



用户需求驱动的低满意度预测：特征构建与主题建模融合方法

刘毅¹, 张叶江¹, 胡坚¹, 唐俊², 董欢欢², 张兆强¹

(1. 中国移动通信集团云南有限公司, 云南 昆明 650228;

2. 中国移动通信集团设计院有限公司, 北京 100080)

摘要: 随着移动通信市场竞争日趋激烈及用户增长红利逐渐消退, 用户满意度已成为维系存量用户价值与实现差异化竞争的核心要素。当前各地运营商正将满意度提升策略从解决现存低满意度问题转向对潜在低满用户的预测与主动关怀。基于低满用户回访录音的处理与分析, 深入挖掘用户需求, 构建用户感知数据的衍生特征, 完成用户不满原因的主题分类, 并分别训练预测模型。实验表明该方法在准确率与召回率上均显著优于基线模型。

关键词: 低满用户; 满意度提升; 用户感知; 预测模型; 特征挖掘

中图分类号: TN929.5

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2025302

0 引言

随着我国移动通信市场进入存量竞争阶段, 用户渗透率趋于饱和, 新增用户空间日益收窄, 运营商竞争焦点已从“规模扩张”转向“存量价值深耕”。用户满意度作为衡量服务质量和用户忠诚度的关键指标, 直接影响存量用户的留存率、每用户平均收入 (ARPU) 及口碑传播效应。目前行业主要依赖月度抽样问卷进行“回溯式”满意度评估, 导致低满用户识别滞后, 且潜在不满用户难以被覆盖, 无法有效支撑主动服务需求。因此业界亟须通过预测模型精准识别潜在低满用户, 并实施针对性干预。本文提出了一种融合用户需求挖掘与特征构建的主题建模预测方法, 旨在提升潜在不满用户的预测准确率。潜在低满用户预测流程如图1所示。

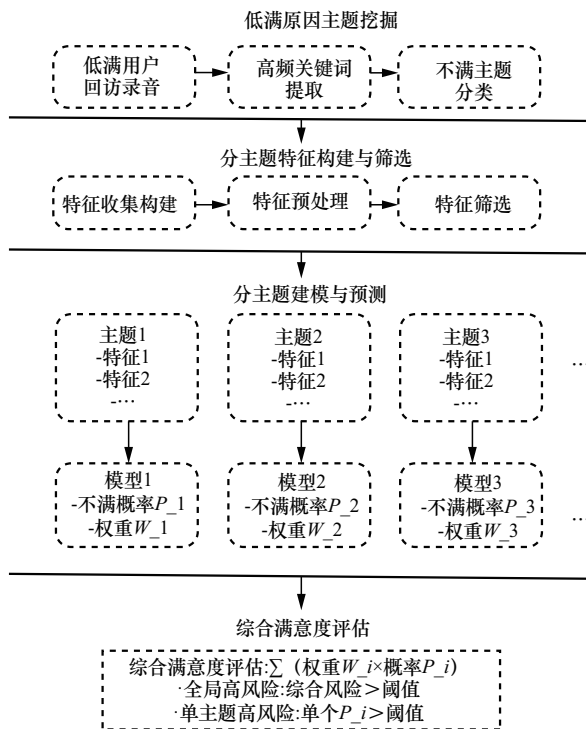


图1 潜在低满用户预测流程



1 满意度现状

1.1 满意度影响因素分析

在满意度调研中，用户对语音和上网两类业务进行评分（10分制，从1分至10分，表示不满意至满意的程度），若语音或上网评分低于9分，则将该用户定义为低满用户。对近两个月的低满用户回访分析表明，仅38.01%的用户因网络问题不满，其余用户的不满主要源于资费、宽带、服务等问题，表明用户不满归因呈现多维度复杂性特征，需深入挖掘特征指标以提升满意度分析与预测的准确性。用户不满原因分类统计如图2所示。

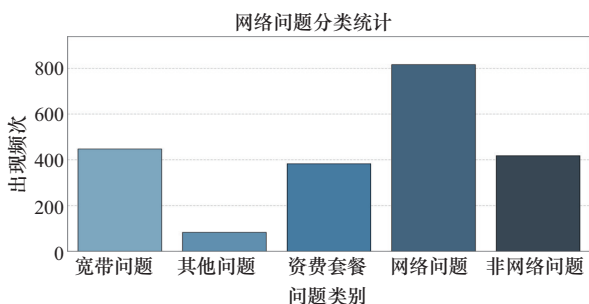


图2 用户不满原因分类统计

1.2 满意度预测研究现状

当前用户满意度预测研究在方法构建与实际应用之间仍存在显著差距。多数已有研究在特征构建与问题识别层面存在明显不足，具体表现为：张鹏等、吴迪、张硕伟等、董莹莹等、何均均等采用高度简化的建模思路，将用户多维不满场景（如语音、上网、套餐、家宽）简化为单主题或双主题的“满意/不满意”二元标签。该做法导致多主题特征混杂耦合，关键判别信号被稀释，模型难以识别用户不满的具体原因，更无法为运营改进提供可解释的依据。

其二如屈新东等、吴亚楠等、赵东明等、黄霜聆等、朱剑伟等、吴娜等虽构建了预测模型，但输入特征多依赖于基础的周期性的统计指标，

未能嵌入移动业务高动态、多场景的特性，导致模型与实际业务感知脱节，解释性和落地性较差。

针对上述局限，本文基于真实场景下的用户回访语音数据，通过文本挖掘实现细粒度不满主题抽取，并融合多域动态指标构建用户状态感知特征，从而增强模型对不满成因的辨别能力和业务可解释性。在此基础上，本研究采用分主题独立建模策略，有效避免了不同不满类型之间的特征干扰，显著提升低满意用户的识别效率与精准度。

