



电费回收风险预测的大数据方法应用

赵雅迪¹, 吴钊², 李庆兵¹, 陈小峰¹, 王宝亭¹

(1. 北京国网信通埃森哲信息技术有限公司, 北京 100032;

2. 国网信息通信产业集团有限公司, 北京 102211)

摘要: 基于电力客户的历史数据, 采用客户的基本属性、用电行为、缴费行为、客户信用、行业前景信息等多个维度确定模型所需指标体系。通过指标的相关系数矩阵及信息值 (information value, IV) 筛选出最终进入模型的指标变量, 同时采用最优分组的方法对变量进行分组, 并进行证据权重转化 (weight of evidence, WOE)。基于处理后的数据, 运用逻辑回归算法构建用电客户电费风险预测模型, 并依据得到的模型结果量化输出变量标准评分卡表, 从而将客户划分为高风险、中风险和低风险用户, 为不同的用户采取差异化的营销措施提供依据。

关键词: 电费回收; 逻辑回归算法; 指标体系

中图分类号: O213.1

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019040

Application of big data method in forecasting the risk of tariff recovery

ZHAO Yadi¹, WU Zhao², LI Qingbing¹, CHEN Xiaofeng¹, WANG Baoting¹

1. State Grid Information & Telecommunication Accenture Information Technology Co., Ltd., Beijing 100032, China

2. State Grid Information & Communication Industry Group Co., Ltd., Beijing 102211, China

Abstract: Based on the historical data of electricity customers, the model index system was determined according to the customers' basic attributes, the electricity consumption and the payment behavior, the customers' credit, the industry prospects' information and so on. Through the correlation coefficient matrix and the information value of the index, the index variables that enter the model were selected. At the same time, the best grouping method was used to group variables and WOE (weight of evidence) transformation was carried out. Based on the processed data, the logic regression algorithm were used to construct the electricity cost risk forecasting model of the electric customers, and output variable standard score card was quantified according to the model results. Thus the customers were divided into high, middle and low risk users that could provide the basis for taking differential marketing measures to the different customers.

Key words: tariff recovery, logical regression algorithm, index system



1 引言

电费回收管理一直是供电企业的工作重点, 且是供电企业经营活动中最为重要的环节之一。长期以来, 供电企业一直采用先用电后缴费的市场规则, 因此存在电费回收周期长及催费措施落后等隐患, 使电费回收逐渐成为困扰电力企业的一大问题^[1-2]。为了解决此问题, 各电力公司纷纷提出了各种行政管理手段和技术手段, 并建立了基于客户风险的电费回收策略, 以防范电费回收风险^[3-4]。但是, 电网企业对于客户欠费风险的预判能力不足, 不能根据客户欠费风险等级提前采取具有针对性的防范措施, 是目前防范电费回收风险面临的一大难题。为了有效地提高电力企业的风险防控能力、降低企业经营风险, 高效准确的欠费风险预测尤为重要^[5-6]。

随着信息技术的发展, 国内开始使用数据挖掘技术分析海量用户的用电特征, 并试图找出用户用电特征与欠费间的联系^[7-9]。赵永良等^[10]在对用户欠费因素进行统计分析的基础上, 运用逻辑回归模型, 对影响高压用户电费回收风险的关键因素及其影响程度进行建模分析, 验证了电费回收预测关键因素在加强资金管理、提高经济效益中的作用, 但未对模型输入变量的预处理做很好的说明。黄文思等^[11]在分析引起电力客户拖欠电费原因的基础上, 深入挖掘分析客户缴费、欠费行为, 设计了欠费风险识别相关的关键影响变量, 运用决策树算法建立了客户欠费风险识别模型。吴漾等^[12]基于理论分析选择了改进的 LR-Bagging 算法来预测用电用户欠费风险, 但未使用多种算法模型进行演示分析。张禄等^[13]引入组合建模方法, 综合比较多种算法性能, 得出最优模型。该模型对客户欠费行为有良好的解释性, 可积极运用于电费回收风险管理中, 有效地指导营销业务健康运行。

