官网访客流失建模说明文档
Version2.7

****有限公司

2017 年 07 月

修订历史记录

日期	版本	描述	作者
2017-03-16	1.0	初稿	A 员工，B 员工，C 员工
2017-03-30	1.1	根据需求会议要求，进行修改	A 员工，B 员工，C 员工
2017-04-01	1.2	完善聚类模型部分	A 员工，B 员工，C 员工
2017-04-06	1.3	完善文档	A 员工，B 员工，C 员工
2017-04-07	1.4	合并客户名单，更新价值模型说明	A 员工，B 员工，C 员工
2017-04-16	1.5	按客户例会需求，调整聚类结果	A 员工，B 员工，C 员工
2017-04-20	1.6	更新价值得分分段统计，完善客户分群类别定义描述	A 员工，B 员工，C 员工
2017-04-27	1.7	完善文档，新增流失模型介绍	A 员工，B 员工，C 员工
2017-05-05	1.8	价值模型分析	A 员工，B 员工，C 员工
2017-05-10	1.9	完善文档	A 员工，B 员工，C 员工
2017-05-11	2.0	流失模型目标值定义更改并重跑所有模型	A 员工，B 员工，C 员工
2017-05-12	2.1	加入 R、F、M 三个输入变量价值模型，并和原模型进行对比，调整流失模型，完善文档细节	A 员工，B 员工，C 员工
2017-05-17	2.2	完善价值模型得分统计	A 员工，B 员工，C 员工
2017-05-24	2.3	调整 RFM 关键访问模块	A 员工，B 员工，C 员工
2017-05-27	2.4	增加 RFM 模型加入关键模块访问后的价值得分迁移情况	A 员工，B 员工，C 员工
2017-06-09	2.5	5 月份数据验证	A 员工，B 员工，C 员工
2017-06-23	2.6	增加混淆矩阵相关指标解释	A 员工，B 员工，C 员工
2017-07-03	2.7	修正表 14 中的混淆矩阵结果	A 员工，B 员工，C 员工

目录

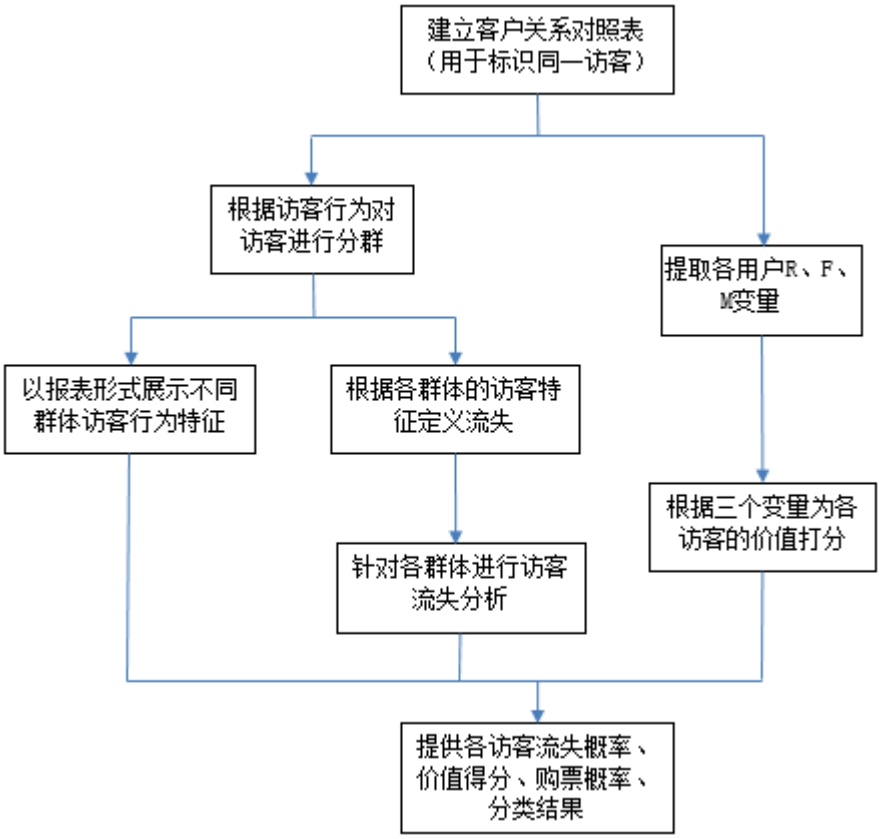
第一章摘要

1.1 背景介绍

本报告针对所提供的 2017 年 1 月 1 日-2017 年 4 月 30 日期间 webtrends、et、cki、订座 PNR、订单数据等源数据，使用 webtrends、et、cki、订单数据等源数据，通过数据挖掘，对访客进行分群，对各访客群体进行流失分析，以便于后续营销活动的制定。

1.2 方案内容

1.2.1 方案总体设计



注：R、F、M 将在 1.2.2 中有详细介绍

方案总体设计简介：

1、针对登录过官网的会员整理出客户关系对照表，将访客的客户号和信息（手机号、邮箱、qq 号、微信号、webtrends_id 等）整合到一张宽表中。

2、针对所有访客进行价值打分；同时根据访客行为对访客进行分群，根据各群体的访客特征定义流失规则，进行各群体的流失分析。

3、向营销相关部门提供各访客流失概率、价值得分、购票概率、行为特征等信息。

1.2.2 访客价值

综合 R、F、M 三个变量和关键模块访问情况变量（收银支付页面访次、预定行程页面访次），对所有访客的价值进行打分。其中 R 为访客最近一次购票时间到观察结束时间的天数，F 为访客在观察时间段内成功购票的次数，M 为访客在观察时间段内总购票金额。

1.2.3 访客分类

基于访客行为(包括官网访问及购票行为、乘机行为等)通过聚类模型将访客进行类别划分，生成各类别访客行为特征的报表。

1.2.4 访客流失

根据 1.2.3 各类别访客的不同行为特征，定义各不同类别访客流失规则，分别进行访客流失分析，最终得到各类访客流失概率，结合访客的价值、购票概率、行为特征提供给营销相关部门。

1.2.5 购票概率

结合影响购票概率的可能因素，依据访客历史访问信息，对访客购票概率进行预测分析，最终训练得到各类访客购票概率，同访客的价值、流失概率、行为特征一并提供给营销相关部门。

1.3 价值意义

1、对访客进行细分，更有效地判断有价值访客，了解他们的特征和实际需求，对不同价值的访客采取不同的营销策略，将有限的资源投放到最有价值的访客身上，实现精准化营销，提高企业的竞争力，最终实现提升航空客运的上座率目标。

2、分析官网访客的行为模式，跟踪访客关注的航班航线但没有形成订单的行为，挖掘流失的访客，结合航班剩余座位，进行二次销售，提升活动响应率，减少虚耗座位。

1.4 报告内容说明

1、第二章针对本报告所使用的数据源、数据清洗思路、模型算法进行简要说明。

2、第三章主要包括分析结果的说明。

3、第四章 相关产出文档，包括提数脚本、代码脚本、访客价值、分类、流失概率、购票概率名单（部分名单）。

第二章数据源说明

2.1 数据处理思路

本次分析的数据源为：2017 年 01 月 01 日至 2017 年 04 月 30 日的官网日志访问记录，其中观测窗口为 2017 年 01 月 01 至 2017 年 03 月 31 日，验证窗口 2017 年 04 月 01 日至 2017 年 04 月 30 日。数据清洗分为两个阶段：

1） 第一阶段数据清洗过程在 hive 中完成，初步形成各个模型所要求的数据，共有 85 个字段，包括新特征构建、缺失值处理。

2） 第二阶段数据清洗过程在 R 中完成，主要在第一阶段数据的基础上对数据质量进行提升处理，包括异常值处理、离群值处理、数据归一化。

2.2 数据清洗过程中涉及的内容

1、缺失值处理：对变量中存在的缺失数据，根据实际业务意义对数据进行插补。

2、异常值处理：访客数据中一些变量存在逻辑上的异常，直接剔除该记录。

3、离群值处理：访客数据中一些变量数据存在一些数据分布的离群值点，处理方法为分离出大于均值加若干倍标准差外的数据。

4、数据归一化：对输入变量单位进行归一化，使得变量间具有可比性，以满足聚类模型建立要求。

2.3 变量总览

经过第一阶段的清洗，得到的 85 个相关初步变量目录如下。

输入变量.xlsx

2.4 模型算法介绍

2.4.1 访客价值分析

1、选用模型：RFM+关键模块访问模型、PCA 算法

2、输入变量：最近一次购票时间到观察结束时间的天数 R，观察时间段内成功购票的次数 F，观察时间段内总购票金额 M，收银支付页面访次，预定行程页面访次。

3、输出变量：访客价值得分

3、模型介绍：1）RFM+关键模块访问模型是航空领域里结合实际业务背景，识别客户价值应用最广泛的模型。即通过航空公司访客数据中的：最近行为的时间间隔（Recency）、行为频率（Frequency）、行为价值（Monetary）、收银支付页面访次、预定行程页面访次对客户进行细分，识别出潜在价值高的客户；2）主成分分析（PCA）是一种应用广泛的线性降维方法，该方法通过将特征集缩减成一小部分能代表原始特征集最主要变化的主要特征分量，来实现高维数据到低维数据空间的映射。

4、算法过程： 1）先从样本数据进行标准化；2）执行 PCA 算法，对数据进行压缩，得到一维得分列；3）生成相应得分权重值；4）生成客户价值得分名单。

5、模型调优：无

6、模型评估：随机取得的样本集，根据访客价值计算公式计算得到的值跟模型的值要完全一致。

2.4.2 访客行为分群

1、选用模型：k-means 聚类

2、输入变量：最后一次访问时间距观测窗口结束时间天数、最后一次购票时间距观测窗口结束时间天数、观测窗口内访次数、观测窗口内成功购票次数、观测窗口会话总时长、观测窗口内消费金额、机票查询访问次数、航班选择访问次数、旅客信息访问次数、机票查询总停留时间、航班选择总停留时间、旅客信息总停留时间、付费搜索次数、非付费搜索次数、非会员手机登陆次数、会员登陆次数、是否会员、会员日访问次数、PC 端访问次数、移动端访问次数、访问间隔时间、支付订单访问次数、支付出错次数、头等舱次数、公务舱次数、明珠经济舱次数、经济舱次数、节假日飞行次数、官网购票国际出行次数、网上值机次数、乘机和购票总间隔时间、折扣票总价、票面价总和、EDM 来源访次、百度 SEM 来源访次、360SEM 来源访次、SM-SEM 来源访

