队伍编号	MCB2201213
赛道	B

北京移动用户体验影响因素研究

摘 要

移动通信技术飞速发展，中国已经步入 5G 时代。但中国国土面积大，人口居住较为集中，因此通信体验成为各移动运营商关注的问题。

根据北京移动的调查问卷，我们小组首先对数据进行处理，主要进行重复值和缺失值检查。我们注意到附件一、附件二未对异常值做处理，我们将对象型数据进行编码处理，使用 3 种编码形式，包括但不限于有序编码、lable 编码等，并删除缺失值大于 80%的指标。考虑到变量值的特点，我们小组将变量分为连续变量和离散变量，并通过相关性分析或满意度分析，选择了 20 个与满意度具有强相关性的连续因素和 26 个离散因素。我们小组使用差异选择方法，对 50 多个独立变量和满意度建立了 LightGBM 回归模型，并对独立变量的贡献进行了排序，以找到显著影响满意度的最佳 23 个因素。考虑到 23 个自变量之间可能存在多重共线性，以确保对变量的高水平解释，计算自变量之间的相关系数，消除自变量之间相关性高的变量，最终得到对满意度影响最大的 20 个变量。最后，计算了所选变量的 MIC 和 Spearman 值，这表明所选变量具有弱相关性和良好的独立性。同时，选择的 20 个变量具有良好的可解释性，这表明 20 个变量的选择是合理的。

针对问题二，我们小组给出了两个解决思路。第一个是迁移学习应用，选择语音和上网共同特征，将语音的样本通过一定的规则选择同与上网有

共同分布的样本一起放入模型训练来预测上网。并且通过类似变量就可以组成一个训练集去预测，从而进行多方案对比。第二个思路是使用 Kmeans 聚类特征生成。分析发现，上网测试集中大部分特征缺失，且这部分特征重要性很高，因此可以使用无监督 Kmeans 聚类将其补全或者训练一个二分类模型补全。进一步分析发现发现′网络覆盖与信号强度′,′手机上网速度′,′手机上网稳定性′与′手机上网整体满意度′呈现高相关，而这类特征只出现在训练集，并未在测试集出现，因此将测试集进行聚类，补充这类特征。

通过研究，我们小组得出研究影响客户语音业务和上网业务满意度的主要因素是分别是语音通话清晰度和上网速度，并通过相关性热图给出各因素对客户打分影响程度的量化分析和结果。

关键词：*LightGBM 回归　　无监督 Kmeans 聚类　　机器学习*

目录

1. 导论

1.1 问题背景

随着移动通信技术的高速发展，人们对带来各种便利的移动通信技术越来越依赖。各大移动运营商也更加重视客户的网络使用体验，致力于让网络服务质量进一步提升。

反映了客户的期望与客户实际体验差异的客户满意度是各大移动运营商运营情况的一个重要体现。各大运营商在产品同质化的今天，需要运用数字经济的管理理念和相关的技术手段，建立一个基于客户体验的全方位系统性测评体系，从而实现数字化转型，推动移动网络可持续性发展。

传统用来提高用户满意度的方法是根据用户投诉来逐点解决。但是在用户数量激增，用户需求提高的今天，传统方法不能有效的提升客户的满意度。本研究拟通过分析影响用户满意度的各种因素，达到更早更全面提升用户满意度的目的。

为了提高用户的上网体验，研究影响用户体验的主要因素，中国移动通信集团北京公司让客户对影响通信体验的各个因素进行打分。同时让客户对语言通话的整体体验和整体满意度进行打分，并且统计整理影响客户语言业务体验的因素。同时，该公司也统计整理影响客户上网体验的因素，希望可以分析影响客户上网业务体验的主要因素

1.2 问题重述与分析问题

已知语言业务用户满意度数据以及预测数据和上网业务用户满意度数据以及预测数据，要求通过数据分析与建模的方法帮助中国移动北京公司分

析影响客户上网业务体验的主要因素。具体地，本文考虑并解决以下两个问题：

问题 1：在语言业务用户满意度数据和上网业务用户满意度已知的情况下，分别研究影响满意度的主要因素，并且给出各因素对客户打分影响程度的量化分析和结果。

问题 2：在问题 1 的基础之上，分别建立基于影响因素的数学模型，并基于此对语言业务用户满意度预测数据和上网业务用户满意度预测数据中的客户打分进行预测，并将结果上传到竞赛平台，说明预测的合理性。

2. 一般性假设和问题简化

假设北京移动提供的数据是真实可靠的，也就是说填写调查问卷的人并没有胡乱填写调查问卷。为简化问题，我们小组将附件所给的变量分为连续型变量和离散型变量，方便后续分析。

3. 变量描述

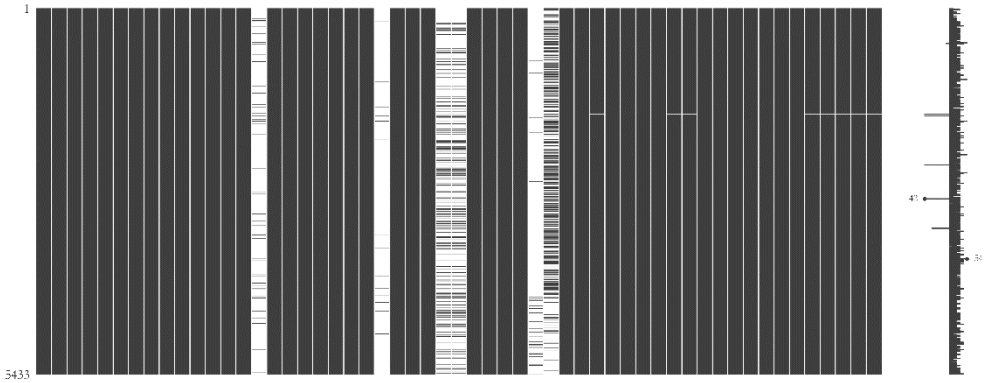
本论文无需引入其他变量，变量名使用附件所给的变量名。

4. 问题一

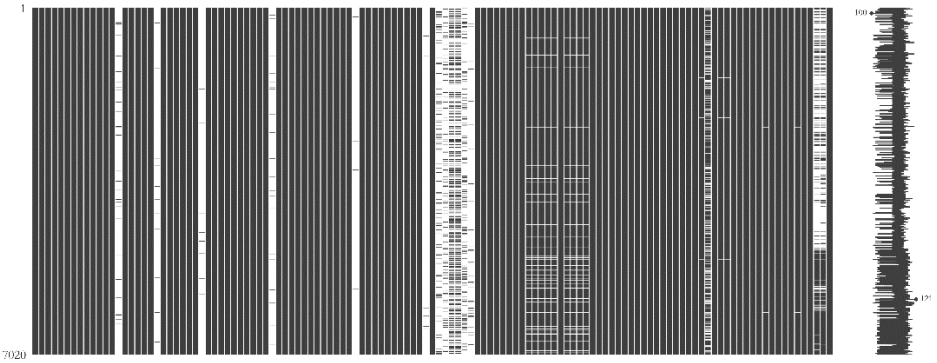
4.1 数据处理

通过宏观查看数据，发现‘用户描述’，‘用户描述.1’，‘重定向次数’，‘重定向驻留时长’，‘是否关怀用户’，‘是否去过营业厅’，缺失值较多，我们选择不要这个字段的列。

‘是否 4G 网络客户（本地剔除物联网）’，‘终端品牌’，‘终端品牌类型’，‘外省流量占比’，‘是否 5G 网络客户’，‘是否实名登记用户’，‘客户星级标识’，‘当月欠费金额’，‘前第 3 个月欠费金额’，缺失值较少，可以选择填充或者删除该行缺失值。