实际上售电收入的形成较为复杂, 电力用户有不同类别且执行的电价标准不同^[14-16]。本文从低压居民客户基础数据、电费信息、历史缴费记录等方面对用户的历史数据进行统计分析, 深度挖掘影响客户电费回收率的关键因素, 构建低压用户影响因素指标体系, 基于逻辑回归算法构建电费风险量化分析模型, 预测客户的欠费风险等级, 为制定差异化措施及电费回收管理办法、提高电费回收率提供理论依据。

2 主要方法

2.1 最优变量分组方法

变量分组是将分类变量的某些类别合并以降低其基数, 或者将数值型变量分段将其转换为分类变量的过程。变量分组的方法是基于决策树模型的分裂找到最优的分组方案, 通过合并变量的类别使得预测力指标最大化, 即首先以使某预测力指标最大为原则找出最优的二元分割点, 然后在每个子类别中重复上一步骤, 当达到最大分组个数时停止分割。

2.2 WOE 证据权重转化方法

将分类变量转化为数值型变量, 以降低建模程序的复杂性, 同时可以将 Logistic 回归模型转变为标准评分卡格式, 以利于后续模型结果的解释及应用。对于某分类自变量的第 i 组, WOE 的计算式如下。

$$\text{WOE}(x_i) = \ln \left(\frac{p_i}{q_i} \right) = \ln \left(\frac{n_{i1} / n_{*1}}{n_{i2} / n_{*2}} \right) \quad (1)$$

其中, x_i 为某组分类自变量, p_i 和 q_i 分别是 x 中第 i 类在目标变量 Y 中的响应个数占总体响应个数概率和未响应个数占总体未响应个数概率, n_i 和 n_* 均表示个数值。信息值表示的是当前这个组中响应的客户和未响应客户的比值和所有样本中这个比值的差异。WOE 越大, 信息值越大, 这个分组里的样本响应的可能性就越大。经过 WOE

转换, 将分类变量转化为数值型变量, 同时不影响原始分布。

2.3 逻辑回归方法

逻辑回归是一种研究二分变量 Y 与一系列影响因素 x_n 之间关系的多变量分析方法, 是在线性模型基础上的进一步发展。其一般形式如下:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, P = \frac{1}{1 + e^{-p}} \quad (2)$$

其中, P 是变量 Y 发生的概率, 取值为 0~1, β 为模型建立后得到的自变量系数。逻辑回归模型的计算速度快, 结果显性化且拟合效果好, 目前在大数据、机器学习、经济学等领域得到了广泛应用。

2.4 评分卡函数方法

一个用户的评分可以表示为:

$$\text{score} = A + B\beta_0 + (B\beta_1\omega_{11})\delta_{11} + (B\beta_1\omega_{12})\delta_{12} + \cdots + (B\beta_r\omega_{p1})\delta_{r1} + (B\beta_r\omega_{p2})\delta_{r2} + \cdots \quad (3)$$

其中, $B = \text{PDO} / \ln 2$, $A = \text{BaseScore} - B \ln(\text{odds})$, β 是模型参数。 ω 是 WOE 转化值, δ 是二元变

量 (1 或者 0), 表示变量是否取某个值。 r 是模型变量个数, P 为每个变量分类个数。BaseScore 为基准分, 可以根据实际情况给出; PDO 为比率翻倍的分数变化, odds 为基础比率。用户得分控制在 0~100 分, 基于标准评分卡的算法将逻辑回归模型结果转化为评分卡的形式, 用户的最终得分即各变量对应分值的和。

3 模型构建与应用

3.1 目标客户

结合对业务的理解及现状分析, 明确电费回收风险客户的定义。不同的电费回收风险客户定义对应不同的业务表现, 也会匹配不同的业务应对策略。本文选取的目标客户为某市区违约金金额大于 0 的低压居民客户。

