呼叫中心虚假客户来电定位方法

潘瑞平　安宁　常利建
国家电网有限公司客户服务中心北方分中心　天津　300300

摘　要：随着信息技术的飞速发展，呼叫中心行业技术已经更成熟。但仍然有些虚拟改号从运营商网络接入发起呼叫，从而产生虚假主叫。某呼叫中心服务区域内客户量较多，投入了巨大的服务资源，但其中存在着部分非真实的客户来电，此类虚假客户极大占用了服务资源，且存在一定的服务风险。该呼叫中心对于虚假客户的判断一直以来均为事后分析判断，无法第一时间定位并有效应对。本文通过对历史数据的精准标注，通过决策树判断模型，建立了一种判断虚假客户的方法，支撑呼叫中心的业务管理工作。

关键词：呼叫中心；虚假客户；决策树；判断模型

某呼叫中心服务区域内客户量较多，为保证快速响应客户各类诉求，提高客户体验，该呼叫中心投入了巨大的服务资源。然而在实际运营管理中发现，存在部分客户来电为虚假客户，因此快速定位虚假客户来电，并给予差异化的服务策略，能有效释放服务资源。

1. 虚假客户现状

该呼叫中心 7×24 小时受理服务区域内客户各类诉求，受业务特点影响，话务量有着明显的月度波动规律，每月 10 日左右迎来话务高峰。因此呼叫中心根据话务波动规律，动态调整人员排班，保证最大限度受理客户诉求。

在大量真实客户中，夹杂着部分虚假客户，此类客户无固定的客户信息，经业务经验判断疑似为中介、信贷等虚假客户。经初步估计，虚假客户诉求占全量诉求约 10%，极大的占用了客户服务资源，造成真实客户的用电诉求得不到及时解决，同时也对客户信息安全产生一定的风险。

目前该呼叫中心对于虚假客户主要通过对其诉求、重复来电情况进行判断，属于事后研判，无法第一时间定位虚假客户从而立即采取相应策略进行应对。本文试图通过呼叫中心服务数据，研究一种事前或事中判断方法，提前定位虚假客户来电。

2. 虚假客户来电定位方法

2.1 确定典型日并对样本进行人工精标

经由业务专家研讨，结合历史数据，发现 A 省客户中虚假客户占比相对较多，故选定 A 省作为典型省，同时选取话务峰值日 2022 年 11 月 9 日作为典型日进行后续研究。

为保证后续研究数据的准确性，我们组织 10 名业务骨干区分不同业务场景，对 A 省当日的全量 8894 件诉求进行数据标注，精准标注了是否虚假客户及其相关特点，为后续定位模型的建立及风险评估做好了数据基础。

2.2 虚假客户特点

根据人工精标结果，真实客户诉求 7503 件，占比 84.36%，虚假客户诉求 1375 件，占比 15.46%，无法判断 16 件，占比 0.18%。经总结，虚假客户的主要特点如下：

2.2.1. 来电时间集中为工作时间。超九成虚假客户来电时间分布在 8:00–12:00 及 13:00–20:00，峰值出现在 9:00–12:00 及 14:00–17:00。

2.2.2. 来电号码为座机且同号段号码较多。虚假客户来电中，座机号码来电占比 58.47%，而在呼叫中心全量数据中，该占比仅为 5.22%。对座机号段进行统计，共涉及 116 个号段，平均来电 6.93 通，同号段来电 5 次及以上的共 39 个，来电量占比 82.59%。

2.2.3. 归属地集中且与来电地址不一致问题较多。号码归属地主要集中在 A 省省会城市，占比 61.19%。对归属地与来电地址进行比对，两者不一致比例为 43.91%，全量数据中，该比例为 17.60%。

2.2.4. 重复来电占比高、时间间隔短。来电 2 次及以上的号码占 46.22%，诉求量占 75.49%。超三成重复来电集中在 10 分钟内，60 分钟内重复来电占比近八成。

2.2.5. 诉求业务类型相对集中。虚假客户诉求主要集中在咨询类业务，占比 97.07%。

2.2.6. 提及“身份证”关键词较多。通过对客户来电表述的关键词进行分析，虚假客户提及关键词“身份证”占比 36.87%，高出真实客户 30.77 个百分点。

2.2.7. 部分虚假客户来电背景有明显杂音。52.48% 的虚假客户有明显背景杂音，如电话按键音、他人电话交谈声、房屋买卖对话等非用电相关信息。从静默时语音波形看，部分虚假客户静默时呈现持续小幅震荡的特点，见图 1。