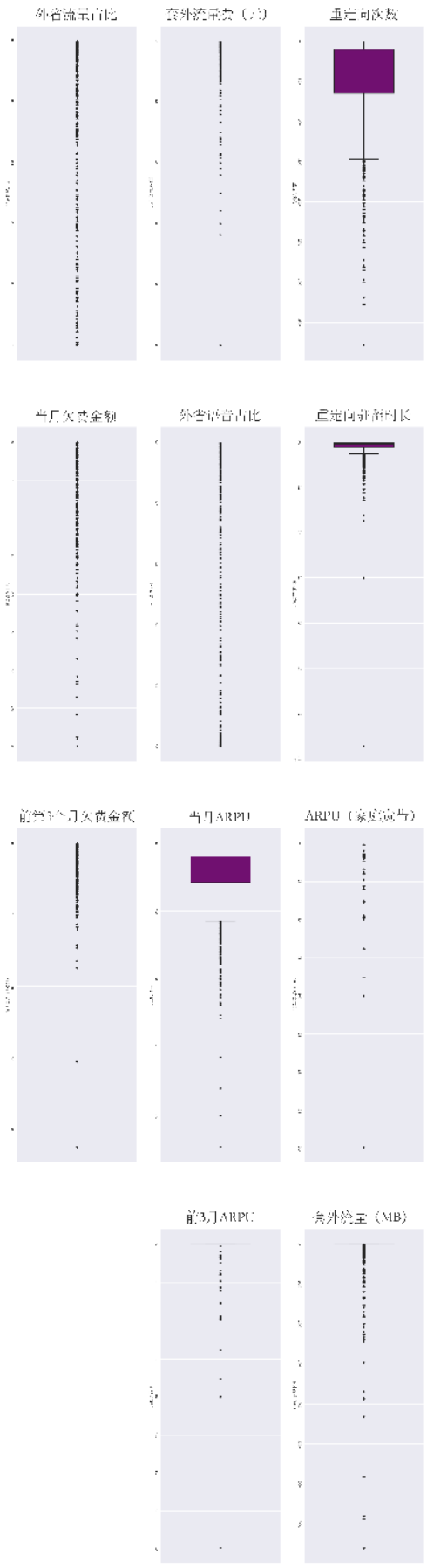


图表 1 附件 1 缺失值

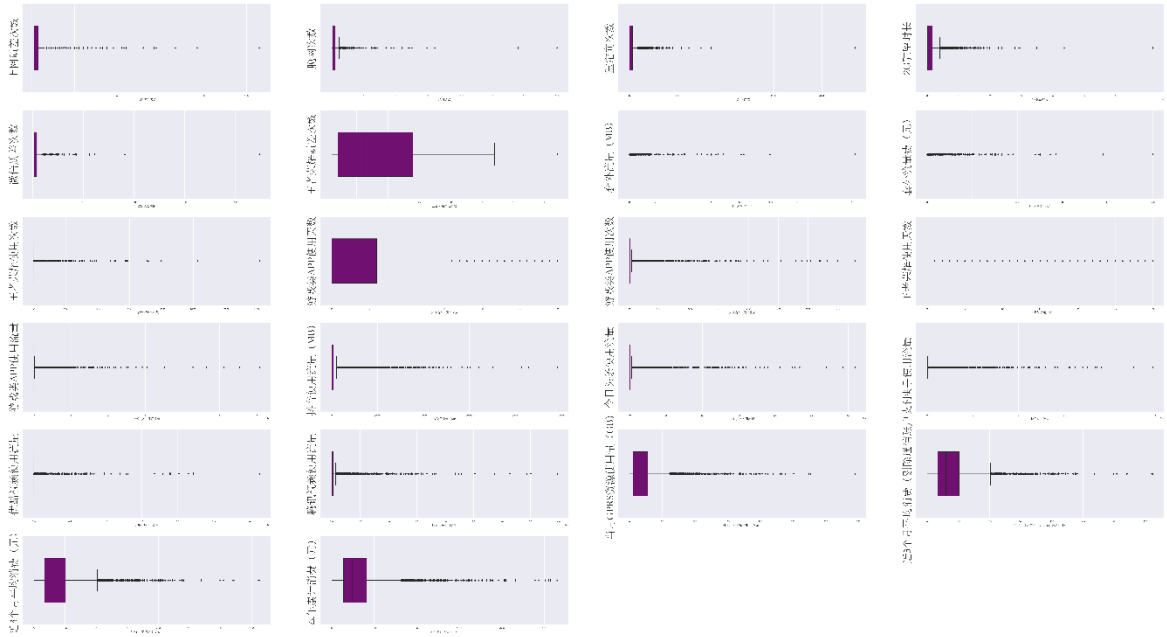


图表 2 附件 2 缺失值

发现异常值较多，进一步查看异常值。



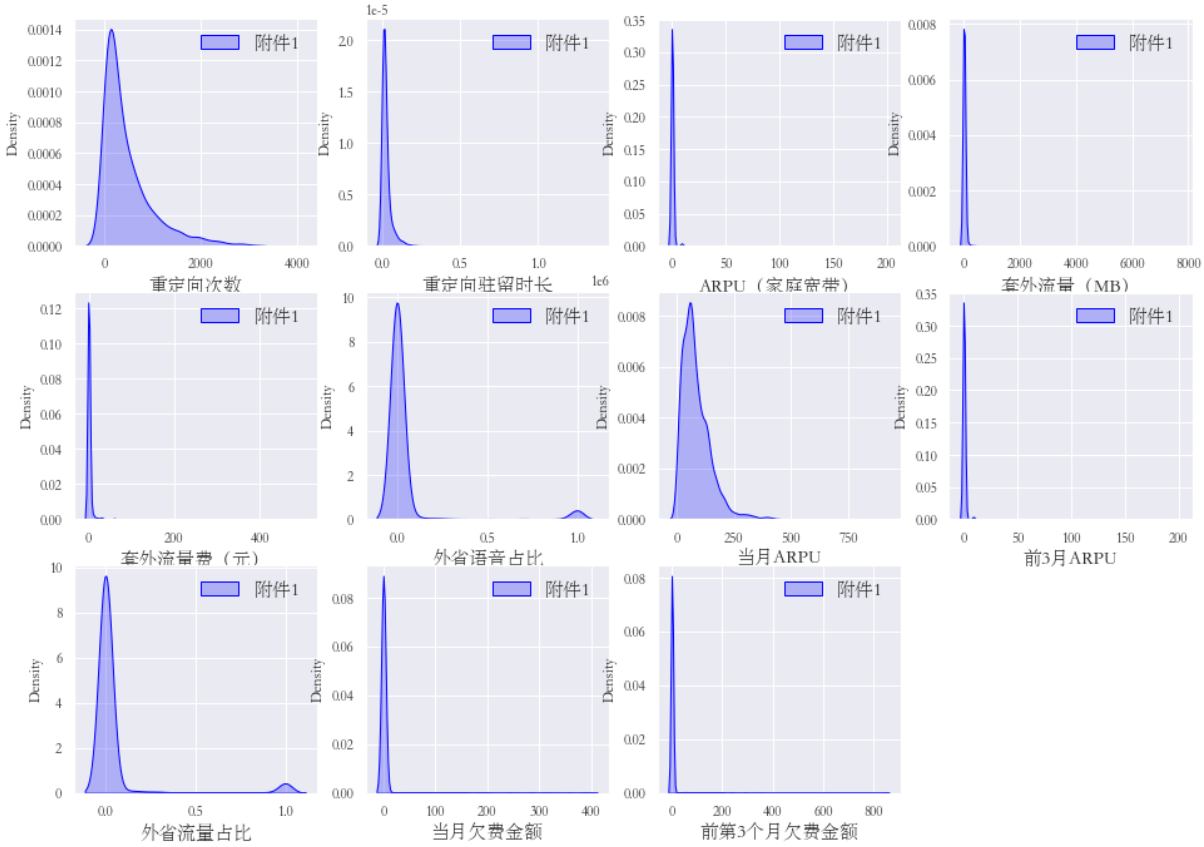
图表 3 附件 1 箱线图



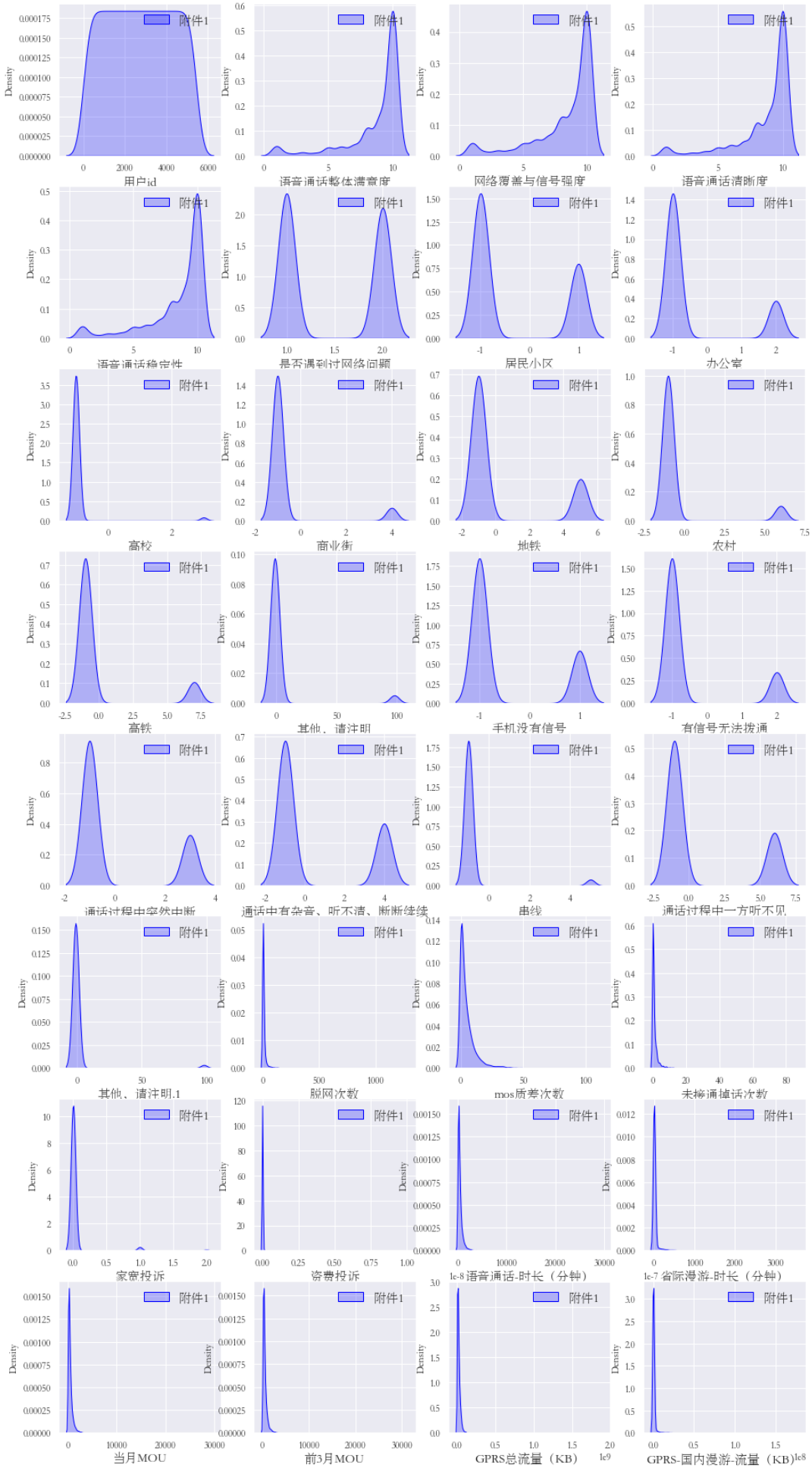
图表 4 附件 2 箱线图

4.2 查看分布

4.2.1 查看附件 1

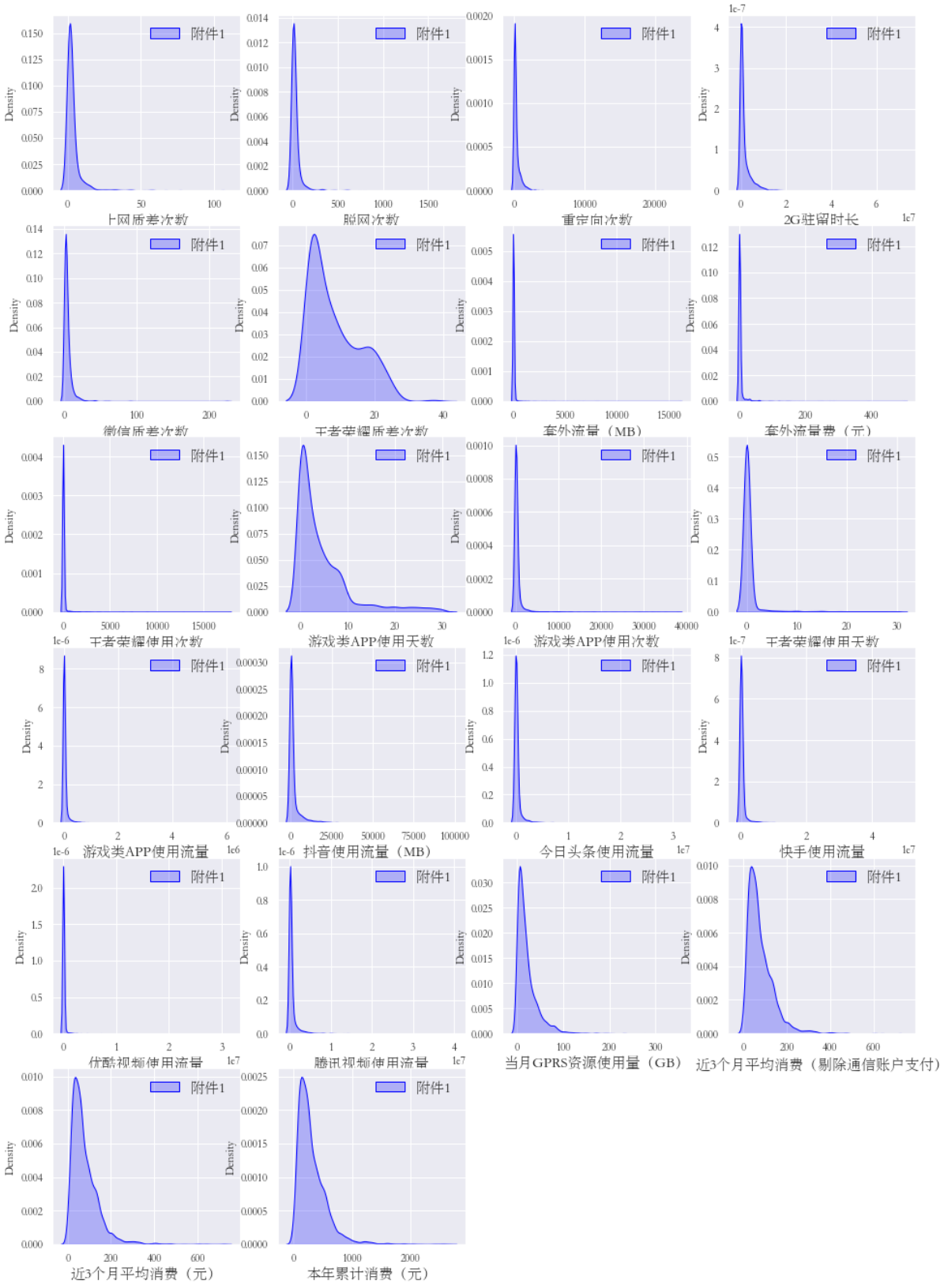


图表 5 附件 1 所有浮点型字段分布

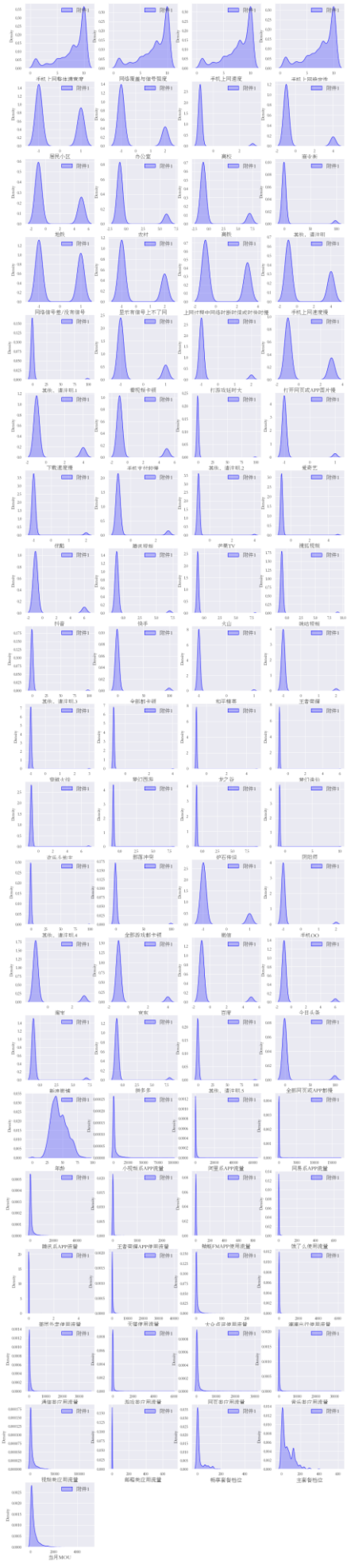

附件1
Density
用户id
语音通话整体满意度
网络覆盖与信号强度
语音通话清晰度
语音通话稳定性
是否遇到过网络问题
居民小区
办公室
高校
商业街
地铁
农村
高铁
其他，请注明