次、谷歌 SEM 来源访次、搜狗 SEM 来源访次、AD 来源访次、LIST 来源访次、移动官网页面访次、预定行程页面访次、首页，服务大厅页面访次、明珠会员页面访次、收银支付页面访次、机票预定页面访次、员工专区页面访次、提前选座页面访次、网上值机页面访次、明珠商城页面访次、抽奖等营销活动页面访次、其他页面访次。

3、模型介绍：聚类分析是在没有给定划分类别的情况下，根据数据相似度进行样本分组。它可以建立在无类标记的数据上，是一种非监督学习算法。K-means 聚类算法是典型的基于距离的非层次聚类算法，在最小化误差函数的基础上将数据划分为预定的 k 类，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度越大。

4、算法过程：1）先从样本数据中随机选取 k 个对象作为初始聚类中心；2）分别计算每个样本到各类聚类中心的距离，将对象分配到距离最近的聚类中；3）所有对象分配完成后，重新计算 k 个聚类中心；4）与前一次计算得到的 k 个聚类中心比较，如果聚类中心发生变化，转 2），否则转 5）；5）当中心不发生变化时停止并输出聚类结果。

5、模型调优：当训练数据变量数量（维度）比较多时，需要先对训练数据进行 PCA 降维，因此需要确定最佳的降维个数，主要根据碎石图来寻找拐点确定降维数；k-means 聚类中需要确定的参数是聚类数目 k,该文档综合考虑下面两种方式加以确定:1)根据实际业务确定类目数 k；2）根据 Within-Cluster Sum of Squaresn 图寻找拐点确定聚类最优类目数。

关于拐点确定方法，就是寻找折线由陡峭变为平稳的点位置，结合下图，具体寻找拐点思路如下：1）绿线为需要寻找拐点的折线；2）用反比曲线进行拟合，对图中绿色折线进行平滑处理，即红色曲线；3）连接绿色折线的首尾点，计算两点之间的斜率值 slope_0，即图中蓝色直线斜率；4）计算红色拟合曲线在各个整数点位置的切线斜率值 slope_k，图中未画出切线；5）寻找一个切线斜率和 slope_0 最接近的 slope_k 所对应的点位置，作为绿色折线的拐点位置。

上述拐点确定方法的应用场景在于 PCA 降维个数和 kmeans 聚类最优数目 k 的确定。

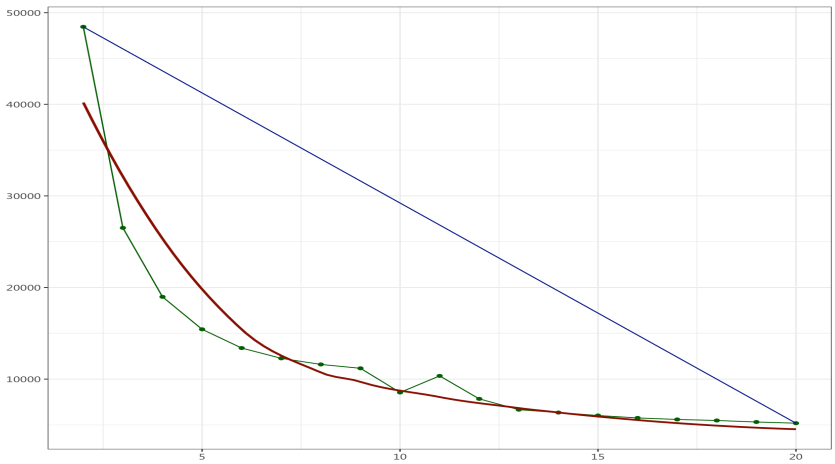


图 1 拐点确定方法示意图

6、模型评估：1）聚类优度标准，R 方尽可能达到 0.6～0.8 以上；2）模型应该具有稳定性，对于随机抽取的不同检验样本，形成的分群结果，应该具有相似的分群结构（分群数、各特征群体所占比例等等），随着时间的推移，分群结果也应该相对稳定，如果分群结构具有周期性变化的规律，应给予合理的说明；3）得出的分群结果应具有明显的特征，且要有业务指导意义；4）模型应该有监测机制以及更新机制。

2.4.3 访客流失分析和购票概率预测

1、选用模型：lasso 算法+logit 回归

2、输入变量：最后一次访问时间距观测窗口结束时间天数、最后一次购票时间距观测窗口结束时间天数、观测窗口内访次数、观测窗口内成功购票次数、观测窗口会话总时长、观测窗口内消费金额、机票查询访问次数、航班选择访问次数、旅客信息访问次数、机票查询总停留时间、航班选择总停留时间、旅客信息总停留时间、付费搜索次数、非付费搜索次数、非会员手机登陆次数、会员登陆次数、是否会员、会员日访问次数、PC 端访问次数、移动端访问次数、访问间隔时间、支付订单访问次数、支付出错次数、头等舱次数、公务舱次数、明珠经济舱次数、经济舱次数、节假日飞行次数、官网购票国际出行次数、网上值机次数、乘机和购票总间隔时间、折扣票总价、票面价总和、EDM 来源访次、百度 SEM 来源访次、360SEM 来源访次、SM-SEM 来源访次、谷歌 SEM 来源访次、搜狗 SEM 来源访次、AD 来源访次、LIST 来源访次、移动官网页面访次、预定行程页面访次、首页，服务大厅页面访次、明珠会员页面访次、收银支付页面访次、机票预定页面访次、员工专区页面访次、提前选座页面访次、网上值机页面访次、明珠商城页面访次、抽奖等营销活动页面访次、其他页面访次。

3、目标值定义：1）转化率角度：验证窗口范围内，访客购票次数占访客访问次数的比例大于 0，定义为非流失（有购票），记为 1；比例等于 0，定义为流失（未购票），记为 0；2）留存率角度：观测窗口内最后一个月内，访客访次数大于 0，且验证窗口范围内，访客访次数大于 0，定义为非流失（留存访客），记为 1；否则，定义为流失（非留存访客），记为 0。

4、模型介绍：1)lasso 算法通过构造一个惩罚函数获得一个精炼的模型，通过最终确定一些指标的系数为零，实现了指标集合精简的目的。基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化，从而能够产生某些严格等于 0 的回归系数，得到解释

力较强的模型；2）logit 回归主要用于因变量为分类变量（如流失概率、购票概率）的回归分析，自变量可以为分类变量，也可以为连续变量。他可以从多个自变量中选出对因变量有影响的自变量，并可以给出预测公式用于预测。

5、算法过程：1）按 4：1 划分训练集和测试集；2）用交叉验证计算 lasso 算法中的 lambda 值；3）得到 2）lambda 下的输入变量集合；4）用训练集训练 logit 模型；5）用测试集对模型进行评估。

6、模型调优：lasso 算法中，lambda 的值需要用交叉验证法进行确定，其中使用 10 折 leave-one-out 法估计 lambda。

7、模型评估：1）模型应具有一定程度的准确性及稳定性。命中率在总流失率尚不清楚的情况下，无法设定明确目标。ROC 曲线应该呈现上凸形态，ROC 曲线下的面积，即 AUC 值应该尽量达到 0.75，前 10 百分位的提升度应达到 5 倍以上，回归模型拟合的 R 方应达到 0.75 以上；2）模型应避免过拟合，在同期多个检测样本上的预测效果，应与训练样本的效果基本一致。模型应该试运行若干周期（或者在建模时预留足够的检测窗口），预测结果及各项统计指标，应与训练样本基本一致；3）在评估模型准确率时，设置预测的时间窗口为某段历史时段，用历史数据来判断是否流失，流失预测准确率要达到 75%以上。

第三章分析结果

3.1 访客价值分析

3.1.1 建模流程

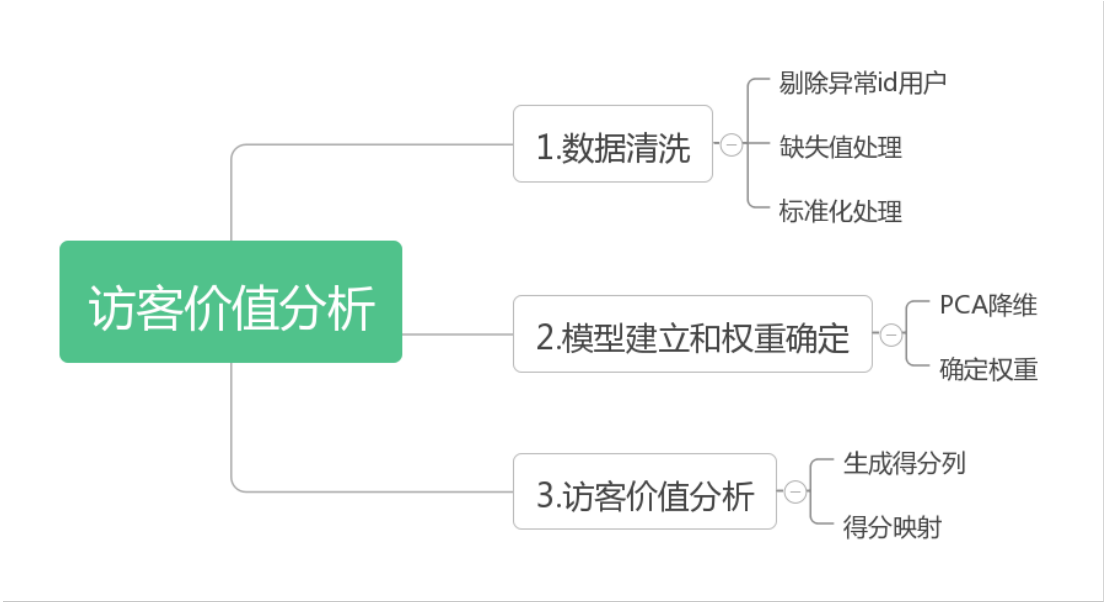


图 2 客户价值分析流程图

3.1.2 数据清洗

导入数据量大小为 1801629 条用户行为数据。

1、缺失值处理：对于最近一次购票时间到观察结束时间的天数，用观测窗口长度+1 进行插补；

2、最后的数据量大小为 1801619 条。

3.1.3 数据分布特征

经过清洗的分布情况如表 1 所示。其中，R 为最近一次购票时间到观察结束时间的天数，F 为观察时间段内成功购票的次数，M 为观察时间段内总购票金额。

表 1 输入变量描述统计

变量名	R	F	M	收银支付页面访次	预定行程页面访次
均值	89.20853	0.021987	41.66712	0.041615	0.474853
标准差	6.948284	0.479596	1126.202	0.571514	2.18105
最小值	0	0	0	0	0
最大值	90	418	751380	426	1218
极差	90	418	751380	426	1218
标准误	0.005177	0.000357	0.839044	0.000426	0.001625