3.2 建设思路

建设思路如图 1 所示。首先, 对数据进行加工和处理, 对于相关性较强的指标, 保留部分即

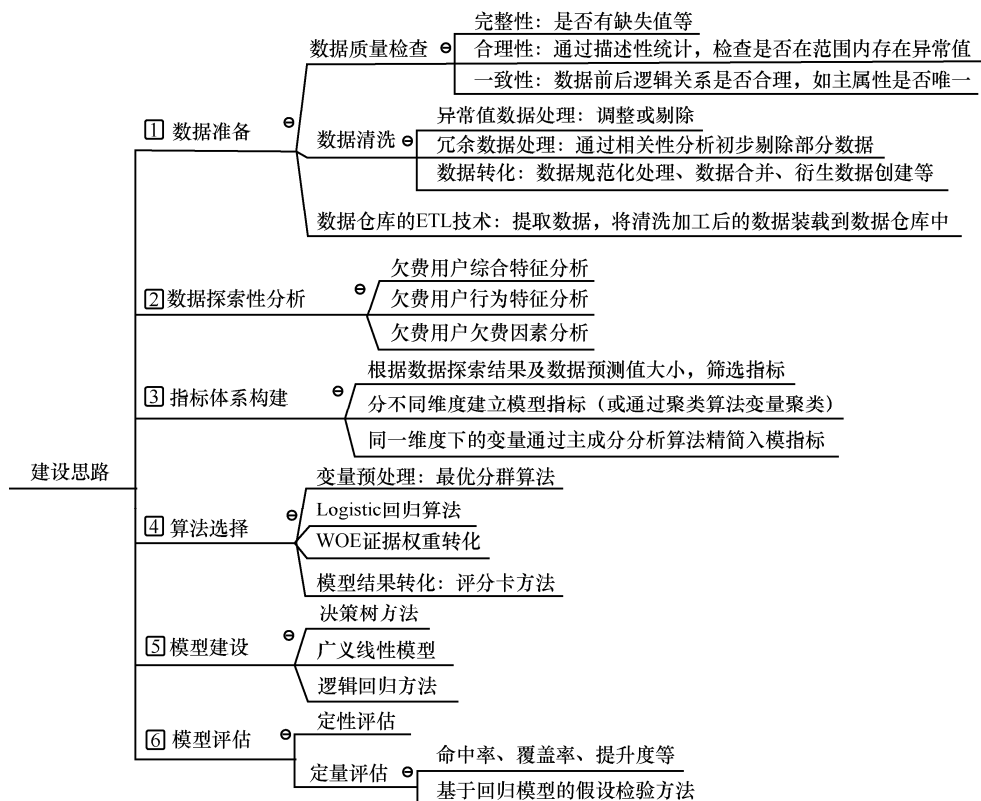


图1 电费回收风险预测模型建设思路



可，同时创建衍生变量，为建模做好数据准备。其次，对数据进行初步的探索性分析，通过对欠费用户的用电特征分析、用电趋势分析、渠道偏好分析等相关描述性统计进行分析，形成初步建设思路，为选取指标做准备。再次，选取相关指标，建立指标体系。按照逻辑回归模型的入模要求，进行聚类分析、关联分析和主成分分析，目的是在降低入模变量维度的同时获得各个指标权重，并通过变量内部分类和 WOE 权重转化等方法进行数据转化，以适应模型建模的数据要求。最后，对模型进行训练和检验，通过模型评估进行评价，重复训练获取最佳模型。

3.3 数据准备

以 2017 年 1—6 月某市区目标客户数据作为基础数据，以 2017 年 7 月和 8 月是否产生违约金作为目标变量，开展模型训练。基础数据主要包

括如下几种。

- 基本属性：用户编号、供电单位、抄表段号等。
- 用电数据：用户分类、行业分类、供电电压、合同容量、负荷程度等。
- 用电行为：用电量、电费、违约金金额，欠费次数、计费次数等。
- 缴费行为：缴费方式、缴费变更次数等。

3.4 探索性分析

3.4.1 欠费用户的综合特征分析

基于决策树算法构建欠费用户的分类模型，掌握欠费用户的综合特征。分类结果示例见表 1，其中拥有特征规则一或规则二的用户为高风险用户。

3.4.2 欠费用户的行为特征分析

基于构建的欠费用户分类模型，根据 IV 值输出重要性指标变量如图 2 所示。

表 1 欠费用户综合特征分类结果示例

规则一			规则二		
变量名	特征规则		变量名	特征规则	
现金缴费占比	>0.4 且 <0.71		现金缴费占比	>0.96	
缴费渠道数	>1		缴费次数	>2 且 ≤4	
6—10 日缴费次数	>0		5 日之前缴费次数	>0 且 ≤2	
6 个月内计费次数	>2		6—20 日缴费次数	=0	
			21—25 日缴费次数	≤2	
			历史 6 个月电费平均值	≤257.853	
			供电单位编号	*****06	
			缴费渠道	支付宝	