手机没有信号
有信号无法拨通
通话过程中突然中断
通话中有杂音、听不清、断断续续
串线
通话过程中一方听不见
其他，请注明.1
脱网次数
mos质差次数
未接通电话次数
家宽投诉
资费投诉
1e-8 语音通话-时长（分钟）
1e-7 省际漫游-时长（分钟）
当月MOU
前3月MOU
GPRS总流量（KB）
1e9
GPRS-国内漫游-流量（KB）
1e8

图表 6 附件 1 所有 int 型字段分布

4.2.2 查看附件 2



图表 7 附件 2 所有浮点型字段的分布


手机卡同每生病常常
网络覆盖与信号强度
手机上网速度
手机卡网络中断
是否小区
办公室
离校
看书类
地铁
饮时
离校
同路信号差/没有信号
是不是信号之不了网
上网过程中网络时断时续或时快时慢
手机上网速度慢
某市、道详职上
看驾校卡顿
打游戏延时大
打开网页或APP图片慢
下载速度慢
手机平时时慢
某市、道详职上
爱看艺
开配
糖果移植
苹果TV
视频剪辑
抖音
快手
火山
微软剪辑
某市、道详职上
全部数卡顿
地区模糊
王者荣耀
管理卡顿
梦幻西游
生产业
梦幻海长
农业卡顿不
部落冲突
炉石传说
阴阳师
某市、道详职上
全部游戏断卡顿
酷跑
手机OCO
腾讯
京东
百度
今日头条
购买前慢
拼多多
某市、道详职上
全部网页或APP都慢
简易
小型办公APP流量
浏览器APP流量
同事乐APP流量
腾讯系APP流量
王者荣耀APP使用流量
物联网APP使用流量
饿了么使用流量
图形处理使用流量
下载使用流量
大型办公软件流量
图像类文件使用流量
高速率应用流量
高应急应用流量
网页类应用流量
聊天类应用流量
视频类应用流量
邮箱类应用流量
主要包包应
当月MOU

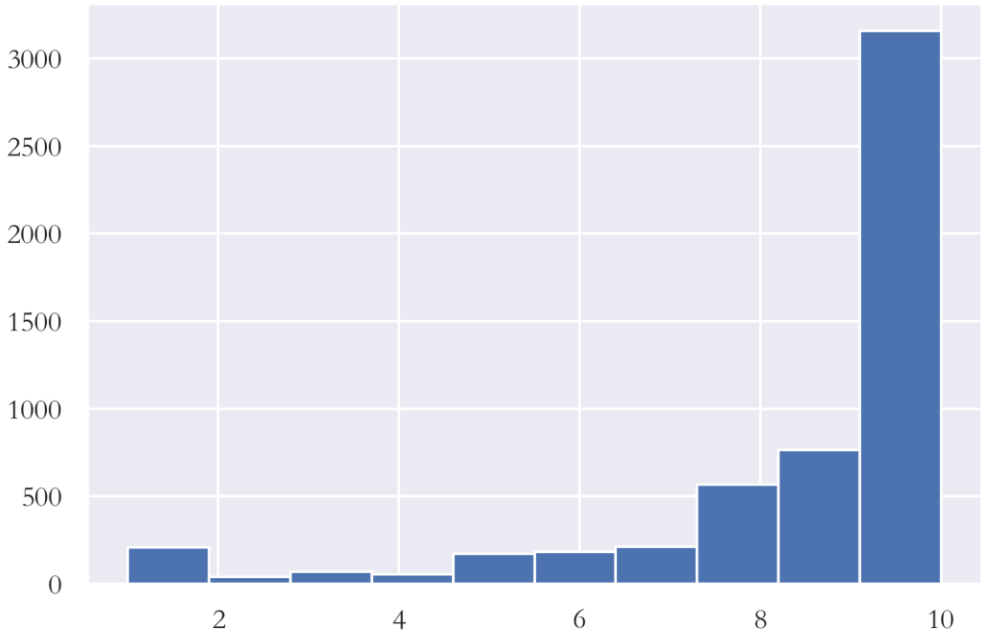
图表 8 附件 2 所有浮点型字段的分布

如果出现某个特征的分布不符合正态分布，需要对数据进行转换，针对右偏和左偏的转换方法不同。

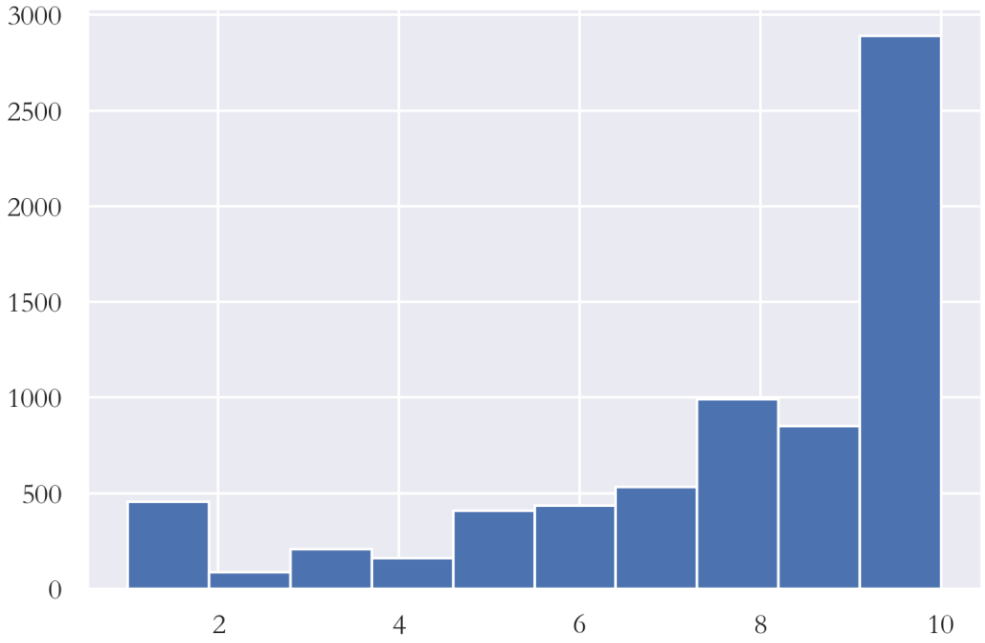
- 大多数右偏数据可以正态化
- 对数变换后呈正态分布，则方差稳定
- 对于不太严重的右偏，使用平方根变换
- 对于严重右偏，使用倒数变换