2 低满原因主题挖掘

2.1 回访录音文件处理

以云南移动2024年7月至2025年3月脱敏回访录音为数据源，抽样显示方言占比低于5%，因此采用OpenAI的Whisper语音识别模型（基于Transformer结构），在Python环境中批量完成语音到文本转换。识别过程表示为：

$$\hat{y} = \arg \max_y p_{\theta}(y|X) \quad (1)$$

其中， $X \in \mathbf{R}^{80 \times T}$ 是由录音波形st经80维Mel滤波器得到的频谱特征矩阵（ T 为帧数）， p_{θ} 表示参数为 θ 的Transformer编码器—解码器模型，将 X 映射为上下文表示并计算文本序列 y 的条件概率， $\hat{y}$ 为模型最终输出的转录文本。

2.2 文本文件分词

采用“结巴分词”工具进行文本预处理，包括分词、去除停用词、符号与数字，仅保留词数大于3的文本。分词结果示例见表1。

表1 分词结果示例

文本编号	词语集合						
文本1	网络	断线	家	经常	解决	不满	
文本2	流量	扣费	增购	用完	超	贵	通知
文本3	反馈	网络	态度	客服	处理	差	
...	...	...	...	...	...	...	...

2.3 关键词提取

采用词频-逆文档频率算法 (term frequency-inverse document frequency, TF-IDF) 提取关键词, 设文档集 $D=\{d^1, d^2, \dots, d^n\}$, 其中 d^1 为文档 1, d^2 为文档 2, 词 t 在文档 d 中的 TF-IDF 值的计算式为:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \quad (2)$$

其中, $\text{TF}(t, d)$ 表示词语 t 在文档 d 中的词频, 分母为文档 d 中所有词语出现次数的总和, $\text{IDF}(t, D)$ 表示词语 t 的逆文档频率, 用于衡量词语 t 在语料库 D 中的普遍程度, 其中, n 是语料库 D 中文档的总数, n' 是包含词语 t 的文档数量, 根据 TF-IDF 值排序选取关键词, 用户不满关键词 TF-IDF 值示例见表 2。

表2 用户不满关键词 TF-IDF 值示例

词语	词频 (TF)	逆文档频率 (IDF)	TF-IDF 值
网络	5 234	1 087	0.875
信号	5 432	1 100	0.872
上网	5 678	3 234	0.821
卡顿	5 432	3 123	0.815
慢	4 567	2 987	0.803
通话	4 321	2 056	0.778
流量	2321	901	0.685
...	...	...	...

2.4 主题-关键词输出

使用 Python 的 sklearn 库根据 TF-IDF 值构建一个文档-词语矩阵, 其中每一行代表一个文档, 每一列代表一个词语, 矩阵中的元素是该词语在该文档中的 TF-IDF 值, 用于后续主题建模。主题建模聚类原理示意图如图 3 所示。基于 TF-IDF 的文档-词语矩阵见表 3。

表3 基于 TF-IDF 的文档-词语矩阵

文本编号	网络	断线	家	经常	解决	不满	流量	扣费	增购	用完	超	贵	通知	反馈	态度	客服	处理	差
文本 1	0.88	0.87	0.86	0.86	0.85	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
文本 2	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.68	0.68	0.67	0.67	0.66	0.66	0.00	0.00	0.00	0.00	0.00
文本 3	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.85	0.84	0.84	0.83

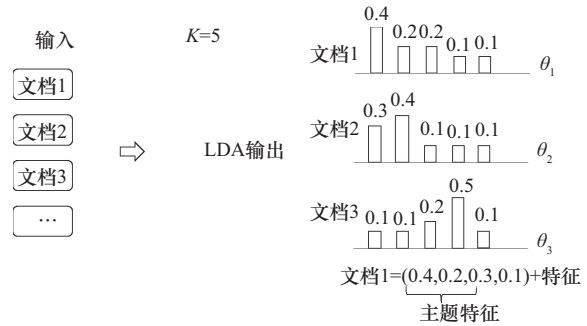


图3 主题建模聚类原理示意图

在文本分析中, 使用隐含狄利克雷分布 (latent Dirichlet allocation, LDA) 通过构建“文档-主题-词语”的 3 层概率模型, 提取潜在语义特征。每篇文档 d 视为 K 个主题的混合体, 主题比例由多项分布 $\theta_d \sim \text{Dir}(\alpha)$ 控制, 每个主题 k 对应词分布 $\phi_k \sim \text{Dir}(\beta)$, 对于文档 d 中的每个词位置 i 主题采用为 $z_{d,i} \sim \text{Mult}(\theta_d)$, 词语生成成为 $w_{d,i} \sim \text{Mult}(\phi_{z_{d,i}})$, 词生成过程为:

$$p(w, z, \theta, \phi | \alpha, \beta) =$$

$$\prod_{k=1}^K p(\phi_k | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(z_{d,i} | \theta_d) p(w_{d,i} | \phi_{z_{d,i}}) \quad (3)$$

经过参数寻优确定主题数为 5 时, 模型具有最优泛化能力, 将用户不满问题内容相似的归为一类主题并输出代表性词汇, 用户不满主题-关键词见表 4。

各主题词语都表现出一类用户不满的细分领域, 主题 1 是关于用户资费套餐消费的问题, 主题 2 是关于宽带机顶盒等业务使用问题, 主题 3 是关于用户通话异常问题, 主题 4 是关于上网业务卡顿问题, 主题 5 是活动营销相关问题, 各主题关键词分布具有显著的业务区分度, 能有效捕获用户反馈