3.1.4 模型建立与权重确定

从购票价值角度对客户价值进行评估，只需对 R、F、M、收银支付页面访次、预定行程页面访次五个输入变量执行一次 PCA 算法，压缩为一维得分列，贡献率情况如下表所示，作为对比，也给出单独对 R、F、M 三个变量执行 PCA 算法结果。

表 2 价值模型 PCA 贡献度情况

RFM+关键模块访问	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.79255	0.92375	0.777293	0.501905	0.278120
Proportion of Variance	0.64265	0.17066	0.12084	0.05038	0.01547
Cumulative Proportion	0.64265	0.81331	0.93415	0.98453	1
RFM	**PC1**	**PC2**	**PC3**		
Standard deviation	1.419491	0.903155	0.411529		
Proportion of Variance	0.67165	0.2719	0.05645		
Cumulative Proportion	0.67165	0.94355	1		

经计算，得到的计算公式如下：

PCA_SCORE_6=3.0687401038989-0.0362518556943907*R+1.10210906263923*F+0.000430207332266411*M+0.907252886951932*收银支付页面访次+0.179679718561159*预定行程页面访次

PCA_SCORE_6 表示基于购票价值和官网关键模块访次的访客价值得分。为方便评估，对上述 PCA 得分进行变换，将 PCA 得分转化为 0-100 之间的分数，用于最终评估访客价值得分，变换公式共分为两步：

1）PCA 得分=log(PCA_SCORE_6-min(PCA_SCORE_6)+0.0001)

2）访客价值得分=(PCA 得分-min(PCA 得分))/(max(PCA 得分)-min(PCA 得分))*100

其中，第一步主要把得分分布区分开，使得分更分散，而第二步将得分映射至 0-100 区间。将最后生成相应客户价值名单见附件 4.3。

3.1.5 访客价值分析

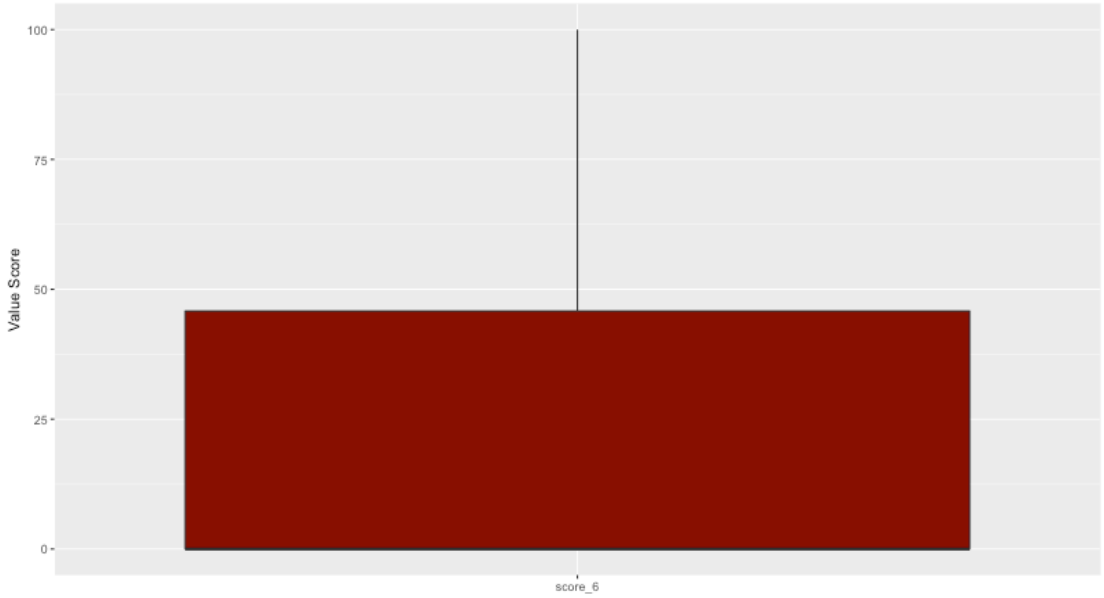


图 3 客户价值得分箱线图

图 3 给出了客户价值得分箱线图，从中可以看出，得分分布情况总体呈现出右偏形态，大多数访客均处在价值得分的中间偏低区域。访客价值得分的频数分布如表 3 所示：

表 3 价值得分频数分布表

RFM+关键模块访问		RFM	
得分区间	访客数量	得分区间	访客数量
0	1178383	0	1771646
(0,10]	0	(0,10]	0
(10,20]	0	(10,20]	0
(20,30]	0	(20,30]	0
(30,40]	0	(30,40]	0
(40,50]	508045	(40,50]	0

(50, 60]	80214	(50, 60]	83
(60, 70]	32295	(60, 70]	26450
(70, 80]	2544	(70, 80]	3321
(80, 90]	131	(80, 90]	112
(90, 100]	7	(90, 100)	7

下表给出 RFM 模型各区间的价值得分，在 RFM+关键模块访问的价值模型下的价值得分分布，如表 4 所示。列为 RFM 模型得分区间，行为加入关键模块访问后的 RFM 价值得分。可以发现，在加入关键模块访问变量后，价值得分分布发生了一定的变化。

表 4 访客 RFM 零价值得分分布情况

得分区间	0 得分迁移情况	(50, 60] 迁移情况	(60, 70] 迁移情况	(70, 80] 迁移情况	(80, 90] 迁移情况	(90, 100] 迁移情况
0	1178383	0	0	0	0	0
(0, 10]	0	0	0	0	0	0
(10, 20]	0	0	0	0	0	0
(20, 30]	0	0	0	0	0	0
(30, 40]	0	0	0	0	0	0
(40, 50]	508045	0	0	0	0	0
(50, 60]	80207	0	7	0	0	0
(60, 70]	4906	83	26100	1206	0	0
(70, 80]	103	0	343	2083	15	0
(80, 90]	1	0	0	32	96	2
(90, 100)	1	0	0	0	1	5

选择收银支付页面访次和预定行程页面访次作为访问关键模块变量的原因是，这两个变量对购票次数和购票金额存在较强的正相关关系。如图 4，行为购票次数(gpcs)、购票金额(gp_amt)，列为各访问模块页面访次数，相对应的数字为相关系数，颜色越深表示相关关系越强。其中，收银支付页面访次（syzf_fc)、预定行程页面访次（ydxc_fc）对购票次数（或购票金额）存在较强的正相关，说明收银支付页面访次和预定行程页面访次会直接影响到购票次数和购票金额。因此认定，收银支付页面访次和预定行程页面访次是影响购票行为的两个主要的访问关键模块。

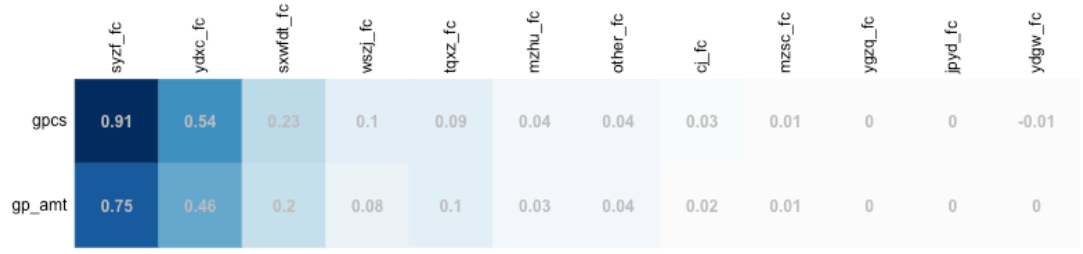


图 4 购票次数(gpcs)、购票金额（gp_amt）和各访问模块的相关系数矩阵图

3.2 访客行为分群

3.2.1 建模流程

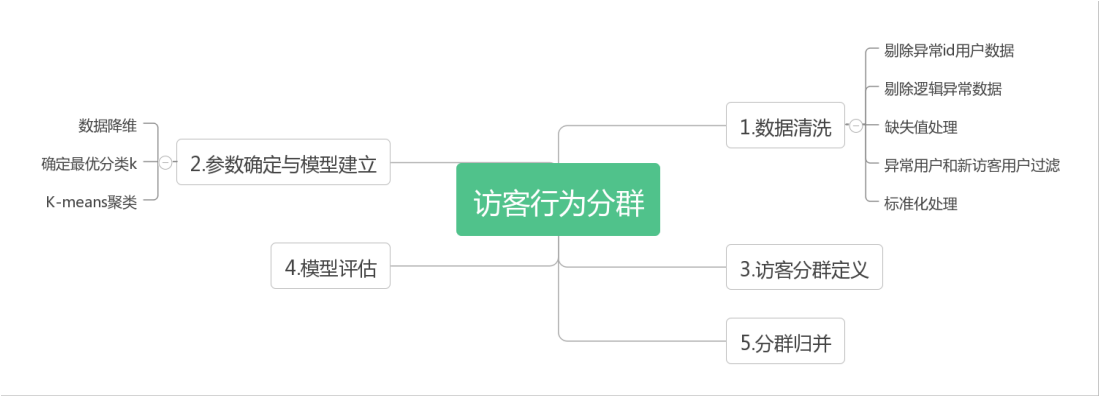


图 5 访客行为分群建模流程

3.2.2 数据清洗

导入数据量大小为 1801629 条。

1、剔除异常 id 用户：对于客户号异常的用户，删去记录；

2、剔除逻辑异常数据：对于一些特征存在逻辑上错误，如乘机和购票总间隔时间出现负值，应删除相应记录，此步删去 10 条记录；

3、缺失值处理：最近一次购票时间到观察结束时间的天数、乘机和购票总间隔时间存在部分缺失值，处理方法如下：1）最近一次购票时间到观察结束时间的天数：用观测窗口长度+1 进行插补；2）乘机和购票总间隔时间：0 值插补；

4、低频访客过滤：定义低频访客规则（访次数=1 次，购票次数=0 次，会话总时长<1 分钟），对于观测期间内，对符合低频访客规则的用户过滤出去，并归于低频访客群体，过滤 580099 条记录；