发现附件 1‘语音通话整体满意度’存在异常值 109000。

count	30000
mean	18.062
std	629.44
min	0.05
25%	6.1
50%	10.48
75%	18
max	109000



图表 9 语音通话整体满意度分布

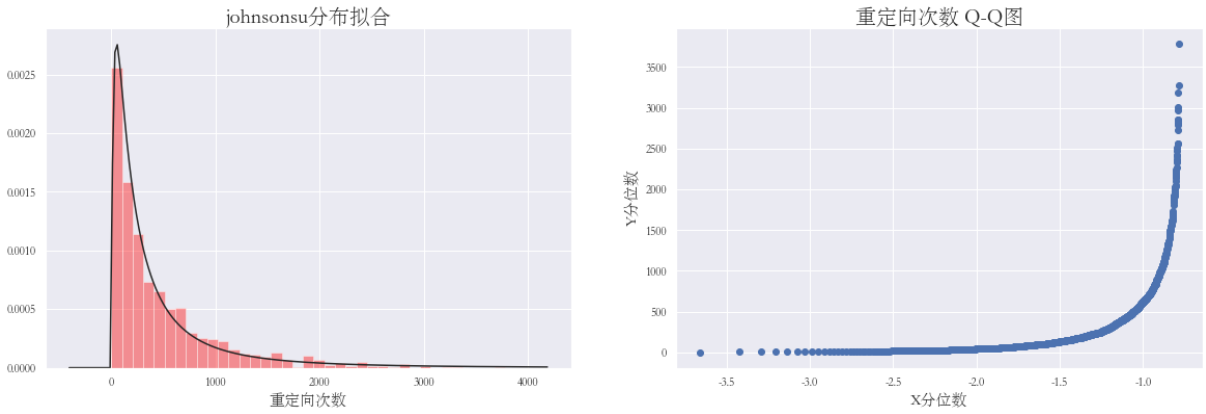


图表 10 手机上网整体满意度分布

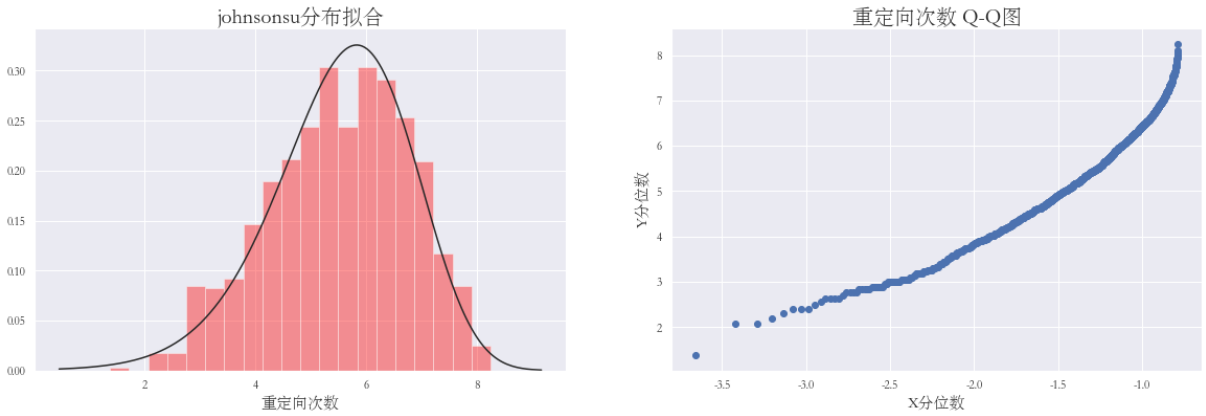
4.3 正态分布转换

我们发现数据分布并不符合正态分布，是左偏数据，回归中对数据分布较为敏感，如果不符合正态分布需要进行数据转换成近似正态分布。

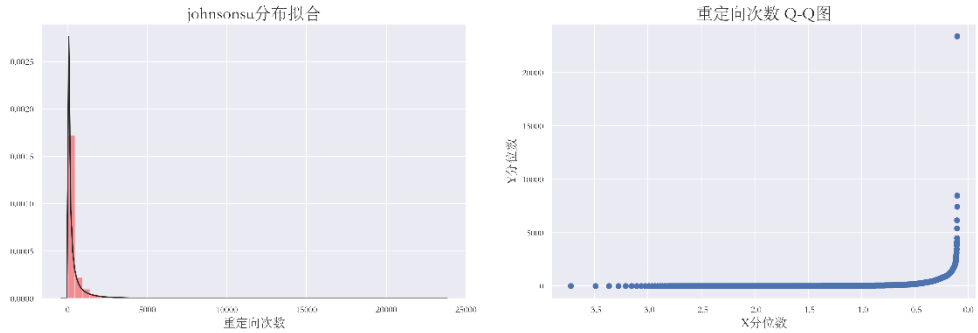
以符合左偏分布的‘重定向次数’，使用对数转换举例。



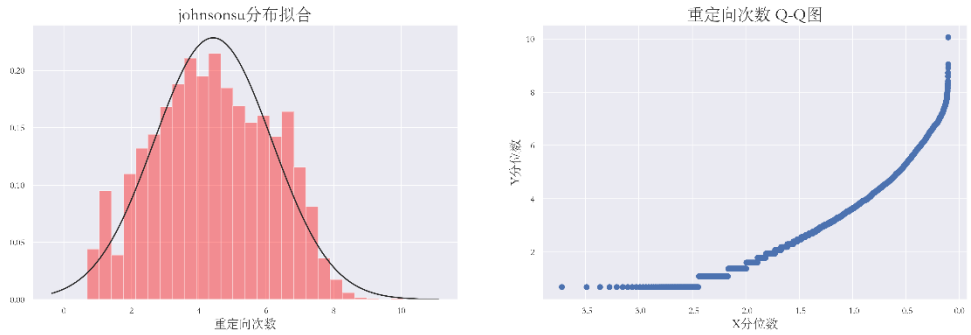
图表 11 附件 1‘重定向次数’未转换前的分布



图表 12 附件 1‘重定向次数’转换后的分布



图表 13 附件 2‘重定向次数’未转换前的分布