表4 用户不满主题—关键词

主题	词语
主题 1	流量, 业务, 本月, 费用, 余额, 解释, 贵, 缴费, 流量包, 扣费
主题 2	宽带, 办理, 机顶盒, 连接, 错误, 电视, 接入, 卡顿, 没网, 重启
主题 3	网络, 信号, 打不通, 电话, 断断续续, 异常, 提示, 电梯, 覆盖, 断线
主题 4	上网, 视频, 慢, 卡顿, 经常, 室内, 商场, 不好, 扫码, 一直
主题 5	积分, 兑换, 参加, 活动, 礼品, 营业厅, 无效, 办理, 变更, 收费

的核心诉求, 为后续预测提供可解释的特征依据。

2.5 满意度特征收集

通过“主题-关键词”输出的用户关注 TOP 词汇及主题, 从现有运营 (operational, O) 域、业务 (business, B) 域、市场 (marketing, M) 域指标中找到相近指标以及特征挖掘得到的场景级感知指标进行汇聚得到全量满意度特征, 由于用户关注点不同, 针对每一类主题单独进行特征筛选及模型训练。满意度特征变量分类见表 5。

3 低满用户不满特征挖掘

对移动用户的不满原因进行主题分类和关键词两方面的文本挖掘和分析, 找到用户感知主题下特征挖掘的方向和维度, 如用户上网关键词中用户有“上网、视频、游戏、扫码”不满的业务类型、有“室内、商场”不满的使用场景、有“经常、一直”的使用感受, 需要针对这些用户主要关注点进行特征衍生构建以贴合用户使用感受。

基于扩展详单 (extended detail record, XDR) 单据中小区占用、流量使用、基础业务指标等数据

基础上进行应用、业务、场景、时间维度的衍生计算得到相关特征, 以实现对用户感知的全面、精准评估。满意度用户感知评估算法原理如图 4 所示。

3.1 用户常驻场景识别

商贸娱乐及交通出行场景可使用小区对应场景表进行关联匹配, 用户居住地、工作地、高流量使用场景各有不同, 通过 XDR 单据用户各小区下驻留及流量使用信息完成识别

高频小区计算方式见式 (4), 其中用户驻留过的小区 1 计为 c^1 , 近 7 天所有的驻留小区集合为 $C=\{c^1, c^2, \dots, c^m\}$, $V(u, c, t)$ 表示用户 u 在时间段 t 于小区 c 的总流量使用量 (MB), $D(u, c, t)$ 表示用户 u 在时间段 t 于小区 c 的业务持续时长 (s), 设定流量阈值 T_{vol} 和持续时长阈值 T_{dur} :

$$Chigh(u, t) = \{c \in C \mid V(u, c, t) \geq T_{vol} \wedge D(u, c, t) \geq T_{dur}\} \quad (4)$$

居住地/工作地识别算法见式 (5), 其中引入了流量权重 α 和时长权重 β , 对流量+持续时长加权, 计算综合得分, t 表示时间分段, 包含白天 (9:00-17:00) 和晚上 (22:00-次日 7:00):

表5 满意度特征变量分类

基础信息	年龄、性别、星级、在网时长、用户类型、终端品牌、终端使用时长、外呼接受度、主套餐多久未更换、一机双卡标识...
上网体验	无线侧超时次数、寻呼失败比例、初始注册失败比例、服务失败比例、寻呼失败比例、HTTP_TCP 建立时延过大比例、VIDEO_下行平均往返路程时间 (RTT) 过大比例...
语音感知	语音通话次数、语音始呼接通率、语音终呼接通率、语音切换成功率、上行高丢包通话次数、下行高丢包通话次数、单通次数、掉话率、高接通时延通话次数...
家庭宽带	是否电视盒子用户、是否移动宽带用户、近 3 个月均宽带流量、网关性能、业务承载、Wi-Fi 性能、智能组网设备性能、接入网性能、宽带业务质量、电视业务质量、服务质量...
营销活动	营销活动参与数、10086 系统呼入时间、掌厅登录次数、是否交互式语音应答 (IVR) 自助服务、是否 IVR 查询、是否 IVR 办理、使用人工服务次数、接通排队时长、人工服务通话时长、挂机满意度评价结果...
流量套餐	近 3 个月月均每用户每月平均收入 (APRU)、近 3 个月月均每户每月上网流量 (DOU)、近 3 个月月均每户每月通话时间 (MOU)、是否不限量套餐、是否超套、超套金额、流量饱和度、有无主资费变动、主资费价格环比、订购的增值业务数...

