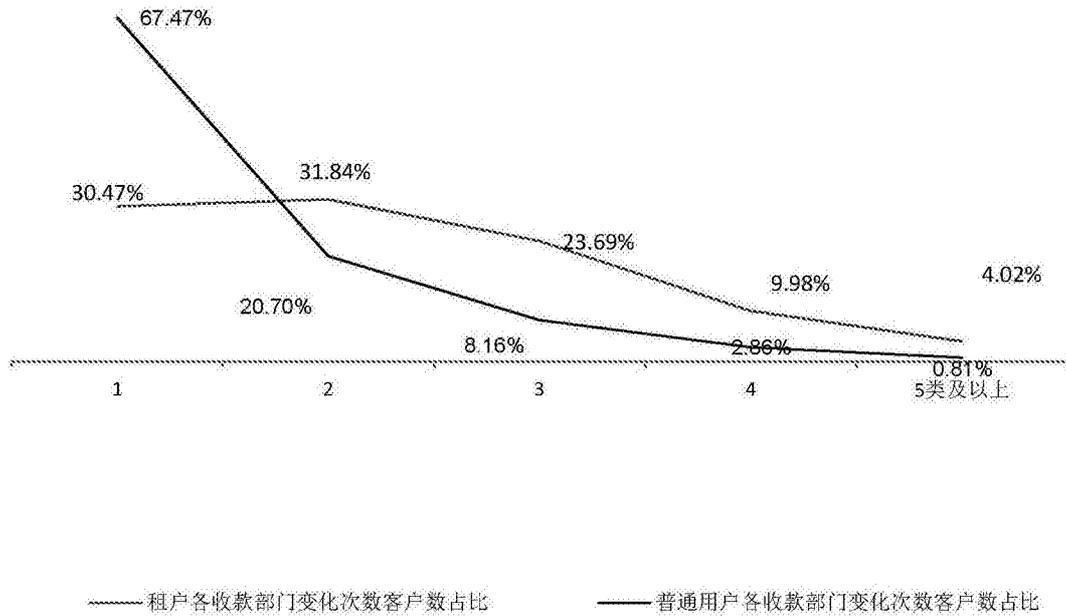


图2

Variable Importance

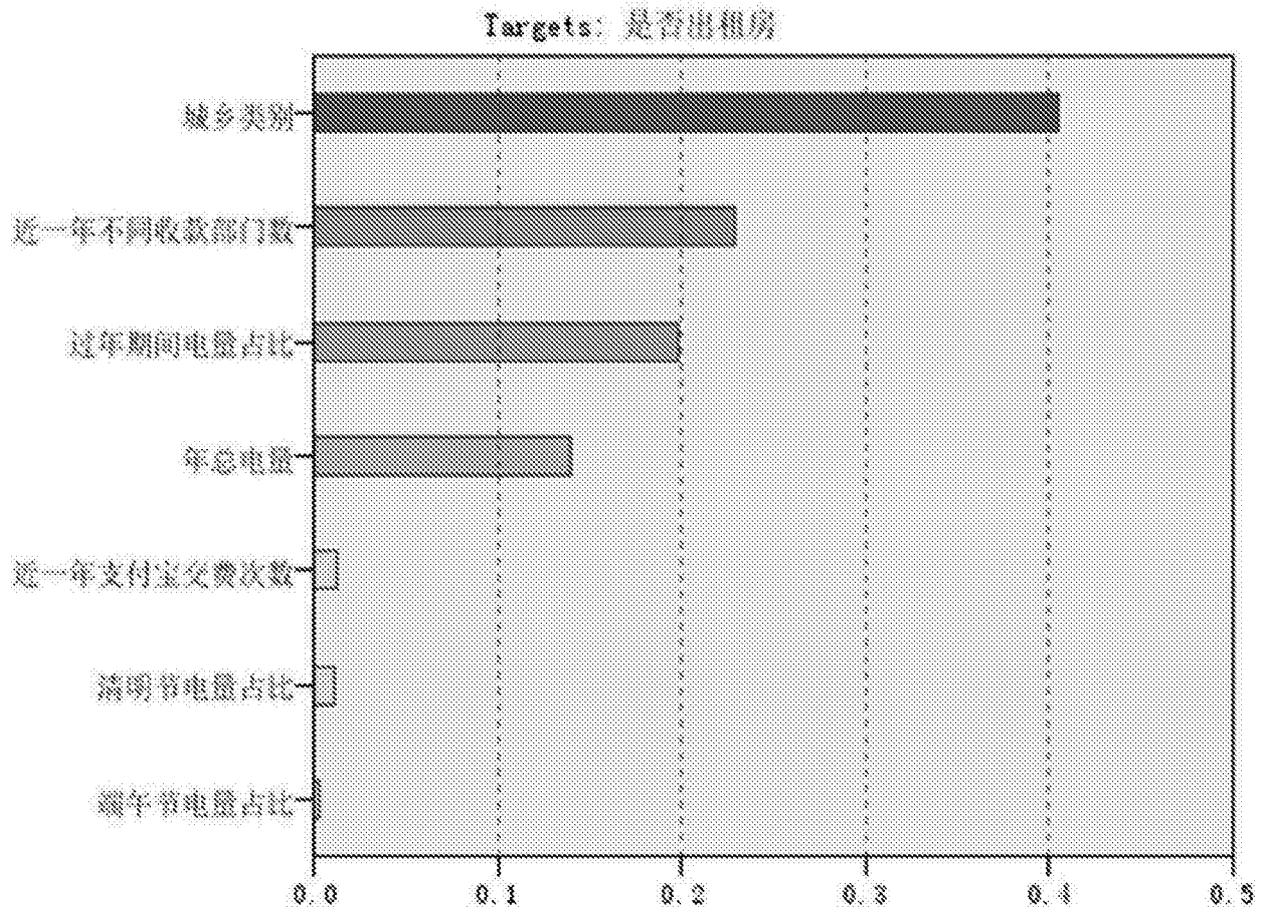


图3