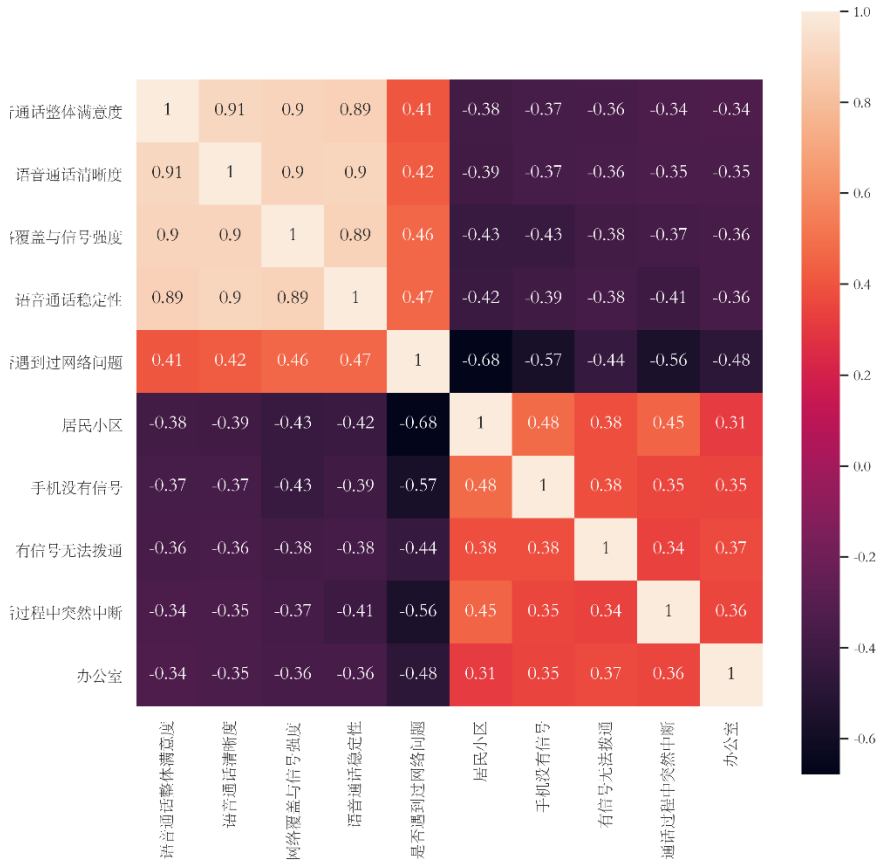


图表 14 附件 2‘重定向次数’转换后的分布

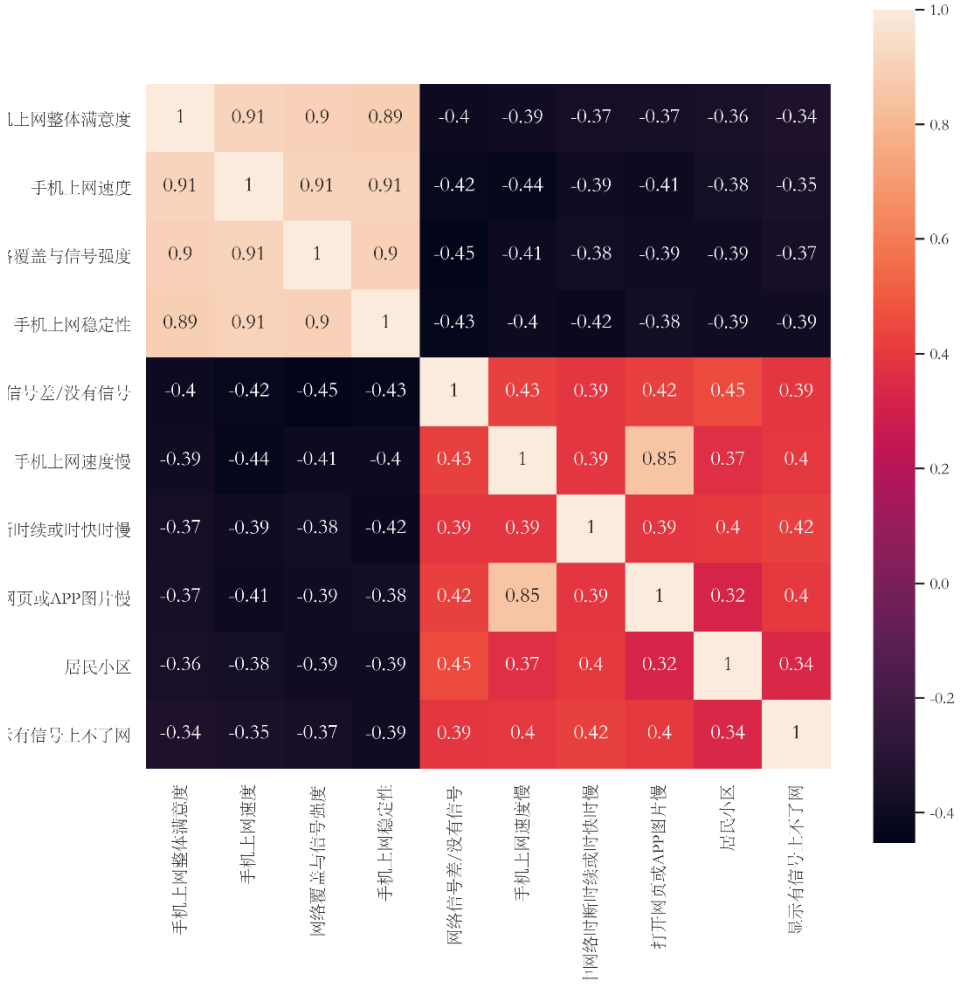
4.4 分析和结论

重定向次数是属于数值特征，存在异常值，数据分布不符合正态分布，需要对数转换后，便可以近似为正态分布。

4.5 查看相关性



图表 15 附件 1-Top10 相关性热图



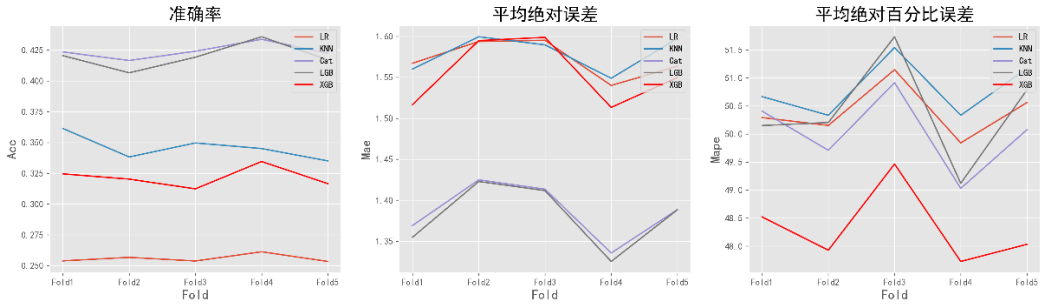
图表 16 附件 2-Top10 相关性热图

我们小组得出研究影响客户语音业务和上网业务满意度的主要因素是分别是*语音通话清晰度*和*上网速度*，相关性热图见上。

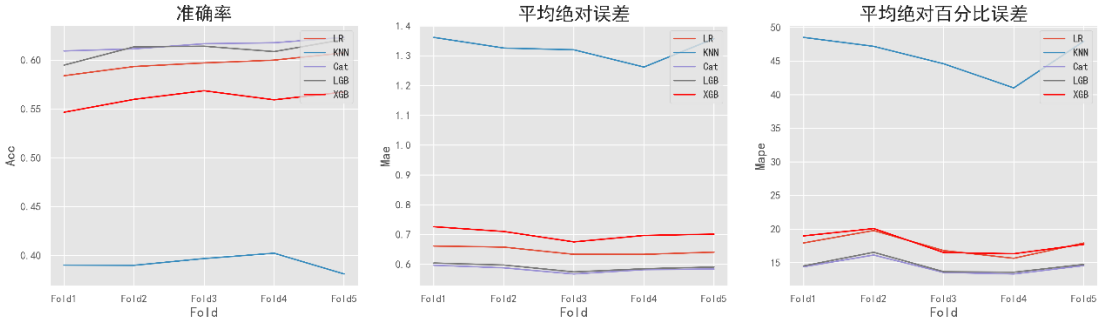
5. 问题二

5.1 使用迁移学习应用分析解决问题

选择语音和上网共同特征，将语音的样本通过一定规则选择与上网有共同分布的样本放入多个模型训练，预测上网。通过类似变量就可以组成一个训练集去预测。经过模型训练和评价，我们得到如下结果：