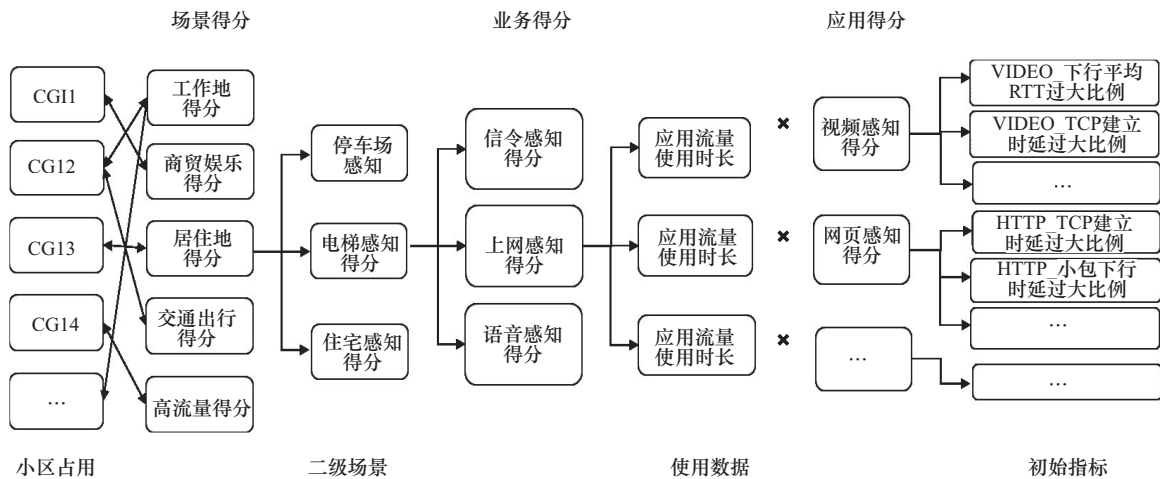


图4 满意度用户感知评估算法原理

$$S(u, c, t) = \alpha \cdot \frac{V(u, c, t)}{\max_c V(u, c, t)} + \beta \cdot \frac{D(u, c, t)}{\max_c D(u, c, t)} \quad (5)$$

统计所有用户的居住地/工作地分布小区的数据计算式为：

$$\text{居住地分布: Hist}_{\text{home}}(c) = \sum_{u \in U} I(c_{\text{home}}(u) = c) \quad (6)$$

$$\text{工作地分布: Hist}_{\text{work}}(c) = \sum_{u \in U} I(c_{\text{work}}(u) = c) \quad (7)$$

3.2 用户感知得分计算

为了计算用户的中长期业务感知得分，需要结合业务使用权重、业务指标得分、时间衰减权重，以此构建一个分层加权模型。用户感知得分分层加权模型见表6。

表6 用户感知得分分层加权模型

层	标准化范围	原因	时间权重
业务权重 w_b	用户自身当周数据	反映用户个体当周业务偏好	周权重影响最终聚合
指标得分 $S(u, b)$	全量用户当周数据	确保网络质量评价标准一致	当周数据需独立标准化
时间衰减 ω_k	月粒度指标 (近4周汇聚计算)	近期数据对当前体验影响更大	按预设衰减系数加权
月聚合得分 $S_{\text{month}}(u)$	加权聚合各周得分	平衡长期稳定性与短期波动	各周得分乘时间权重后求和

3.2.1 业务权重计算

业务权重 $w_b(u)$ 反映用户自身业务偏好，基于用户自身业务量计算，计算式见式(8)其中分

母为同一用户所有业务的加权和， α, β, γ 是全局超参数，体现流量、时长、请求次数的相对重要性 ($\alpha=0.5, \beta=0.3, \gamma=0.2$):

$$w_b(u, k) = \frac{\alpha V(u, b) + \beta T(u, b) + \gamma N(u, b)}{\sum_{b' \in B} (\alpha V(u, b') + \beta T(u, b') + \gamma N(u, b'))} \quad (8)$$

3.2.2 业务指标得分

网络质量指标基于所有用户数据进行全局标准化处理，因为所有用户对相同网络指标的感知差异是一致的，其中指标标准化处理计算式见式(9)，其中分子表示当前值 D 指标 (u, b) 与所有数据中最小值的差距，分母表示数据的整体范围 (最大值减最小值)，用于计算相对位置，分数部分结果是当前值在数据范围内的相对位置 (取值范围为 $[0, 1]$ ，最小值对应0，最大值对应1)，整体公式“1-分数”，对归一化结果取反，即反向归一化，原始数据的最小值对应1，最大值对应0，中间值线性映射到 $(0, 1)$ 之间。

$$\widehat{D}_{\text{指标}}(u, b) = 1 - \frac{D_{\text{指标}}(u, b) - \min(D_{\text{指标}})}{\max(D_{\text{指标}}) - \min(D_{\text{指标}})} \quad (9)$$

周粒度感知得分见式(10)，计算用户 u 的业务权重 $w_b(u, k)$ 再根据全局范围标准化网络指标，计算 $S(u, b, k)$ ，最后加权求和得到 $S_{\text{total}}(u, k)$ ：



$$S_{\text{total}}(u, k) = \sum_b w_b(u, k) \cdot S(u, b, k) \quad (10)$$

设定每周的权重系数 ω_k ，反映数据新鲜度最近一周权重最高，且每周轮询计算捕捉用户感知变化；最近第1周（ $k=0$ ）： $\omega_0=0.5$ ，第2周（ $k=1$ ）： $\omega_1=0.25$ ，第3周（ $k=2$ ）： $\omega_2=0.15$ ，第4周（ $k=3$ ）： $\omega_3=0.1$ ，将最近K周的得分按时间权重加权求和，此方法既能体现用户长期行为感知，又能体现近期网络变化，见式（11）：

$$S_{\text{month}}(u) = \sum_{k=0}^{K-1} \omega_k \cdot S_{\text{total}}(u, k) \quad (11)$$

通过周粒度感知数据进行聚合得到最终用户感知得分详表，对用户中长期使用过程中在不同细分场景的使用情况进行数据化评估，作为特征数据纳入模型训练。用户感知得分输出截取前7个样本见表7。

表7 用户感知得分输出截取前7个样本

用户	居住地— 上网得分	居住地— 语音得分	居住地— 信令得分	居住地— 电梯—得分	...	高流量— 上网得分	高流量— 语音得分	高流量— 信令得分	...
1	100.00	88.21	100.00	100.00	...	100.00	100.00	100.00	...
2	85.14	71.93	99.60	92.84	...	88.29	100.00	100.00	...
3	83.11	95.44	93.08	93.88	...	74.34	95.00	100.00	...
4	88.36	91.14	81.51	93.88	...	87.93	96.14	80.00	...
5	81.71	80.32	95.00	72.61	...	82.11	97.05	100.00	...
6	99.82	99.73	100.00	97.82	...	95.71	78.67	90.00	...
7	99.82	97.07	92.20	83.55	...	96.17	96.40	99.23	...
...	...	...	...	...	...	...	...	...	...