5、离群访客处理：对消费金额、会话总时长、机票查询总停留时间、航班选择总停留时间、旅客信息总停留时间和票面价总和进行遍历，计算每个特征变量是否大于均值+1.5 倍标准差，至少一个特征符合的记录就将其过滤出去，并归于离群访客群体，此步过滤 33945 条记录；

6、对剩余数据进行标准化处理，最后的数据量大小为 1187575 条记录。

3.2.3 数据分布特征

表 5 输入变量描述统计

变量名	均值	标准差	最小值	最大值	极差	标准误
最后一次访问时间距观测窗口结束时间天数	40.0408	26.3379	0	89	89	0.01962
最后一次购票时间距观测窗口结束时间天数	89.2085	6.9483	0	90	90	0.00518
观测窗口内访次数	1.5377	5.4857	1	5633	5632	0.00409
观测窗口内成功购票次数	0.0220	0.4796	0	418	418	0.00036
观测窗口会话总时长	43.7110	629.1943	0	528973.2	528973.2	0.46876
观测窗口内消费金额	41.6671	1126.2022	0	751380	751380	0.83904
机票查询访问次数	0.5817	1.7275	0	585	585	0.00129
航班选择访问次数	0.5257	1.8888	0	649	649	0.00141
旅客信息访问次数	0.0826	0.7709	0	444	444	0.00057
机票查询总停留时间	330.0607	5338.1948	0	1924335	1924335	3.97707
航班选择总停留时间	494.4290	11485.7108	0	6480298	6480298	8.55710
旅客信息总停留时间	71.6902	3867.2765	0	4011346	4011346	2.88120
付费搜索次数	0.1922	1.1770	0	581	581	0.00088
非付费搜索次数	0.0005	0.0747	0	60	60	0.00006
非会员手机登陆次数	0.0051	0.1682	0	92	92	0.00013
会员登陆次数	0.1781	2.9643	0	3208	3208	0.00221
会员日访问次数	0.0694	0.2542	0	1	1	0.00019
PC 端访问次数	0.1175	0.4900	0	198	198	0.00037
移动端访问次数	0.6564	5.1795	0	5570	5570	0.00386
访问间隔时间	0.8875	1.8455	0	605	605	0.00137
支付订单访问次数	3.1499	11.4500	0	89	89	0.00853
支付出错次数	0.0615	0.7511	0	479	479	0.00056
头等舱次数	0.0018	0.1683	0	131	131	0.00013
公务舱次数	0.0000	0.0091	0	7	7	0.00001
明珠经济舱次数	0.0009	0.0545	0	18	18	0.00004
经济舱次数	0.0024	0.0675	0	15	15	0.00005
节假日飞行次数	0.0203	0.6191	0	495	495	0.00046
官网购票国际出行次数	0.0029	0.0848	0	39	39	0.00006
网上值机次数	0.0020	0.0659	0	30	30	0.00005
乘机和购票总间隔时间	0.0020	0.0951	0	77	77	0.00007
折扣票总价	0.6269	16.1803	0	6520	6520	0.01205
票面价总和	0.2348	3.5444	0	1700	1700	0.00264
EDM 来源访次	37.4445	1238.7360	0	889880	889880	0.92288
百度 SEM 来源访次	0.0057	0.0981	0	22	22	0.00007
360SEM 来源访次	0.1478	0.8807	0	307	307	0.00066
SM-SEM 来源访次	0.0287	0.5000	0	225	225	0.00037

谷歌 SEM 来源访次	0.0002	0.0218	0	15	15	0.00002
搜狗 SEM 来源访次	0.0000	0.0086	0	8	8	0.00001
AD 来源访次	0.0163	0.5250	0	547	547	0.00039
LIST 来源访次	0.0133	0.1410	0	19	19	0.00011
移动官网页面访次	0.0273	0.2010	0	26	26	0.00015
预定行程页面访次	0.6483	1.5032	0	203	203	0.00112
首页，服务大厅页面访次	0.4749	2.1811	0	1218	1218	0.00162
明珠会员页面访次	0.8031	4.7753	0	5275	5275	0.00356
收银支付页面访次	0.1032	1.5073	0	1543	1543	0.00112
机票预定页面访次	0.0416	0.5715	0	426	426	0.00043
员工专区页面访次	0.0222	0.2578	0	63	63	0.00019
提前选座页面访次	0.0183	0.6588	0	499	499	0.00049
网上值机页面访次	0.0459	0.7863	0	838	838	0.00059
明珠商城页面访次	0.0446	0.8774	0	605	605	0.00065
抽奖等营销活动页面访次	0.0164	0.2596	0	92	92	0.00019
其他页面访次	0.0315	0.2508	0	38	38	0.00019

3.2.4 参数确定与模型建立

最后进行训练模型的数据共 1187575 条记录，输入变量特征有 52 个，需要先对数据进行降维，这里采用的降维算法为 PCA，关于 PCA 算法思想，详见 2.4.1 中的介绍。

由于输入变量维数较大，且变量之间存在一定的冗余信息，利用 2.4.2 节中介绍方法确定 PCA 方差值构成的碎石图中的拐点，寻找最优降维数，碎石图见图 6。最终确定拐点位置为 7，降维数量为 7，将输入变量由原来的 52 维降至 7 维，相应的贡献度如下表所示。

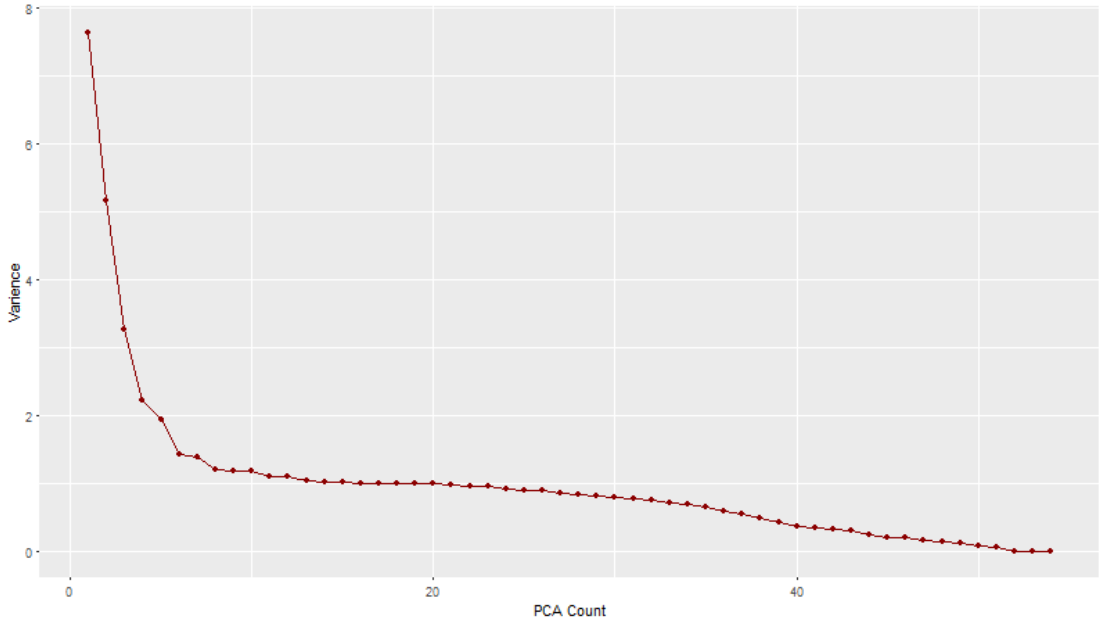


图 6 PCA 碎石图

表 6 聚类模型 PCA 降维贡献度情况

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	2.7639	2.2705	1.8080	1.4945	1.3956	1.1931	1.1759	1.1004	1.0878	1.0873	1.0536	1.0462	1.0226
Proportion of Variance	0.1415	0.0955	0.0605	0.0414	0.0361	0.0264	0.0256	0.0224	0.0219	0.0219	0.0206	0.0203	0.0194
Cumulative Proportion	0.1415	0.2369	0.2975	0.3388	0.3749	0.4013	0.4269	0.4493	0.4712	0.4931	0.5137	0.5339	0.5533
	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25	PC26
Standard deviation	1.0139	1.0060	1.0026	1.0007	1.0000	0.9976	0.9949	0.9850	0.9802	0.9747	0.9578	0.9491	0.9428
Proportion of Variance	0.0190	0.0187	0.0186	0.0185	0.0185	0.0184	0.0183	0.0180	0.0178	0.0176	0.0170	0.0167	0.0165
Cumulative Proportion	0.5723	0.5911	0.6097	0.6282	0.6467	0.6652	0.6835	0.7015	0.7193	0.7369	0.7538	0.7705	0.7870
	PC27	PC28	PC29	PC30	PC31	PC32	PC33	PC34	PC35	PC36	PC37	PC38	PC39
Standard deviation	0.9272	0.9186	0.9076	0.8906	0.8783	0.8737	0.8446	0.8338	0.8020	0.7693	0.7375	0.6962	0.6557
Proportion of Variance	0.0159	0.0156	0.0153	0.0147	0.0143	0.0141	0.0132	0.0129	0.0119	0.0110	0.0101	0.0090	0.0080
Cumulative Proportion	0.8029	0.8185	0.8338	0.8485	0.8628	0.8769	0.8901	0.9030	0.9149	0.9259	0.9359	0.9449	0.9529
	PC40	PC41	PC42	PC43	PC44	PC45	PC46	PC47	PC48	PC49	PC50	PC51	PC52
Standard deviation	0.6091	0.5856	0.5735	0.5488	0.4840	0.4576	0.4438	0.3933	0.3777	0.3443	0.2807	0.2448	0.0536
Proportion of Variance	0.0069	0.0064	0.0061	0.0056	0.0043	0.0039	0.0037	0.0029	0.0026	0.0022	0.0015	0.0011	0.0001
Cumulative Proportion	0.9597	0.9661	0.9722	0.9778	0.9821	0.9860	0.9896	0.9925	0.9951	0.9973	0.9988	0.9999	1