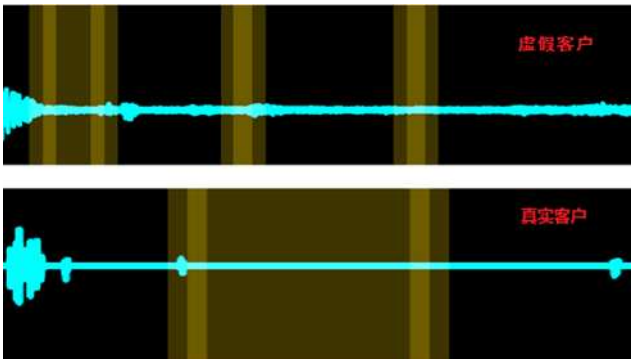


图 1 虚假客户与真实客户来电音频对比

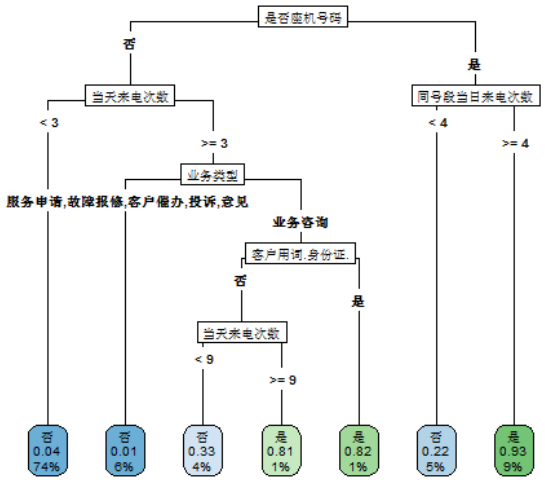
3. 模型建立

基于样本精标结果及虚假客户典型特点，采用决策树方法归纳形成虚假客户判断模型。首先，将数据样本按 7:3 的比例随机分为训练集（6230 件）与测试集（2664 件），训练集用于模型训练，测试集用于模型验证。其次，确定模型选取的 9 个变量，1 个为分类结果的目标变量，8 个为自变量，具体见表 1。[1]

表 1 模型变量选取

序号	变量	样例	备注
1	是否虚假客户	是	分类结果目标变量
2	业务类型	业务咨询	
3	业务子类型	客户信息查询	
4	来电时间（是否 8–20 点）	是	
5	当日来电次数	4	
6	是否座机号码	是	
7	是否本地号码	否	是否本市号码
8	客户用词“身份证”	是	客户在通话中表述了“身份证”相关关键词
9	同号段当日来电次数	15	如 0551626538** 号段为 05516265

在 R 语言中对训练集运用决策树模型，剪枝后结果见图 2。[2]



剪枝后

图 2 决策树模型结果

从决策树结果来看，入选虚假客户判断模型的字段包括：是否座机号码、当天来电次数、同号段当日来电次数、业务类型、客户用词“身份证”。其中判断为虚假客户的主要分枝为：

①来电为座机号码，且同号段当日来电 4 次及以上；

②来电为手机号码，业务类型为咨询，当天来电 3 次及以上，且表述“身份证”关键词；

③来电为手机号码，业务类型为咨询，没有表述“身份证”关键词，但当天来电 9 次及以上。

4. 模型验证

利用测试集验证模型结果，模型得到的虚假客户来电 276 件（占比 10.36%）、真实客户 2388 件（占比 89.64%）。模型判断准确率为，真实客户误判率为，虚假客户识别率为，经查看，未识别的虚假客户来电次数较少，且同号段来电次数也较少，相对较难识别。

表 2 模型验证结果

预测值（模型测算） 真实值（人工验证）	否	是
否	2241（判断准确）	16（真实客户误判）
是	147（虚假客户未识别）	260（判断准确）

通过验证，模型的真实客户误判率较低，且模型判断准确率、虚假客户识别率相对较高，可应用于虚假客户来电的初步预判。

5. 模型应用

将模型应用于 2024 年 10 月 A 省全部来电，得出每天虚假客户来电量情况。从趋势上看，虚假客户来电量与 A 省整体话务波动特点基本一致，10 月 7 日 –11 日工作日期间虚假客户来电量占总业务量的 15.31%（月均 8.44%）；从周特点看，每周一至周五虚假客户来电量最高（日均 776 通，平均占比 9.81%），周六次之（日均 349 通，平均占比 5.68%），周日虚假客户来电相对较少（日均 20 通，平均占比 0.38%），三者比值为 39:17:1，表明工作日是虚假客户的主要来电时间，部分中介、信贷等虚假客户为单休工作性质，周六也有较多的来电。总体来说，模型应用结果符合实际业务情况，存在可推广应用的价值。[3]

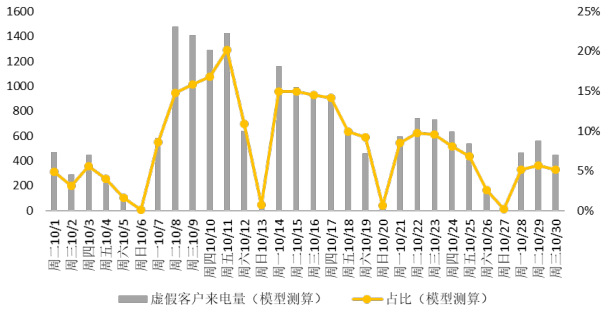


图 3 模型判断虚假来电量及占比日趋势

结束语

本文从呼叫中心实际运营过程中发现的虚假客户来电问题出发，结合虚假客户特点，建立了一种可以实时研判模型，能够相对准确的定位虚假客户，且真实客户的误判率相对较低，对呼叫中心的业务管理有一定的支撑作用。下一步，呼叫中心将不断优化虚假客户判断模型，持续提升研判准确性，同时在针对虚假客户的差异化服务策略上加以推广应用。

参考文献：

[1] 刘章华 , 魏凤歧 , 李琦 , 等 . 基于二叉决策树分类模型的草原牛行为识别 [J]. 黑龙江畜牧兽医 , 2022(000–004).

[2] 赵宁杰 , 李雪飞 . 基于 bagging 思想的决策树分类算法研究 [J]. 北京服装学院学报：自然科学版 , 2020, 40(3):6.

[3] 徐睿 , 基于大数据技术的渠道虚假客户的识别系统与应用 . 广东省 , 中国移动通信集团广东有限公司 ,2016–08–01.

作者简介：潘瑞平（1990.11）男 汉 安徽省黄山市 硕士研究生 高级工程师 研究方向：电力营销