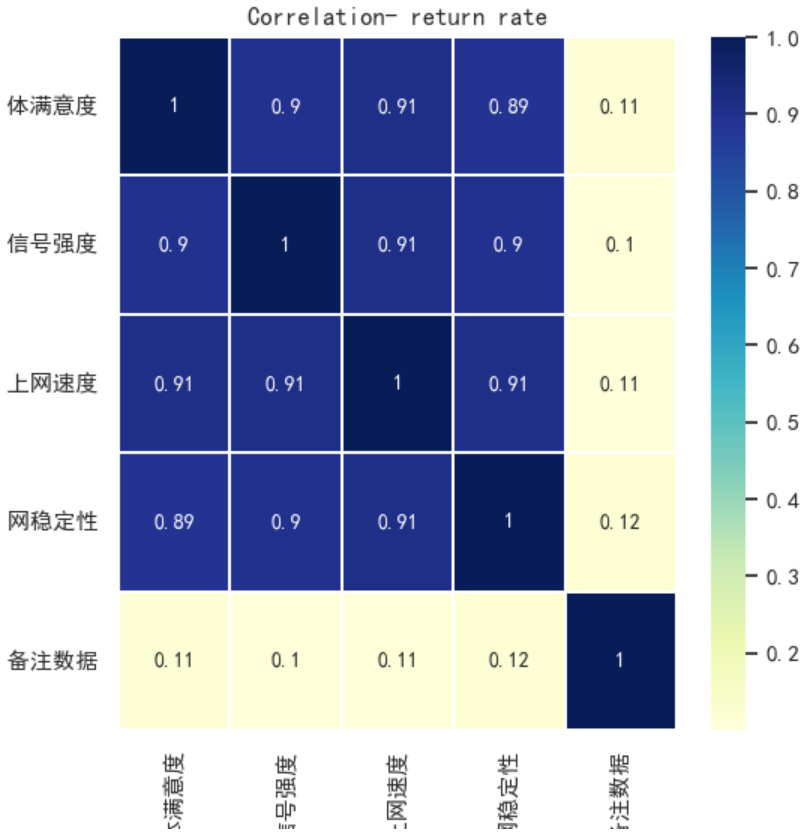
图表 17 迁移学习多模型对比



图表 18 Kmeans 聚类多模型对比

5.2 使用无监督 Kmeans 聚类分析解决问题

通过分析发现，上网测试集中大部分特征缺失，且这部分特征重要性很高，因此可以使用无监督 Kmeans 聚类将其不全或者训练一个二分类模型补全。

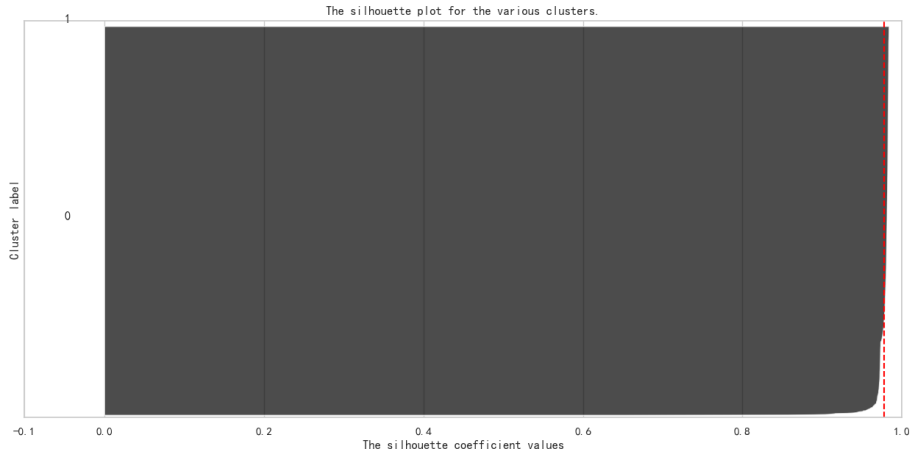


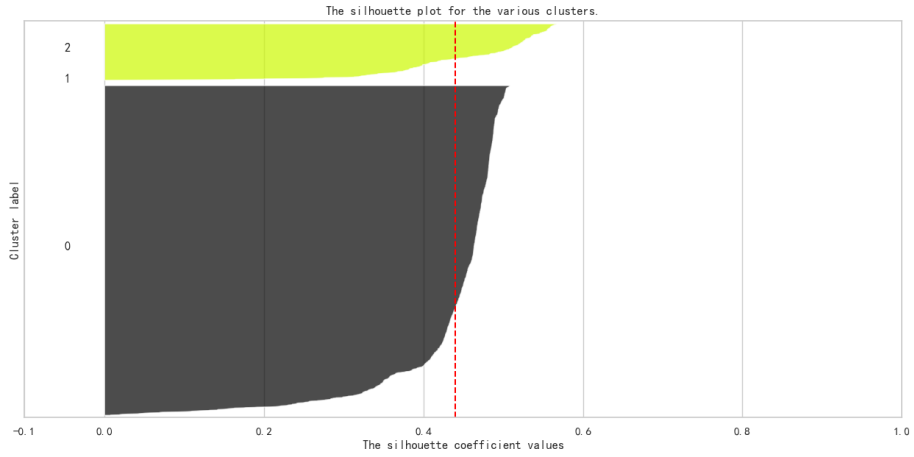
图表 19 相关-回报率

发现′网络覆盖与信号强度′,′手机上网速度′,′手机上网稳定性′与′手机上网整体满意度′呈现高相关，而这类特征只出现在训练集，并未在测试集出现，因此将测试集进行聚类，补充这类特征。

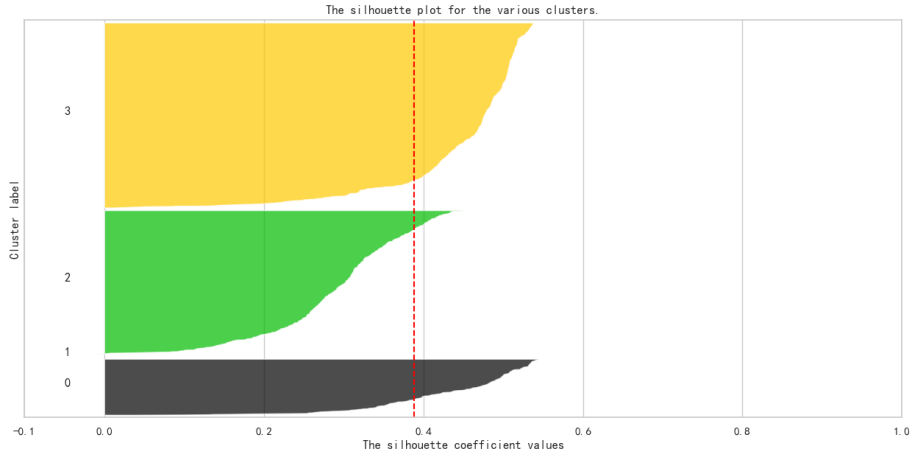
将上网测试集聚类得到新的特征，找到共同特征，先使用 5 个簇尝试聚类实例化 K-Means 算法模型，之后使用数据集 X 进行训练，调用属性 labels_，查看聚类结果。

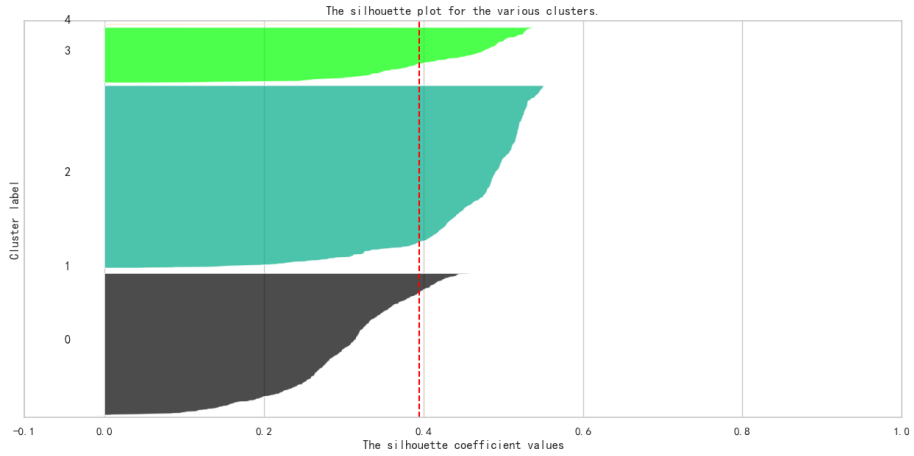
之后查看每一分类结果的数量、 轮廓系数均值、每一样本轮廓系数和样本轮廓系数结果的数组结构。