表8 预处理后的上网数据集前10条数据

用户	上网 满意度	年龄	性别	用户 类型	终端 品牌	终端使 用时长	星级	近3个月 月均APRU	近3个月 月均DOU	外呼 接受度	易受 访概率	是否不 限量套餐	居住地—停车 场感知得分	...
用户1	10	2	1	1	9.7	0.191	0	-0.80	-0.25	0.25	0.35	1	99.95	...
用户2	10	2	1	3	9.7	0.106	3	0.55	1.21	0.11	0.10	0	99.94	...
用户3	10	2	1	3	9.3	0.553	6	2.09	-0.07	0.00	0.23	0	100.00	...
用户4	10	2	2	3	9.5	0.298	1	-0.59	-0.68	0.33	0.46	0	100.00	...
用户5	10	1	1	4	9.7	0.149	1	-0.23	-0.71	0.33	0.41	1	100.00	...
用户6	10	4	1	3	9.8	0.872	3	-0.37	-0.73	0.21	0.21	0	98.80	...
用户7	9	3	2	1	9.3	0.043	4	0.20	0.15	0.10	0.10	0	98.96	...
用户8	9	4	2	3	9.7	0.872	3	-0.63	-0.71	0.21	0.21	1	100.00	...
用户9	9	1	2	3	9.3	0.944	3	0.12	1.42	0.84	0.81	0	99.25	...
用户10	9	4	1	4	9.2	0	3	-0.35	-0.65	0.52	0.48	0	94.25	...

3.2.3 用户特征预处理

基于上述满意度特征，进行基本的数据清洗、数据重置、缺失值填补等处理，对于缺失值超过60%的特征，即用户基础数据中的“一机双卡标识”特征，结合实际意义的考察，删除这个特征，对于存在缺失值的变量，“是否不限量套餐”，根据数据本身特征值分布，用“N”进行填充，对于其余存在缺失值的变量，如“居住地—停车场感知得分”“交通出行—地铁感知得分”等感知类特征，根据其特征含义表明用户未在占用该场景进行业务使用，故用100分进行填充；

对于特征“外呼接受度”中出现的异常值“#DIV/0!”，根据其含义，用0替换，对于字符型的特征变量，为使其能够代入模型完成预测，需进行特征编码。变量值为“Y/N”的特征变量，将其通过{“Y”：1，“N”：0}的对应方式转化为

数值型变量，如“是否不限量套餐”，对特征“终端品牌”进行目标编码计算；

对其余字符型特征，利用 sklearn 库中的 LabelEncoder 进行标签编码，将 n 种不同的取值转化为 0 至 $n-1$ 中的数字，对“年龄”“主套餐多久未更换”“近3个月月均DOU”等特征进行截断异常值：将超过 99% 分位数（假设年龄为 100）的值设为 100，根据各特征值含义进行等频分箱、业务导向分箱（binning），例如，将“在网时长”根据实际意义划分为（0~6个月）新用户、（6~12个月）稳定用户、（12~24个月）老用户等，并编码为有序类别。预处理后的上网数据集前 10 条数据如图所示。

3.2.4 特征筛选

在模型训练前通过构建皮尔逊相关性矩阵，

评估各特征之间的线性相关性，进行特征筛选，通过识别和剔除冗余或不相关的特征，以确保模型在保持解释能力的同时，减少过拟合的风险，并提升计算效率。

首先删去仅作为用户标识序号、对目标变量无解释性的“用户 ID”特征。对数据集的部分特征分别求出其皮尔逊相关系数矩阵，根据图中各变量的相关性可知语音业务特征“易受访概率”与“外呼接受度”完全线性相关，故剔除其中之一“外呼接受度”，“居住地-电梯感知”“5G套餐判断”“是否不限量套餐”与用户上网满意度评分相关性较低故剔除 3 类特征，上网数据集部分特征皮尔逊相关性矩阵如图 5 所示。