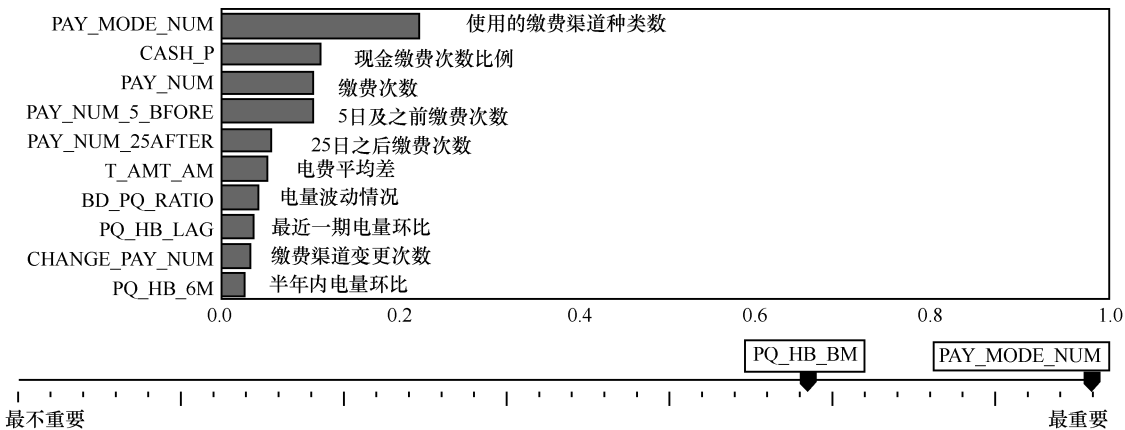


图 2 重要性指标变量

通过图2中的变量重要性列表得出以下结论。

(1) 客户使用的缴费渠道种类数与客户的欠费行为有重要关系。进一步分析该变量对欠费行为产生影响特征,同时给出欠费用户占比排名前5位的渠道变更情况。用户缴费渠道种类如图3所示,排名前五位的渠道变更情况见表2。

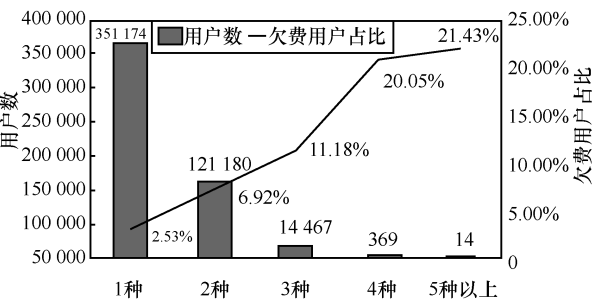


图3 欠费用户缴费渠道种类数

从图3可以发现,半年内使用缴费渠道种类数越多,欠费用户比例越高。对变更渠道用户进一步分析发现,在欠费用户占比排名前五位的渠道变更均发生在线上 and 线下缴费之间,同时这种线上、线下的渠道变更行为展示了客户在现金缴费与网上支付的转换行为,进一步说明了变量重要性中现金缴费次数比例与欠费风险存在重要关系。

(2) 在变量重要性中缴费次数相关变量与欠费行为的发生存在一定关系。对不同时间段缴费用户数及欠费用户占比进行分析,如图4所示。

从图4可以看出,约80%的用户在6—25日

间进行缴费,该类用户中欠费用户约占14%。在5日及之前缴费、25日之后缴费的用户约占19.2%,但是发生欠费概率较高。不同时间段缴费的用户欠费用户占比具有明显差异,从集中在月初、月末的特征可以看出该类用户可能存在拖延、卡点和遗忘的情况。