经过 PCA 降维后，分别筛选降维后的每种特征（共 7 维）中的主要影响变量，筛选依据基于各个主成分的系数绝对值，从大到小依次筛选，详细描述如下：

表 7 各主成分主要影响的输入变量

主成分	主要影响输入变量	主成分定义
PC1	最后一次购票时间距观测窗口结束时间天数、最后一次访问时间距观测窗口结束时间天数	近期互动程度
PC2	观测窗口内访次数、移动端访问次数、最后一次购票时间距观测窗口结束时间天数、访问间隔时间、移动官网页面访次、观测窗口会话总时长、首页，服务大厅页面访次	官网访问互动程度
PC3	PC 端访问次数、预定行程页面访次、首页，服务大厅页面访次	PC 端行为维度
PC4	会员登陆次数、抽奖等营销活动页面访次、明珠会员页面访次、会员日访问次数、网上值机页面访次、提前选座页面访次、明珠商城页面访次	会员成熟度
PC5	付费搜索次数、百度 SEM 来源访次、最后一次访问时间距观测窗口结束时间天数、移动官网页面访次、网上值机页面访次、搜狗 SEM 来源访次、提前选座页面访次	访客来源和提前规划偏好维度
PC6	旅客信息访问次数、折扣票总价、收银支付页面访次、网上值机次数、明珠经济舱次数、支付订单访问次数、明珠商城页面访次	订单行为维度
PC7	抽奖等营销活动页面访次、百度 SEM 来源访次、付费搜索次数	广告与营销相关维度

对降维后的数据集再进行 kmeans 聚类，其中最优的聚类数目 k 同样采用 2.4.2 节介绍的拐点确定方法对 Within-Cluster Sum of Squaresn 曲线图进行拐点确定，得到最优聚类数 k=8。说明从 k=8 开始，曲线开始趋于平缓，如需再增加新类别，其组内差异的变化大体上将越来越小。

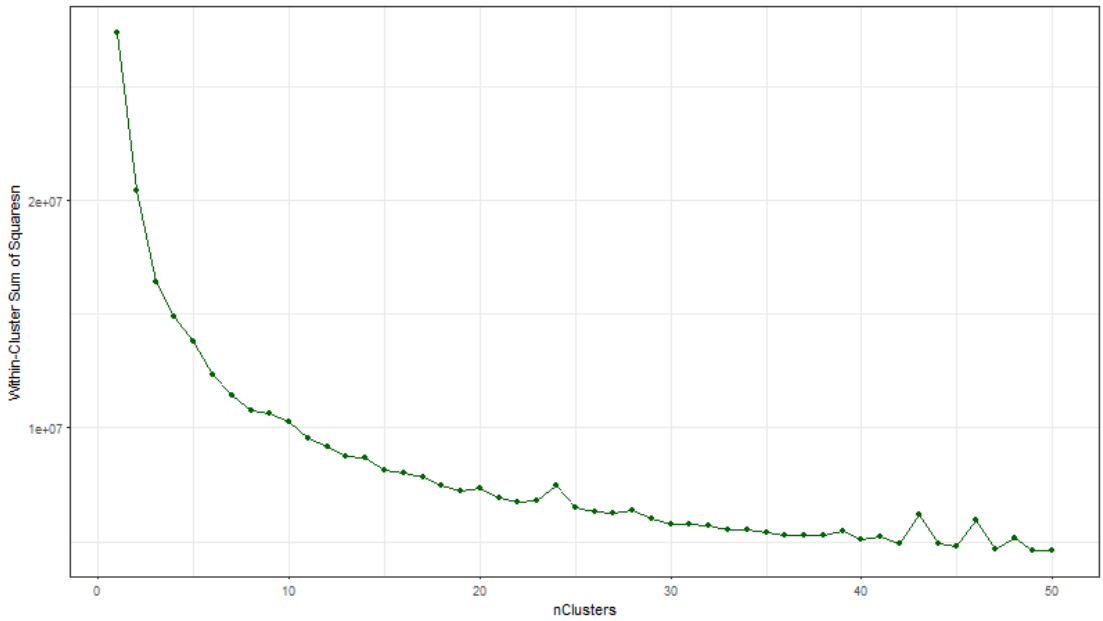


图 7Within-Cluster Sum of Squaresn 曲线图

3.2.5 访客分群定义

首先对聚类后的分群结果加上离群值访客群体进行特征分析，各类 52 个特征的中心值如下：

center_by.xls

对于每个特征，在某个分群中的值越大，数值对应的单元格颜色越红，否则值越低，单元格颜色倾向于绿色。

从 52 个特征选出一些有明显区分类别的特征，如表 9。

注意到类 1、类 8、离群值访客，在各个维度上较其他类有着显著的优势，各特征水平值较高，在表中的表现为红色单元格占据大部分。这类访客经常和官网互动，官网活跃程度较高，对官网各个模块关注度较大，同时在官网上有频繁的购票行为，消费金额高。将这类用户定义为珍贵常客。

类 2、类 4、类 5、类 6，在各个维度上数值略低于上一类珍贵常客，处于中等水平，在表中的表现，黄色单元格居多，并存在部分的红色和绿色单元格。说明这类用户，访问官网较频繁，特别关注机票、航班、旅客信息，同时也存在一定的购票行为。将这类用户定义为频访群体。

再看类 3、类 7，这两类各个维度上没有明显的优势，各个水平值较低，在表中主要以绿色单元格为主。但注意到的是，两类在访问端口上有显著差异，类 3 的移动端访问次数居多，说明这类访客对相对于类 7 移动端访问有偏好。而类 7 在 PC 端访问次数上占优势，说这类访客对类 3 在 PC 端访问相比存在一定偏好。因此，定义类 3 为移动端普通访客，类 7 为 PC 端普通访客。

各个类的特征情况归纳如下表 8 所示。

表 8 各类别特征描述和类别定义

类别	特征描述	定义类别
类 1、类 8、离群值访客	各维度水平值较高。经常和官网互动，官网活跃程度较高，对官网的各个模块关注度较大，同时在官网上有频繁的购票行为，消费金额相对其他类较高。	珍贵常客
类 2、类 4、类 5、类 6	各维度水平值一般。访问官网较频繁，特别关注机票、航班、旅客信息，同时也存在一定的购票行为。	频访群体
类 3	各维度水平值较低。和类 7 比较，该类用户在移动端访问次数有优势。	移动端普通访客
类 7	各维度水平值较低。和类 3 比较，该类用户在 PC 端访问次数有优势。	PC 端普通访客

表 9 各类特征中心值（部分特征）

输入变量名	类 1	类 2	类 3	类 4	类 5	类 6	类 7	类 8	离群值访客
观测窗口内访次数	1.7822	4.1275	1.2480	5.9893	1.3584	3.9345	1.0910	11.4093	10.8112
观测窗口内成功购票次数	1.0409	0.0237	0.0000	0.0083	0.0017	0.0001	0.0004	0.2378	0.5156
观测窗口会话总时长	50.8832	176.8984	14.8079	257.8189	33.1790	174.0496	6.1442	440.2081	1195.6506
观测窗口内消费金额	1014.3115	2.2113	0.0000	1.8048	0.1832	0.0098	0.0319	120.9777	1625.7275
机票查询访问次数	0.6521	1.3049	0.2452	3.7907	0.5207	0.9191	0.9173	3.2353	5.3496
航班选择访问次数	0.9684	1.6217	0.2091	3.4036	0.5144	0.8502	0.9463	4.4471	5.4826
航班选择总停留时间	592.1857	2073.1300	22.1774	2214.4830	204.6561	452.2012	113.8221	4397.6242	19165.0587
旅客信息总停留时间	162.0439	56.9797	5.3794	927.2705	16.5820	17.1657	3.9552	271.9908	3085.9175
付费搜索次数	0.4643	1.3192	0.0056	0.5935	1.1973	0.0938	0.0103	3.2584	1.9887
非付费搜索次数	0.0034	0.0048	0.0000	0.0001	0.0008	0.0004	0.0003	0.0183	0.0084
非会员手机登陆次数	0.0120	0.0698	0.0000	0.0015	0.0049	0.0001	0.0029	0.1606	0.1183
会员登陆次数	1.3619	1.6191	0.0380	0.4458	0.1649	0.4552	0.0604	6.3657	3.1641
会员日访问次数	0.1462	0.2497	0.1166	0.4586	0.0410	0.6109	0.0424	1.6244	0.8498
PC 端访问次数	1.6755	3.7164	0.1020	0.3829	0.7692	0.0560	1.0290	8.4170	5.4153
移动端访问次数	0.1474	0.4759	1.1467	5.6180	0.6041	3.8814	0.0660	3.1593	5.4830
折扣票总价	20.5966	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4119	0.9645
票面价总和	956.9747	0.6855	0.0000	1.1746	0.0765	0.0032	0.0114	96.7927	1439.5040
收银支付页面访次	1.0277	0.0971	0.0101	0.8988	0.0186	0.0239	0.0067	0.4885	0.8234
抽奖等营销活动页面访次	0.0196	0.0602	0.0295	0.0683	0.0088	0.2450	0.0013	0.8907	0.2176
其他页面访次	0.0328	0.2632	0.0213	0.0660	0.0297	0.0522	0.0164	0.5329	0.3542

综上，如果将 3.2.2 节中被过滤掉的低频访客群体纳入，最终将官网访客分为五大类：PC 端普通访客、珍贵常客、频访群体、移动端普通访客、低频访客。各类目中访客占比分布如图 8 所示。