主题数据集特征选择结果见表 9。

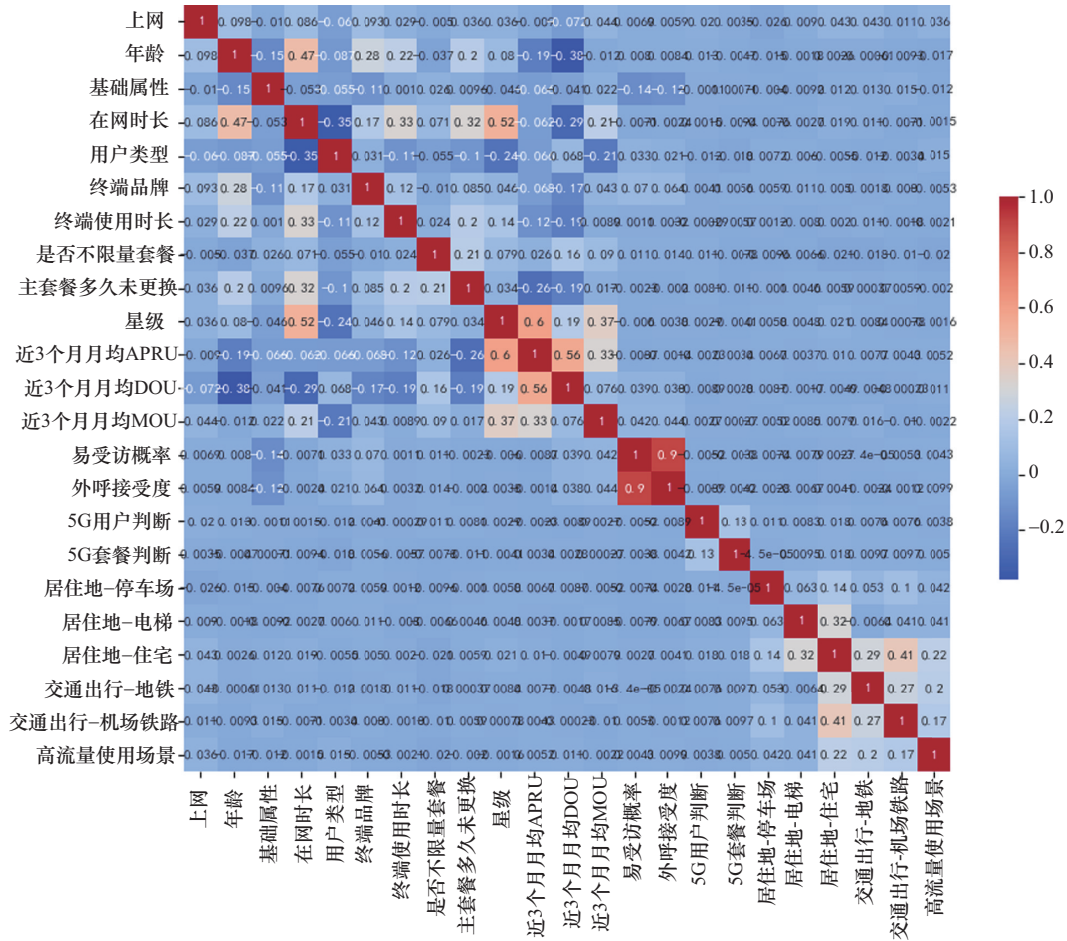


图5 上网数据集部分特征皮尔逊相关性矩阵



表9 主题数据集特征选择结果

主题维度	指标名	主题维度	指标名
基本信息	用户 ID	语音	居住地-语音得分
	年龄		工作地-语音得分
	性别		交通出行-语音得分
	在网时长		商贸娱乐-语音得分
	用户类型		高通话小区-语音得分
	终端品牌		常驻小区-语音得分
	终端使用时长	家庭宽带	是否家宽客户
	是否不限量套餐		是否电视盒子用户
	主套餐多久未更换		Wi-Fi性能
	星级		智能组网设备性能
	近3个月月均APRU		接入网性能
	近3个月月均DOU		网关性能
	近3个月月均MOU		宽带业务质量
	是否投诉客户		电视业务质量
	易受访概率		服务质量
	是否敏感客户	营销活动	10086系统呼入时间
	是否免打扰客户		是否请求人工服务
	是否离网预警客户		人工服务通话时长
	是否高风险携出客户		是否进行挂机满意度评价
	是否资费敏感客户		挂机满意度评价结果
	一机双卡标识		营销活动参与数
	5G用户判断		掌厅登录次数
	2G网络使用时长		是否IVR查询
	4G网络使用时长		是否IVR办理
	5G网络使用时长		是否业务办理满意度评价
	总上网时长		业务办理满意度评价
	4G通话时长	资费使用	是否超套
	2G网通话时长		超套金额
	5G通话时长		是否流量超套
上网	居住地—上网得分		当月超流量费用
	居住地—信令得分		上月超流量费用
	工作地—上网得分		过去3个月流量超套的次数
	工作地—信令得分		过去3个月流量超套月份的月均套外流量金额
	交通出行—上网得分		过去3个月语音超套的次数
	交通出行—信令得分		过去3个月语音超套月份的月均套外语音金额
	商贸娱乐—上网得分		流量饱和度
	商贸娱乐—信令得分		有无主资费变动
	高流量小区—上网得分		主资费价格环比
	高流量小区—信令得分		增值业务退订数量
	常驻小区上网得分		订购的增值业务数
	常驻小区信令得分		近3个月内是否订购提速包
	停车场上网得分		是否限速
	感知异常天数		欠费停机次数
	感知异常持续时长		近3个月内降速后是否提速
	...		当月在省内使用的流量比例

4 预测模型的建立与分析

本文聚焦于用户满意度的二分类预测问题，旨在通过机器学习方法对用户满意度的“是”与“否”进行准确预测，采用 ROC 曲线下面积（area under the curve, AUC）值作为模型预测性能的主要评估指标，AUC 是 ROC 曲线下方的面积，用于衡量分类器性能，取值范围在 0 到 1 之间：AUC=1 表示模型是完美的分类器，能够将所有正例和负例完全正确地分类，AUC=0.5 表示模型性能等同于随机猜测，没有分类能力， $0.5 < \text{AUC} < 1$ 表示模型具有一定的分类能力，AUC 值越接近 1，模型的分类性能越好。ROC 曲线的 AUC 值如图 6 所示。

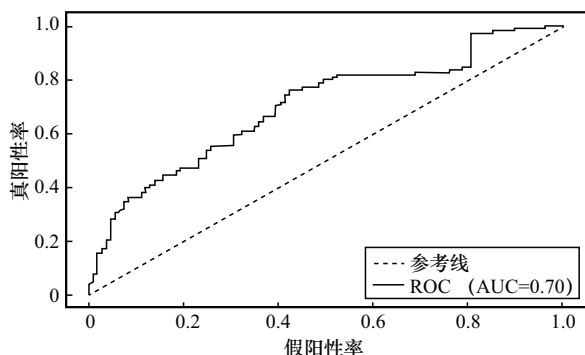


图6 ROC曲线的AUC值

4.1 高效精准的问题定位与分析

全网低满用户（满意度评分<9分）仅占全量用户的35%，其余65%为高满用户，导致数据集中“低满”样本（负样本）数量远少于“高满”样本（正样本），形成典型的类别不平衡现象；在这种正