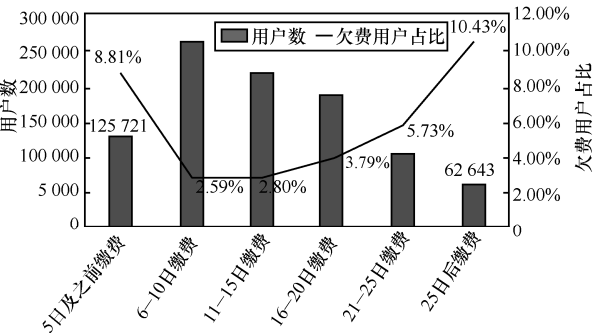


图4 欠费用户缴费时间段占比分布

3.5 指标体系构建

根据探索性分析结果及数据特征构建模型指标体系。从用户基本信息、用电行为和缴费行为3个维度构建。具体如图5所示。

3.6 模型建设

基于R软件^[17]运用逻辑回归算法实现欠费风险预测分析模型。

对于分类指标,通过IV值筛选预测力高的变量进入模型。对于连续性指标,计算变量间的相关系数矩阵,将相关系数大于0.5的指标结合IV值进行筛选。经过筛选,最终进入模型的指标有21个,具体见表3。

表2 缴费渠道变更后欠费用户占比明细

变更前渠道	变更后渠道	变更用户数	欠费用户数	欠费用户占比
微信缴费	电力机构柜台收费	130	81	62.31%
电力机构柜台收费	95598网站缴费	77	39	50.65%
电力机构柜台收费	支付宝缴费	1 363	571	41.89%
电力机构柜台收费	微信缴费	833	311	37.33%
支付宝缴费	电力机构柜台收费	451	147	32.59%