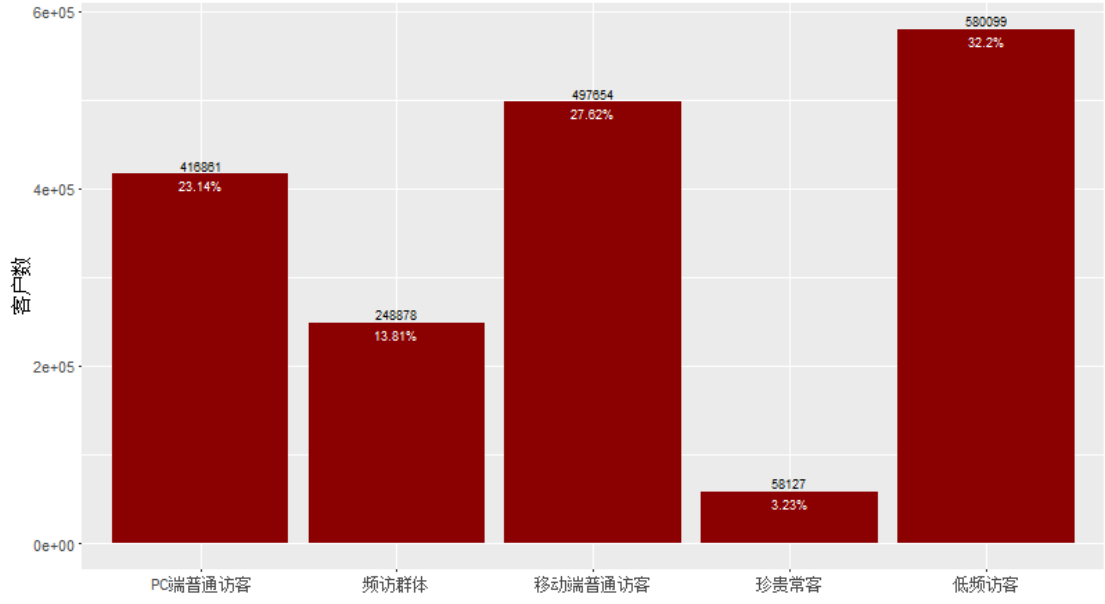


图 8 客户分布图

结合表 9，对各个访客群体的特点进行描述。

1、PC 端普通访客：这类访客在访问和购票方面并没有明显的优势，但保持一定的访问频率，在官网产生一定的操作行为，极少购票，多数情况是 PC 端访问的。

2、珍贵常客：这类访客在访问官网频率高，且在官网购票次数上非常高，购票金额也很大，属于官网的高价值活跃用户。

3、频访群体：这类用户在官网上有一定的访问量，和官网的互动次数也比较多，操作行为多样，同时也会产生一些购票行为。

4、一般访客：这类访客在访问和购票方面并没有明显的优势，但保持一定的访问频率，在官网产生一定的操作行为，极少购票，多数情况是移动端访问的。

5、低频访客：该类用户，在观测区间内访问过一次官网，且会话时间短，并无购买记录，行为暂不明确。

3.2.6 模型评估

R2 的计算公式为：R2=B/W。其中 B 表示每个类之间的差异程度总和，即每类的中心到全量数据中心的距离平方和。W 表示每条数据之间的差异程度总和，即每条记录到全量数据中心的距

离平方和。当 R2 越大，对于同一份数据集，W 是既定的，B 随着类别数目的不同而有相应变化，B 越大表示每个类之间的差异就越大，而每个类中的差异则越小，也就说明 k 个类分得越开。因此 R2 统计量可用于评价合并成 k 个类时的聚类效果，R2 值越大，聚类模型效果就越好。

R2 的取值范围在 0 与 1 之间，它大体上是随着分类个数的减少而变小。

经过多次对模型进行训练，得到 R2 值为 0.606，效果较好。

3.2.7 建议和说明

访客行为分群包含规则分群和聚类分群两个部分，之后又根据群特征进行合并，最终分为 5 类。其中低频访客群体和离群访客群体属于规则划分，低频访客群体可以考虑不需要进行之后的访客流失分析，原因是：低频访客群体中数据几乎没有访问行为信息，无法根据其访问行为来估算流失和购票或留存概率。

在进行规则划分把低频访客群体和离群访客群体分离出来后，对于其余的访客通过 k-means 聚类得到的 8 个分类，根据类特征，将这些分类进行合并，关于划分的示意图如图 9 所示。

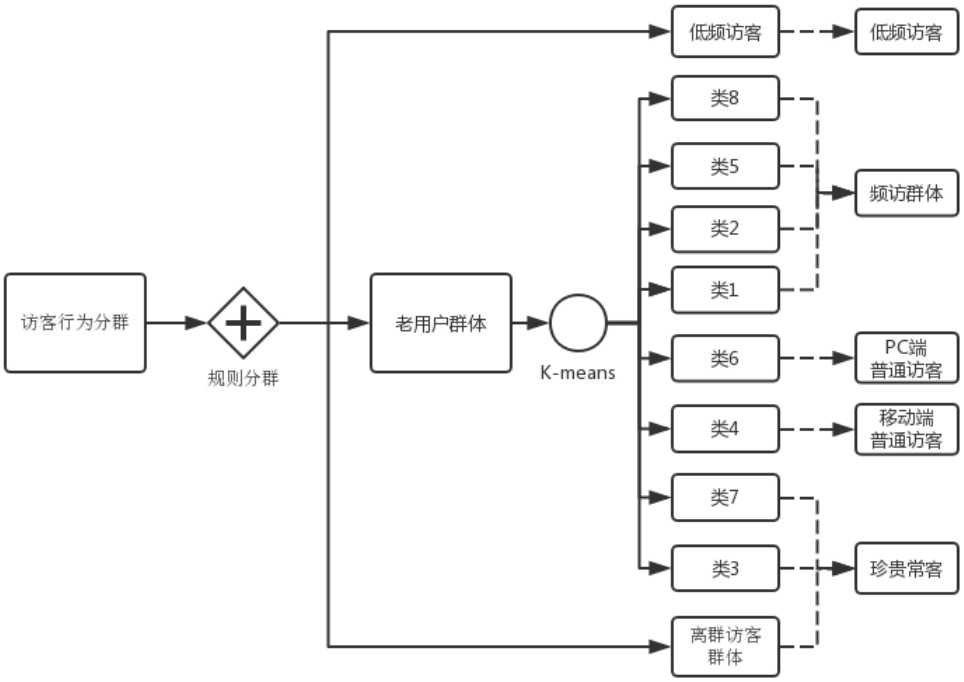


图 9 访客行为分群示意图

在每个客户分群中，会员和非会员的分布占比如图 10 所示，其中 1 表示会员，0 表示非会员。注意到，频访群体和珍贵常客群体的会员人数多于非会员人数。而其他分类，包括低频访客、PC 端普通访客、移动端普通访客，非会员人数占主要部分，同时存在少量的非会员用户。

此外，在所有用户中，付费搜索次数达到 346327 访次，站所有搜索访次数的 99.74%。非付费访次数为 887 次，占 0.26%。如图 11 所示。

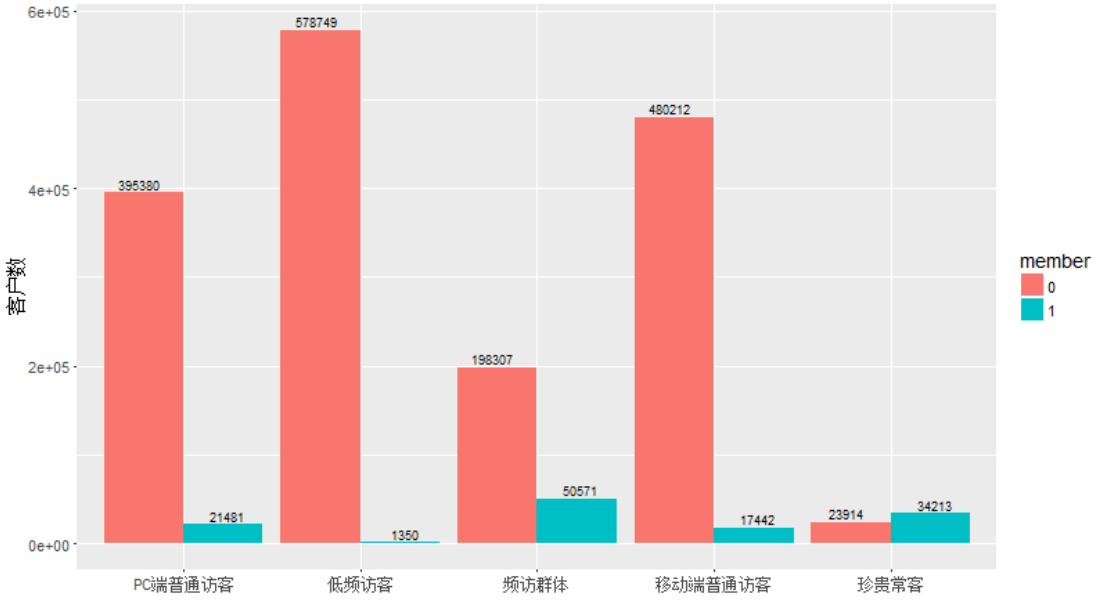


图 10 各分群会员数量分布（0 为非会员，1 为会员）

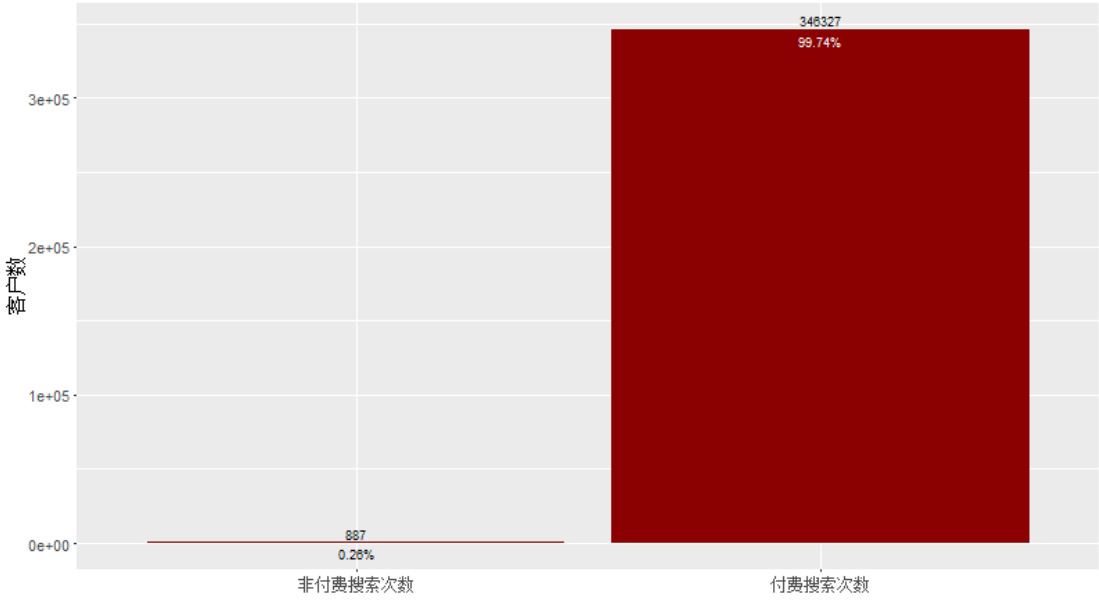


图 11 付费搜索和非付费搜索次数分布图

还需要说明的一点，截止目前全量数据约有 2000 万条数据，将这些数据划分 11 份，每份 6 个数据文件（测试环境下共有 66 个数据文件），分别训练聚类模型。其 PCA 的降维数、聚类最优数以及各类分布情况如下表所示。