负样本比例悬殊的情况下，模型容易偏向多数类而将少数的负样本误判为正样本，因此在建立预测模型前必须对不平衡数据集进行处理。

本文将过采样（synthetic minority oversampling technique, SMOTE）方法与欠采样（edited nearest neighbour, ENN）方法结合，首先在少数类样本点的最近邻连线上随机选择新的样本点扩充少数类样本，然后对新的样本集中的大多数近邻样本点分类与自身分类不同的点进行剔除，得到最终的新样本。用 imblearn 库中的 SMOTEENN 方法对两个数据集进行处理，得到处理后的各主题数据集分布如图 7 所示。

4.2 模型预测结果分析

训练样本平衡后，对各主题业务数据集建立多个机器学习模型，分别求其 AUC 均值，以寻找预测效果最好的模型，各模型训练 AUC 值见表 10。从表 10 中数据可以看出，集成学习模型的预测效果普遍优于传统学习模型。在对上网、语音数据建模中，XGBoost 算法的预测表现最佳，宽带、营销业务中，LightGBM 算法的优势突出，家庭宽带是 Random Forest 算法最优。

表 10 各模型训练 AUC 值

模型名称	上网 体验	语音 感知	家庭 宽带	营销 活动	流量 套餐
随机森林	0.939 1	0.910 2	0.885 3	0.837 1	0.891 5
GBT	0.928 5	0.891 2	0.834 5	0.812 3	0.845 6
XGBoost	0.954 6	0.937 4	0.858 6	0.852 3	0.866 1
LightGBM	0.945 8	0.923 1	0.876 4	0.864 2	0.882 7
Bagging	0.887 5	0.864 2	0.812 3	0.795 6	0.823 4
SVM	0.856 6	0.852 7	0.750 9	0.720 7	0.747 3

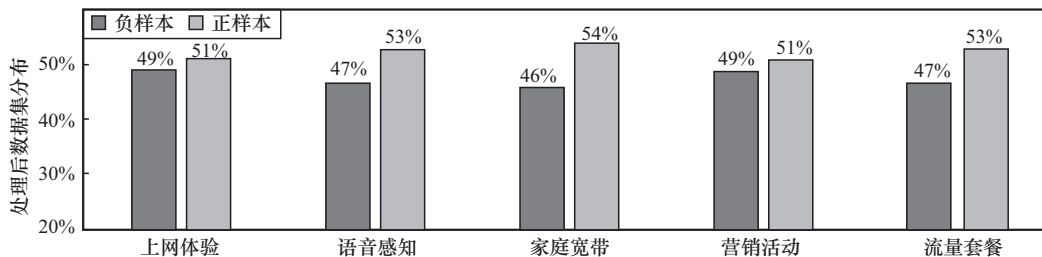


图7 处理后的各主题数据集分布



4.3 消融实验验证特征价值

通过消融实验系统验证特征工程对模型性能的贡献。实验以XGBoost模型为基础,按照“基础感知特征→感知得分特征→分场景感知得分特征”的渐进式特征构建逻辑,在数据集上进行验证:当仅使用基础O域特征时,AUC值为0.836 2,暴露出传统指标的局限性与用户满意度弱相关;添加分层加权感知模型后AUC提升至0.893 0(+6.8%),凸显出时间衰减机制与业务权重设计的价值;继续引入场景化特征工程后,AUC跃升至0.954 6(较基础特征提升14.1%),证实空间维度对用户感知建模的关键作用,实验表明场景化特征解决了传统指标的“空间盲区”问题,感知得分建模有效捕捉了业务使用时变特性,显著提升了预测性能。

本文研究围绕5G时长驻留比提升展开,构建“六层协同”优化体系。通过基础网络提质,挖掘功率、整治下倾,扩大5G覆盖;多频协同组网,优化互操作门限与测量参数,引导终端驻留5G;整治低效及高倒流小区,精准定位问题区域并整改;部署特性功能与版本升级,提升用户感知;开展终端与开关营销,提升5G终端适配率与功能激活率。经试点验证,成功将5G驻留比从91.14%提升至92.97%,实

现网络、业务、用户的深度协同优化。本文研究为5G网络优化提供了系统性方法,对提升5G网络效能、改善用户体验具有重要价值,也为后续5G网络运营优化提供了可参考的实践范式与理论支撑。

5 模型实验验证

在分类模型评估中,准确率(accuracy)和召回率(recall)是衡量模型对正类样本预测能力的核心指标。准确率反映了模型预测为正类的样本中实际为正类的比例,其中,TP(true positive)表示正确预测的正类样本数,FP(false positive)表示错误预测为正类的负类样本数,其计算式为:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

召回率则衡量了模型对所有实际正类样本的覆盖能力,其中,FN(false negative)表示实际为正类但被误判为负类的样本数,计算式为:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

基于云南移动2025年3月-4月用户满意度调研数据(样本量 $N=17\,940$,含低满用户8 825例),本节对3种主题划分方案进行实证对比,采用相同的数据预处理流程和测试集划分策略(训练集:

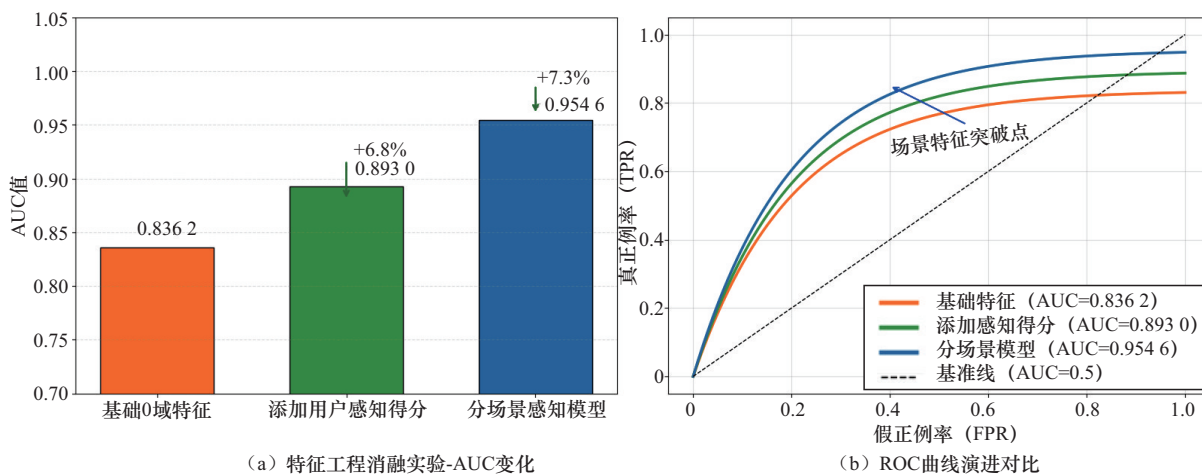


图8 特征工程消融实验对比

测试集=7:3), 结果如下。

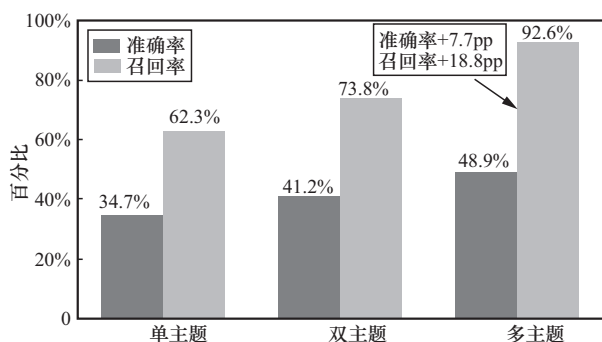


图9 主题模型预测性能对比

多主题模型准确率达48.9%，显著高于双主题模型的41.2%（增加7.7个百分点），验证了细粒度主题划分的预测优势，在“任一主题低满即判定”机制下，召回率提升至92.6%（较双主题增加18.8个百分点），实现更全面的低满用户覆盖。

6 结束语

以本研究的用户低满预测模型为例，通过低满用户回访录音分析、关键词提取、主题建模、用户感知特征构建和机器学习模型训练，实现了潜在低满用户的精准预测，多主题模型通过细粒度主题划分，在召回率提升至92.6%的同时保持准确率48.9%，表明模型在扩大正类覆盖范围时仍维持了合理的预测精度，这一特性使其在需要全面识别潜在低满用户的场景中具有实际应用价值。

参考文献：

- [1] 张鹏, 高源. RFA-XGBoost模型在移动网络潜在投诉用户预测中的应用[J]. 电信科学, 2025, 41(3): 168-178.
- [2] 屈新东, 朱绍柯, 罗凌琦. 基于机器半监督学习的改进自训练

模型在互联网电信业务质差用户识别中的实践研究[J]. 长江信息通信, 2025(4): 218-220.

- [3] 吴迪. 基于机器学习算法的中国移动用户满意度预测研究[D]. 重庆: 重庆大学, 2023.
- [4] 吴亚楠, 尉燕. 基于决策树的低感知用户挖掘方法[J]. 通信与信息技术, 2023(S1): 30-35.
- [5] 张硕伟, 刘贤松, 陈凌, 等. 基于融合AI模型的家庭宽带满意度预测及应用研究[J]. 邮电设计技术, 2025(5): 74-77.
- [6] 董莹莹, 葛阳, 李坤树, 等. 融合AI模型在移网用户满意度预测中的应用研究[J]. 邮电设计技术, 2022(8): 75-78.
- [7] 赵东明, 田雷, 刘静. 事理认知图谱在电信运营商潜在不满预测分析中的应用研究[J]. 通信世界, 2023(7): 42-45.
- [8] 黄霜聆, 谢健, 李荣, 等. 随机森林法对北京移动用户体验影响因素研究[J]. 科学技术创新, 2024(9): 99-102.
- [9] 朱剑伟, 李莹杰, 赵宇韬, 等. 一种应用神经网络和分段模型的家庭宽带低满意度用户识别方法[J]. 江苏通信, 2024, 40(6): 78-82.
- [10] 王珺璇. 运营商用户满意度预测建模与可解释性研究[D]. 北京: 北京邮电大学, 2024.
- [11] 吴娜, 赵东明, 槐秋媛, 等. 政企业务客户感知量化及预测模型的研究[J]. 通讯世界, 2025, 32(3): 7-9.
- [12] 何均均, 王江舟. 基于LightGBM算法的家庭宽带满意度分类研究和实现[J]. 江苏通信, 2023, 39(2): 92-96.

[作者简介]

刘毅（1980-），男，中国移动通信集团云南有限公司工程师，主要研究方向为5G无线网、5G网络优化。

张叶江（1989-），男，中国移动通信集团云南有限公司高级工程师，主要研究方向为5G无线网、5G网络优化。

胡坚（1977-），男，中国移动通信集团云南有限公司高级工程师，主要研究方向为5G无线网、5G网络优化。

唐俊（1991-），男，中国移动通信集团设计院有限公司工程师，主要研究方向为5G无线网、5G网络优化。

董欢欢（1993-），男，中国移动通信集团设计院有限公司工程师，主要研究方向为5G无线网、5G网络优化。

张兆强（1996-），男，中国移动通信集团云南有限公司高级工程师，主要研究方向为5G无线网、5G网络优化。