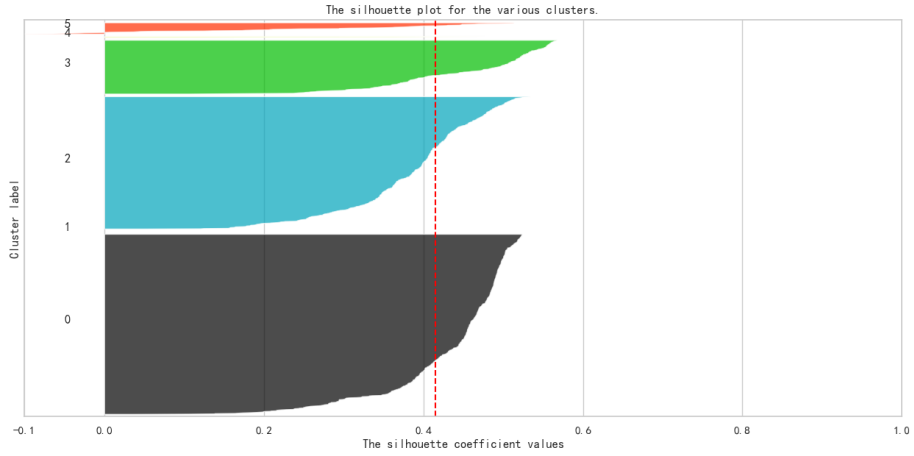


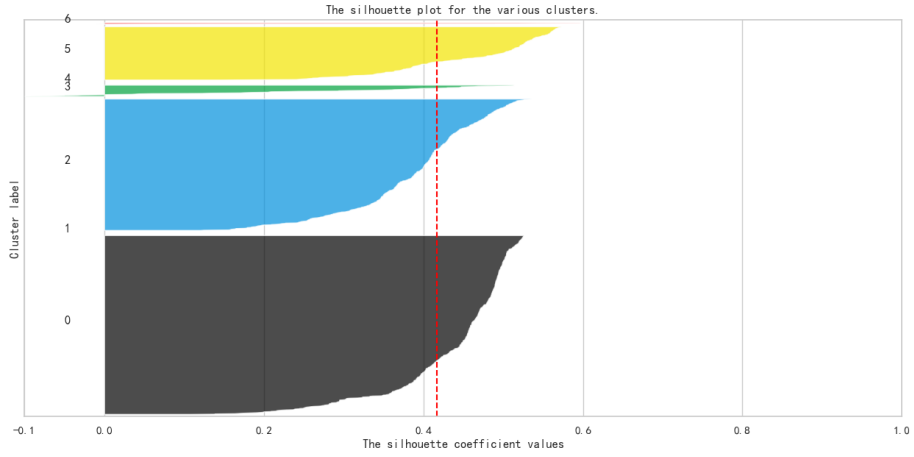
图表 20 簇数为 2 ，轮廓系数均值为 0.9776851414909911



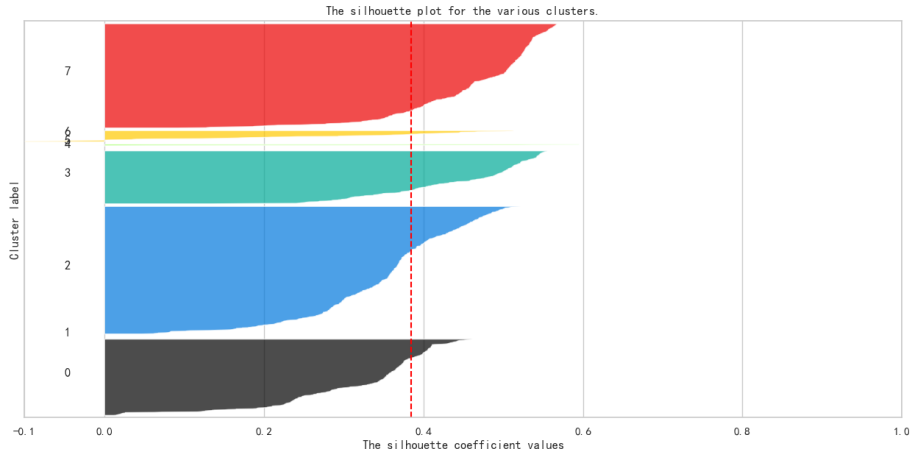


图表 21 簇数为 4 ，轮廓系数均值为 0.3882082232971846





图表 22 簇数为 6 ，轮廓系数均值为 0.41473931428541516



图表 23 簇数为 8 ，轮廓系数均值为 0.3845556444721305

5.3 分析和结论

将迁移学习和 Kmeans 聚类的结果平均后，导出结果如下：

result.xlsx

6. 敏感性分析

假设无监督 Kmmeans 聚类算法中 K 的值是合理的，则本文无需敏感性分析。我们小组认为 K 是合理的。

7. 优缺点分析

7.1 优点

i. 我们剔除了缺失值和异常值，使得结论更加可靠；

ii. 我们直接 LightGBM 回归、无监督 Kmmeans 聚类等热门机器学习方案，模型更加高级，结论更加可靠。具体地说，首先，LightGBM 回归速度

更快，占用内存更小。其次，迁移学习可以充分利用模型之间存在的相似性；无监督 Kmmeans 聚类算法能根据较少的已知聚类样本的类别对树进行剪枝确定部分样本的分类。为克服少量样本聚类的不准确性，该算法本身具有优化迭代功能，在已经求得的聚类上再次进行迭代修正剪枝确定部分样本的聚类，优化了初始监督学习样本分类不合理的地方。最后，由于只是针对部分小样本可以降低总的聚类时间复杂度。

7.2 缺点

i. 我们无法 100%相信北京移动提供的数据是真实可靠的，也就是说没有办法可以保证数据的 100%客观真实；

ii. 我们的模型是有缺点的，具体的说，对于 LightGBM 模型，可能会长出比较深的决策树，产生过拟合。因此 LightGBM 在 Leaf-wise 之上增加了一个最大深度限制，在保证高效率的同时防止过拟合；另外 Boosting 族是迭代算法，每一次迭代都根据上一次迭代的预测结果对样本进行权重调整，所以随着迭代不断进行，误差会越来越小，模型的偏差（bias）会不断降低。由于 LightGBM 是基于偏差的算法，所以会对噪点较为敏感；最后在寻找最优解时，依据的是最优切分变量，没有将最优解是全部特征的综合这一理念考虑进去；对于迁移学习，缺点在于模型参数不易收敛；无监督 K-means 算法中 K 是事先给定的，这个 K 值的选定是非常难以估计的。很多时候，事先并不知道给定的数据集应该分成多少个类别才最合适。其次，在 K-means 算法中，首先需要根据初始聚类中心来确定一个初始划分，然后对初始划分进行优化。这个初始聚类中心的选择对聚类结果有较大的影响，一旦初始值选择的不好，可能无法得

到有效的聚类结果。最后，该算法需要不断地进行样本分类调整，不断地计算调整后的新的聚类中心，因此当数据量非常大时，算法的时间开销是非常大的。

8. 参考文献

[1]赵衡,彭铃,李云飞.基于改进 K-means 聚类算法的上市公司信用风险评估研究[J].内江师范学院学报,2022,37(12):77-83.DOI:10.13603/j.cnki.51-1621/z.2022.12.013.

[2]白雨佳,李靖,高升.基于最优 K 均值聚类算法的负荷大数据任务均衡调度研究[J].电力电容器与无功补偿,2022,43(06):85-91.DOI:10.14044/j.1674-1757.pcrpc.2022.06.013.

[3]张超,闫茹玉,朱晓君.基于用户需求意象的多重 K-means-ELM 侗锦配色模型研究[J].丝绸,2022,59(12):108-118.

[4]曲福恒,潘曰涛,杨勇,胡雅婷,宋剑飞.基于加权空间划分的高效全局 k-means 聚类算法[J/OL].吉林大学学报(工学版):1-8[2022-12-29].DOI:10.13229/j.cnki.jdxbgxb20221338.

[5]周成龙,陈玉明,朱益冬.粒 K 均值聚类算法[J/OL].计算机工程与应用:1-10[2022-12-29].http://kns.cnki.net/kcms/detail/11.2127.tp.20221213.1343.009.html

[6]赵天瑞,刘一鸣,吕鹏召,黎彦良,唐晓秘,郭威,张军,田禹.无废城市建设场景下 LightGBM 垃圾产量预测模型构建[J/OL].环境工程:1-7[2022-12-29].http://kns.cnki.net/kcms/detail/11.2097.X.20221209.1402.005.html

[7]张怡,谢晓金.基于 K-means 聚类与粗糙集的个人信用集成分类模型[J/OL].

软件导刊:1-6[2022-12-29].http://kns.cnki.net/kcms/detail/42.1671.tp.20221207.1121.024.html

[8]何选森,何帆,徐丽,樊跃平.K-Means 算法最优聚类数量的确定[J].电子科技大学学报,2022,51(06):904-912.

[9]欧阳群文.基于 LightGBM 的水质预测模型研究与应用[J].智能城市,2022,8(11):84-87.DOI:10.19301/j.cnki.zncs.2022.11.028.

[10]段央央,陈光宇,张仰飞,邓湘.基于 LightGBM-IndRNN 模型的现货市场日前电价预测[J/OL].水力发电:1-6[2022-12-29].http://kns.cnki.net/kcms/detail/11.1845.tv.20221122.1011.004.html

[11]吴丹,雷珽,李芝娟,王宁,段艳.基于 XGBoost 与 LightGBM 集成的电动汽车充电负荷预测模型[J].电子技术应用,2022,48(09):44-49.DOI:10.16157/j.issn.0258-7998.212316.

[12]孙泉,耿磊,赵奇慧,杨佳昊,吕平,李莉.基于 LightGBM 的温室番茄冠层 CWSI 预测模型研究[J].农业机械学报,2022,53(S1):270-276+308.

[13]王瑶. 基于改进 LightGBM 的分类预测算法研究与应用[D].北京石油化工学院,2022.DOI:10.27849/d.cnki.gshyj.2022.000046.

[14]曾海潇. 基于 LightGBM-GRU 的新能源股票价格预测模型[D].西南大学,2022.DOI:10.27684/d.cnki.gxndx.2022.002389.

[15]田烜瑜,汪旭杰,史恩泽,陈思奇.基于 LightGBM 的在线磁盘故障预测模型[J].电子技术与软件工程,2022(07):249-253.

[16]唐一峰.基于 XGBoost 算法和 LightGBM 算法的贷款违约预测模型研究[J].现代计算机,2021,27(32):33-37.