表 10 分组聚类结果汇总

组别	PCA 降维数	k-means 最优分类数	PC 端普通访客占比	珍贵常客占比	频访群体占比	移动端普通访客占比	低频访客占比
1	7	8	23.30%	3.22%	13.31%	27.94%	32.22%
2	7	8	23.32%	3.37%	13.78%	27.38%	32.15%
3	7	8	23.49%	3.38%	14.12%	26.74%	32.26%
4	7	8	23.53%	3.25%	14.18%	26.78%	32.26%
5	7	7	2.58%	1.80%	49.55%	2.94%	43.14%
6	7	8	23.33%	3.33%	13.63%	27.50%	32.22%
7	7	8	23.34%	3.22%	13.52%	27.69%	32.23%
8	7	8	23.54%	3.29%	14.32%	26.60%	32.26%
9	7	8	22.87%	3.31%	13.79%	27.62%	32.41%
10	7	8	29.74%	3.04%	7.28%	27.83%	32.11%
11	7	8	23.88%	3.16%	14.09%	26.70%	32.17%

11 组的聚类模型训练结果具有以下几个共同点：

1）PCA 降维自动都被降到 7 维

2）k-means 最优分类数除了第 5 组外，其余都自动定位 8 类

3）各个分组的占比总体上较稳定

以上结果表明，对于不同划分的训练集，得到的客户分群模型结果趋于稳定。在后续上线过程中，可以考虑将 PCA 降维数和 k-means 最优分类数进行固定，以减少模型系统的复杂度。

3.2.8 新数据验证

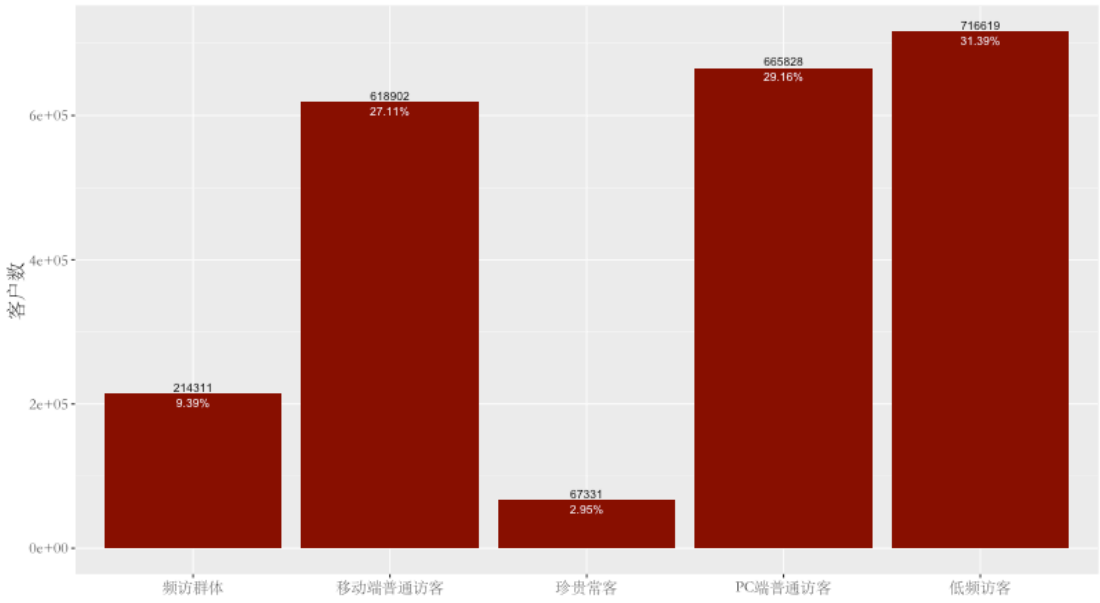


图 12 客户分布图（新数据验证）

引入 2017 年 2 月-2017 年 4 月为观测时间窗口的数据，重新训练聚类模型，得到最优降维数为 7，最优聚类数为 8，R 方等于 0.621，新数据下的聚类模型的客户分布图如图 12 所示。可以发现，和 2017 年 1 月-2017 年 3 月为观测时间窗口相比，聚类模型的客户分布情况差异很小，聚类模型趋于稳定。

3.3 访客流失分析

3.3.1 建模流程

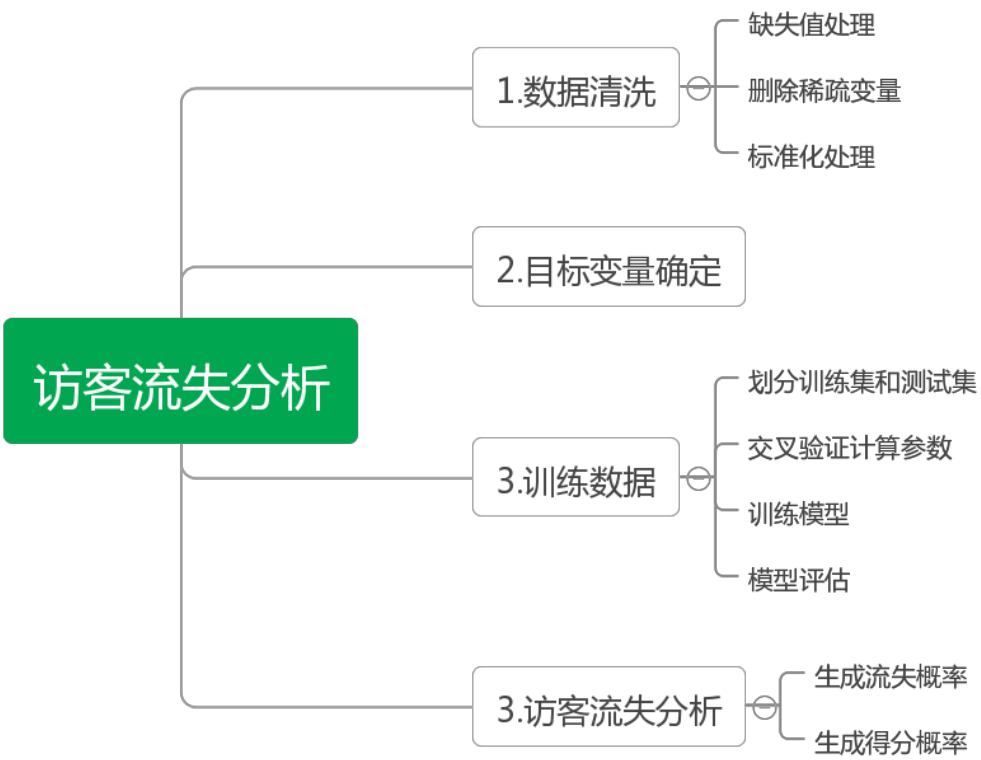


图 13 访客流失分析建模流程

3.3.2 数据清洗

导入数据量大小为 1801629 条。

1、剔除异常 id 用户：对于客户号异常的用户，删去记录；

2、剔除逻辑异常数据：对于一些特征存在逻辑上错误，如乘机和购票总间隔时间出现负值，应删除相应记录，此步删去 10 条记录；

3、缺失值处理：最近一次购票时间到观察结束时间的天数、乘机和购票总间隔时间存在部分缺失值，处理方法如下：1）最近一次购票时间到观察结束时间的天数：用观测窗口长度+1 进行插补；2）乘机和购票总间隔时间：0 值插补；

4、剔除稀疏输入变量：对每个输入变量进行遍历，剔除输入变量值均为常数的输入变量；

5、目标变量设定：验证窗口为2017.4.1-2017.4.30期间，观测区间为2017.1.1-2017.3.31。针对不同的分群场景有以下两种流失模型的定义：

1）珍贵常客场景：验证窗口范围内，访客购票次数占访客访问次数的比例大于 0，定义为非流失（有购票），目标值记为 1；比例等于 0，定义为流失（未购票），目标值记为 0。（转化率角度）

2）频访群体、PC 端普通访客、移动端普通访客场景：验证窗口范围内，访客访次数大于 0，且观测窗口内最后一个月内，访客访次数大于 0，定义为非流失（留存访客），记为 1；否则，定义为流失（非留存访客），记为 0。（留存率角度）

6、按客户分群结果分别从数据集单独分离出来，除掉低频访客群体，共分为 4 大类分别进行训练；

7、划分训练集和测试集：按 3 比 2 的比例将各分群数据划分成训练集和测试集。

8、平衡数据处理：对数据进行欠抽样，即从多数类样本中抽取少数类样本等量的样本数和少数类样本合成训练数据。

3.3.3 数据分布特征

输入变量的分布特征同表 5。目标值的在各分群中的分布情况如图 14 所示。其中，PC 端普通访客、频访群体、移动端普通访客的目标值为留存指标（1 表示留存，0 表示流失）。珍贵常客群体的目标值为购票转化率指标（1 表示有购票，0 表示流失）。

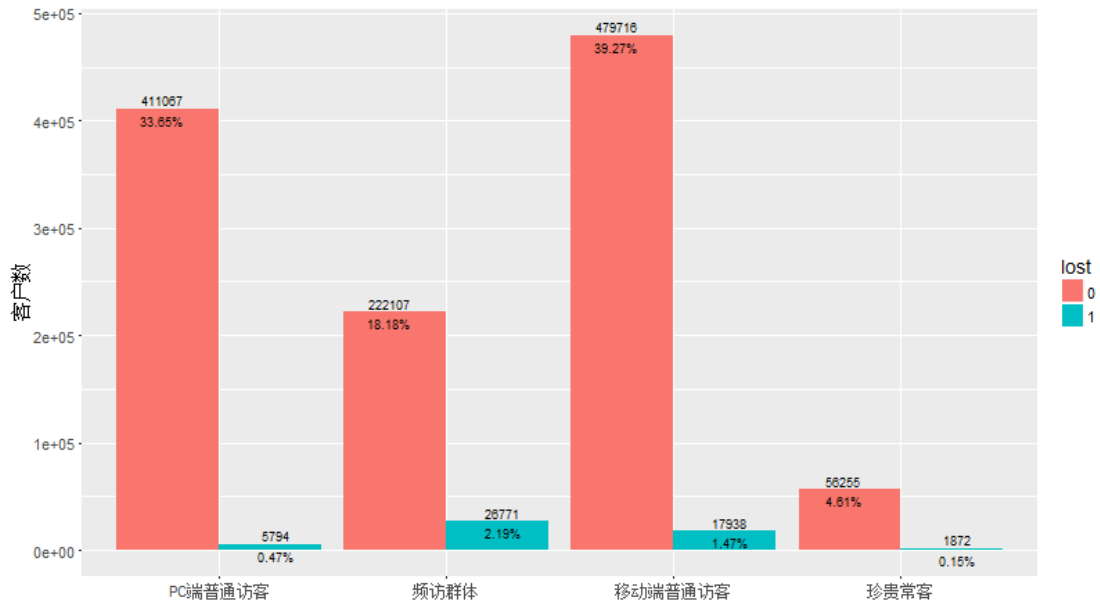


图 14 各客户分群实际流失情况分布

3.3.4 参数确定与模型建立

由于训练集中的数据为非平衡数据，主要表现在目标变量等于 0(负样本)的数据占比居多，因此从目标变量等于 0 的记录中抽取和目标变量等于 1 的记录等量的数据，并和目标变量等于 1 的记录一起作为最终的训练集训练模型。关于各分群特征的数据量情况见下表所示。