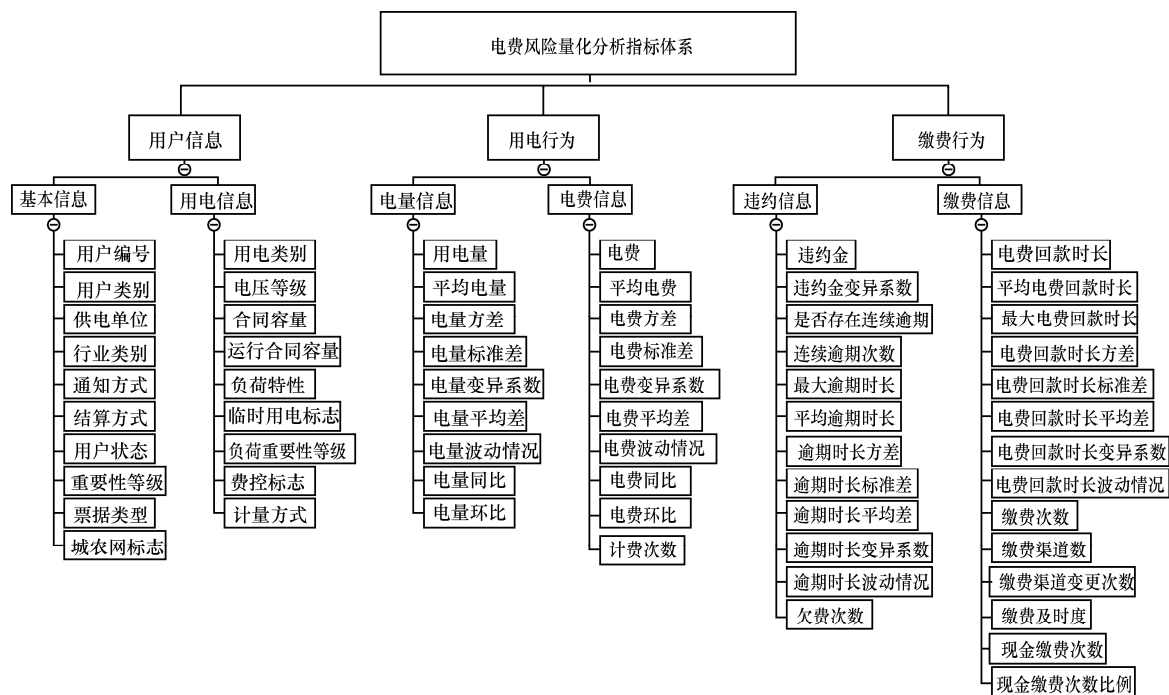


图5 电费风险量化分析指标体系

表3 入模指标

入模指标	指标名称	入模指标	指标名称
近6个月缴费回款时长平均值	PAY_TIME_AVG_6M	近6个月计费次数	RELEASE_NUM_6M
近6个月连续逾期次数	YQ_LAST_NUM_6M	缴费时长波动情况	BD_PAY_RATIO
近6个月平均电量环比	PQ_HB_6M	缴费次数	PAY_NUM
近6个月平均电量同比	PQ_TB_6M	现金缴费占比	CASH_P
近6个月交费及时度	PAY_IN_TIME_6M	电费平均值	T_AMT_AVG
25日之后缴费次数	PAY_NUM_25AFTER	电量方差	T_PQ_S
21—25日缴费次数	PAY_NUM_25_21	本期电量同比	PQ_TB_LAG
11—15日缴费次数	PAY_NUM_15_11	交费渠道变更次数	CHANGE_PAY_NUM
6—10日缴费次数	PAY_NUM_10_6	缴费渠道	PAY_MODE
6日之前缴费次数	PAY_NUM_5_BFORE	供电单位	ORG_NO
缴费回款时长方差	PAY_TIME_S		

对各指标进行最优分组处理和 WOE 证据权重转化, 构建逻辑回归模型, 并对结果进行量化输出, 给出评分卡得分。经验证, 模型通过拟合优度检验, 且各变量系数在 0.05 水平下显著。表 4 展示了部分影响力较强的指标系数, 截距项为-4.13。

3.7 模型效果

3.7.1 模型结果

基于欠费风险预测分析模型结果, 依据评分

卡函数构建欠费信用预测分析的标准评分卡, 部分重要指标得分见表 5。

3.7.2 模型效果

根据模型命中率、覆盖率和提升度 3 个维度画出模型效果评估图, 如图 6 所示, 随着高风险用户得分阈值的增大, 命中率和提升度逐渐增高, 覆盖率逐渐降低。

在图 6 中, 覆盖率和命中率交叉点的位置为

表4 部分影响力较强的指标系数

入模指标	系数
常数项	-4.132 907 9
供电单位	0.926 893 9
近6个月缴费回款时长平均值	-0.551 051 3
电量方差	-0.527 719 6
缴费回款时长方差	-0.329 336
缴费次数	-0.362 532
25日之后缴费次数	-0.232 415 4
11—15日缴费次数	-0.362 621 2
6日之前缴费次数	-0.366 416 4
电费平均值	-0.290 359 2
本期电量同比	-0.279 614
近6个月平均电量同比	-0.288 589 2
缴费渠道	0.184 463 8

表5 指标得分示例

变量名称	变量分组	得分
近6个月缴费回款时长平均值	(-1,0]	1.71
	(0,3.33]	1.47
	(3.33,6.67]	2.02
	(6.67,10.33]	2.33
	(10.33,13.67]	2.59
	(13.67,17.33]	2.93
	(17.33,22]	3.20
	(22,57]	3.84
供电单位	S区A所	2.40
	S区B所	2.46
	S区C所	1.96
	S区D所	2.50
	S区E所	2.62
电量方差	(-1,0]	2.86
	(0,5.33]	2.34
	(5.33,32.33]	2.43
	(32.33,50]	2.20
	(50,544.5]	2.30
	(544.5,722]	2.44
	(722,2 094.33]	2.34
	(2 094.33,inf]	2.45
6日之前缴费次数	(-1,0]	2.44
	(0,1]	2.39
	(1,2]	2.36
	(2,3]	2.17
	(3,inf]	1.82

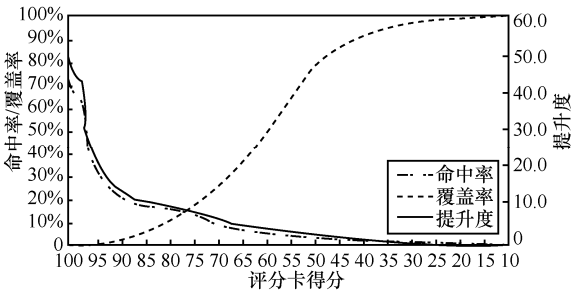


图6 模型效果

收益平衡点，即在交叉点处，考虑营销成本的情况下，可以达到收支平衡，在交叉点左侧，命中率高于覆盖率为，在考虑营销成本的情况下，投入更少的资金可以获得更好的收益；在交叉点右侧，命中率低于覆盖率为，在考虑营销成本的情况下，想要获得更好的收益（效果）需要更多的成本（措施）。

因此，在考虑营销成本的情况下，建议将评分卡得分在90分以上的客户定义为高风险用户，此时的命中率为22.8%，覆盖率为4.1%，提升度为15.4。在不考虑营销成本的情况下，建议将评分卡得分在70分以上的客户定义为高风险用户，此时的命中率为11.4%，覆盖率为26.3%，提升度为7.7。同时将2017年7月和8月数据作为试验集，通过模型验证，得到如下结果：将90分以上的客户定义为高风险用户，在考虑营销成本的情况下，此时的命中率为21.3%，覆盖率为3.6%，提升度为13.9；在不考虑营销成本的情况下，命中率为11%，覆盖率为25.4%，提升度为7.5，与训练集得到的结果基本一致。

3.8 应用场景

依据电费回收风险客户风险等级划分的结果，可在实际业务场景中进行应用，建议可针对抄表员、管理层等不同岗位设计标签及其应用场景，如图7所示。

抄表员在电费回收工作各阶段，利用客户风险等级标签筛选客户群体，指导电费回收工作。一是在抄表阶段，针对高风险、中风险客户，尽量做到优先抄表、当面抄表。并根据移动

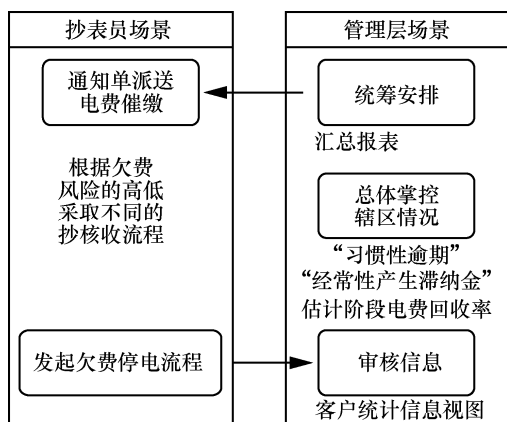


图7 模型业务应用场景

终端的提示,核实用户的联系信息,包括户号、户名、联系电话、通信地址等。二是在派送电量电费通知单工作阶段,针对高风险、中风险客户,应粘贴纸质通知单,并尽可能当面通知客户;针对低风险客户,可以采用发送电子账单的形式通知,包括短信账单、微信账单、邮件账单等。三是在派送催费通知单工作阶段,及时跟进高风险、中风险客户的电费缴纳情况,增加催缴频率和力度。并根据资源情况,采取电话的方式进行催费。对于管理层,可采取按电费风险等级排序,优先开展重点催费措施;或针对高风险用户提高催费短信发送频率,差异化催费内容,对低风险等级用户降低催费短信发送频率等。

4 结束语

利用大数据技术进行数据挖掘和建模分析可以对未来业务进行及时预测,便于有效管控风险和采取针对性措施。本文将逻辑回归模型的运用加以完善,将入模指标加以细化,具有较强的适用性和扩展性,可针对抄表员、管理人员等不同岗位职能设计标签及应用场景。如抄表员可在电费回收阶段,利用客户风险等级标签筛选客户群体,指导电费回收工作;管理层可针对高风险的居民客户,将其不良记录提供给合作个人征信机构,实现用电客户信息风险的动态评估。通过实

践应用,展现了模型对数据关联的整合能力,可广泛地应用在电力回收预警方案的设计及开发工作中,保障电费回收工作有效进行。

参考文献:

- [1] ROBU V, VINYALS M, ROGERS A, et al. Efficient buyer groups for prediction-of-use electricity tariffs[C]//AAAI Conference on Artificial Intelligence, July 27-31, 2014, Québec City, Québec, Canada. Palo Alto: AII Press, 2014: 451-457.
- [2] PEPERMANS G, WILLEMS B. Cost recovery in congested electricity networks[J]. Zeitschrift Für Energiewirtschaft, 2010, 34(3): 195-208.
- [3] 张晓峰. 电力大客户电费回收风险防范体系的构建[J]. 内蒙古科技与经济, 2013(24): 121-123.
ZHANG X F. Construction of risk prevention system for electric power customer's electricity recovery[J]. Inner Mongolia Science Technology & Economy, 2013(24): 121-123.
- [4] 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(1): 214-221.
QIU H D, TU Y, DING Q. Construction of power customer portrait and its credit evaluation and electricity fee risk control based on tag library system[J]. Telecommunications Science, 2017, 33(Z1): 214-221.
- [5] HUANG Y, YAN Y. Research of evaluating credit-risk in power enterprise based on SVM and VIKOR method[C]//IEEE International Conference on Industrial Engineering and Engineering Management, Dec 8-11, 2009, Singapore. Piscataway: IEEE Press, 2009: 1596-1599.
- [6] 刘金. 基于改进层次分析法的电力客户欠费风险评价[J]. 电子世界, 2014(19): 182-183.
LIU J. Risk evaluation of electric power customer arrears based on improved analytic hierarchy process[J]. Electronic World, 2014(19): 182-183.
- [7] WIGAN M R, CLARKE R. Big data's big unintended consequences[J]. Computer, 2013, 46(6): 46-53.
- [8] RAHMAN M N, ESMAILPOUR A, ZHAO J. Machine learning with big data an efficient electricity generation forecasting system[J]. Big Data Research, 2016(5): 9-15.
- [9] 杨华飞, 李栋华, 程明. 电力大数据关键技术及建设思路的分析和研究[J]. 电力信息与通信技术, 2015, 13(1): 7-10.
YANG H F, LI D H, CHENG M. Analysis and research of key technologies and construction ideas of power big data[J]. Electric Power Information and Communication Technology, 2015, 13(1): 7-10.
- [10] 赵永良, 秦莹, 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 Y L, QIN X, WU S Y, 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.
- [11] 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(1): 19-22.
HUANG W S, HAO H Y, LI J H, et al. Prediction of power customer arrear risk based on decision tree algorithm[J]. Electric Power Information and Communication Technology, 2016(1): 19-22.
- [12] 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(4): 70-75.
WU Y, ZHU Z. The arrears risk prediction of power residential customers based on LR-Bagging algorithm improved by feature selection[J]. Electronic Engineering and Products World, 2017(4): 70-75.
- [13] 张禄, 潘鸣宇, 田贺平, 等. 基于数据挖掘技术的电力客户欠费风险预警研究[Z]. 2017.
ZHANG L, PAN M Y, TIAN H P, et al. Risk early-warning of electricity customers'arrears based on data mining technology[Z]. 2017.
- [14] GRANELL R, AXON C J, WALLOM D C H. Predicting winning and losing businesses when changing electricity tariffs[J]. Applied Energy, 2014, 133(10): 298-307.
- [15] KAMILARIS A, KALLURI B, KONDEPUDI S, et al. A literature survey on measuring energy usage for miscellaneous electric loads in offices and commercial buildings[J]. Renewable & Sustainable Energy Reviews, 2014(34): 536-550.
- [16] GRANELL R, AXON C J, WALLOM D C H, et al. Power-use profile analysis of non-domestic consumers for electricity tariff switching[J]. Energy Efficiency, 2016, 9(3): 825-841.

- [17] FELDMAN R, SANGER J. Text mining handbook: advanced approaches in analyzing unstructured data[M]. Oxford City: Cambridge University Press, 2006.

[作者简介]



赵雅迪 (1993-), 女, 北京国网信通埃森哲信息技术有限公司大数据业务咨询顾问, 主要研究方向为业务数据分析和数据挖掘。

吴钊 (1972-), 男, 国网信息通信产业集团有限公司党委组织部(人力资源部)主任、高级工程师, 主要研究方向为电力和信息系统。

李庆兵 (1975-), 男, 北京国网信通埃森哲信息技术有限公司企业管理咨询部副主任、高级信息系统项目管理师, 主要研究方向为企业咨询与信息化咨询。

陈小峰 (1980-), 男, 北京国网信通埃森哲信息技术有限公司管理咨询业务总监, 主要研究方向为企业管理、精益生产和大数据挖掘。

王宝亭 (1981-), 男, 北京国网信通埃森哲信息技术有限公司大数据业务咨询顾问, 主要研究方向为电力营销业务扩展、计量电费和大数据挖掘。