表 11 各分群特征的数据量情况

分群类别	总合计	处理前训练集数量	平衡处理后训练集数量	测试集数量
珍贵常客	58127	35064	2254	23063
频访群体	248878	149340	32162	99638
PC 端普通访客	416861	250084	6992	166777
移动端普通访客	497654	298653	199001	21572

接下来开始对每个分群训练 lasso+logit 回归，其中含有超参数 lambda，使用交叉验证法对 lambda 值进行估计，fold 值取默认值 10，并采用 leave-one-out 法估计。经过训练得到各个分群训练模型参数 lambda.min 值的估计值，如下表所示。

表 12lasso+logit 回归的交叉验证结果

分群类别	Lambda.min
珍贵常客	0.00126
频访群体	0.00079
PC 端普通访客	0.00133

移动端普通访客	0.00095

之后依据上表结果分别训练四个分群的流失模型，得到训练模型参数后，将每个分群的测试集应用于训练模型中，对模型的效果进行评估。

3.3.5 模型评估

各分群中的 lasso 逻辑回归的系数如下：


coef_珍贵常客_lasso.csv


coef_频访群体_lasso.csv


coef_移动端普通访客_lasso.csv


coef_PC端普通访客_lasso.csv

导入各个分群的测试集对模型的泛化能力进行评估，得出 4 个分群下的混淆矩阵和基于混淆矩阵计算的评估指标汇总如下。

表 13 各分群特征的数据量情况

珍贵常客		实际值		频访群体	实际值	
		0	1		0	1
预测值	0	17788	4530	0	71701	17247
	1	163	582	1	757	9933
移动端普通访客		0	1	PC 端普通访客	0	1
预测值	0	136224	55625	0	130992	33487
	1	979	6173	1	307	1991

从表 13 中计算得到的相关评估指标汇总表如下。

表 14 各分混淆矩阵相关指标汇总表

指标名称	珍贵访客	频访群体	移动端普通访客	PC 端普通访客
Sensitivity	0.990920	0.98955	0.992865	0.997662
Specificity	0.113850	0.36545	0.099890	0.056119
Pos Pred Value	0.797025	0.80610	0.710058	0.796406
Neg Pred Value	0.781208	0.92919	0.863115	0.866406
Precision	0.797025	0.80610	0.710058	0.796406
Recall	0.990920	0.98955	0.992865	0.997662
F1	0.883459	0.88846	0.827979	0.885745
Prevalence	0.778346	0.72721	0.689459	0.787273
Detection Rate	0.771279	0.71962	0.684539	0.785432
Detection Prevalence	0.967697	0.89271	0.964060	0.986221
Balanced Accuracy	0.552385	0.67750	0.546377	0.526891

关于上表中各指标解释，假设混淆矩阵如下所示：

		真实值	
		0	1
预测值	0	A	B
	1	C	D

那么，上表中各个指标的公式如下：

Sensitivity = A/(A+C)

Specificity = D/(B+D)

Pos Pred Value = (sensitivity * prevalence)/((sensitivity*prevalence) + ((1-specificity)*(1-prevalence)))

Neg Pred Value = (specificity * (1-prevalence))/(((1-sensitivity)*prevalence) + ((specificity)*(1-prevalence)))

Precision = A/(A+B)

Recall = A/(A+C)

F1 = 2*precision*recall/(precision+recall)

Prevalence = (A+C)/(A+B+C+D)

Detection Rate = A/(A+B+C+D)

Detection Prevalence = (A+B)/(A+B+C+D)

Balanced Accuracy = (sensitivity+specificity)/2

模型评估主要包括以下几个指标：ROC 曲线、AUC 值、前 10 百分位的提升度、预测准确度。各分群模型的 ROC 曲线如图 15-图 18 所示，其余评估指标包括 AUC 值、前 10 百分位的提升度、预测准确度，其结果见表 15。

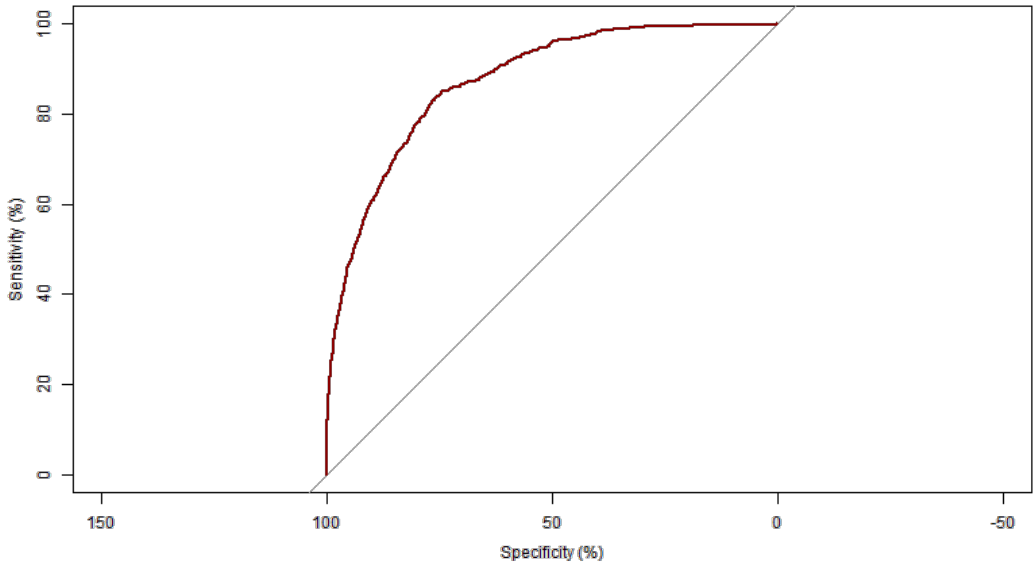


图 15 珍贵常客流失模型 ROC 曲线

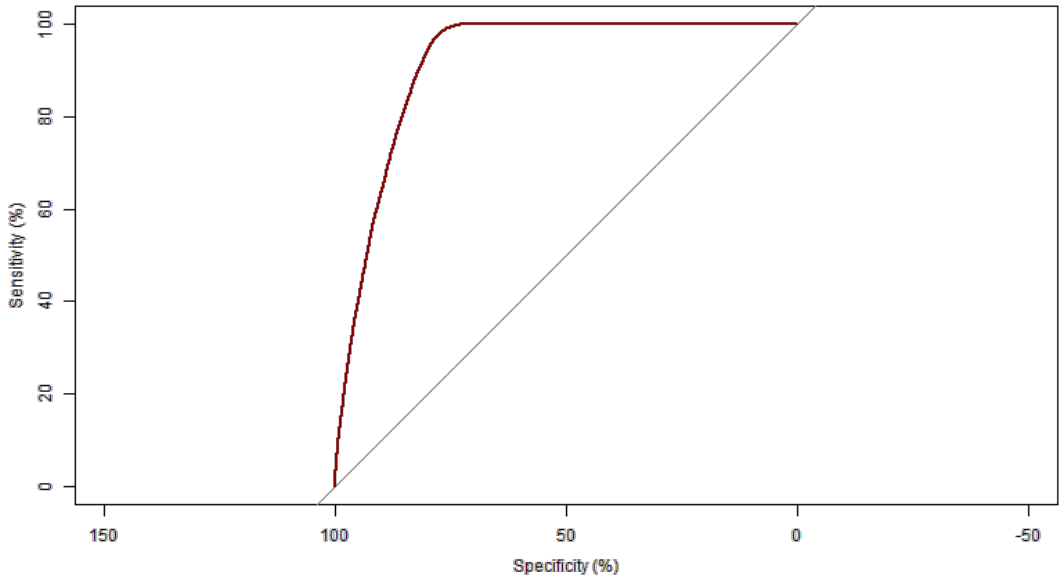


图 16 频访群体流失模型 ROC 曲线

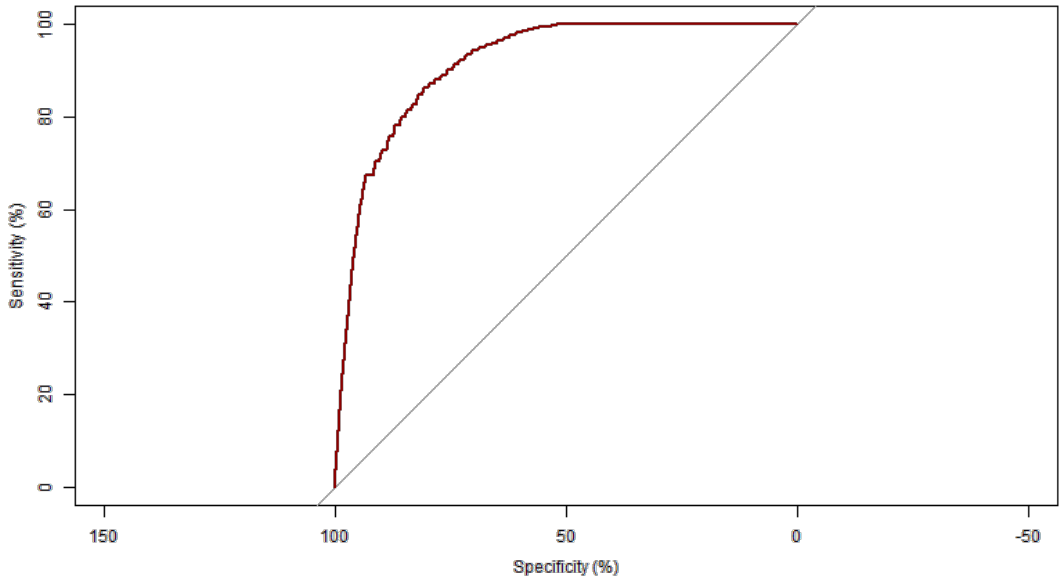


图 17PC 端普通访客流失模型 ROC 曲线

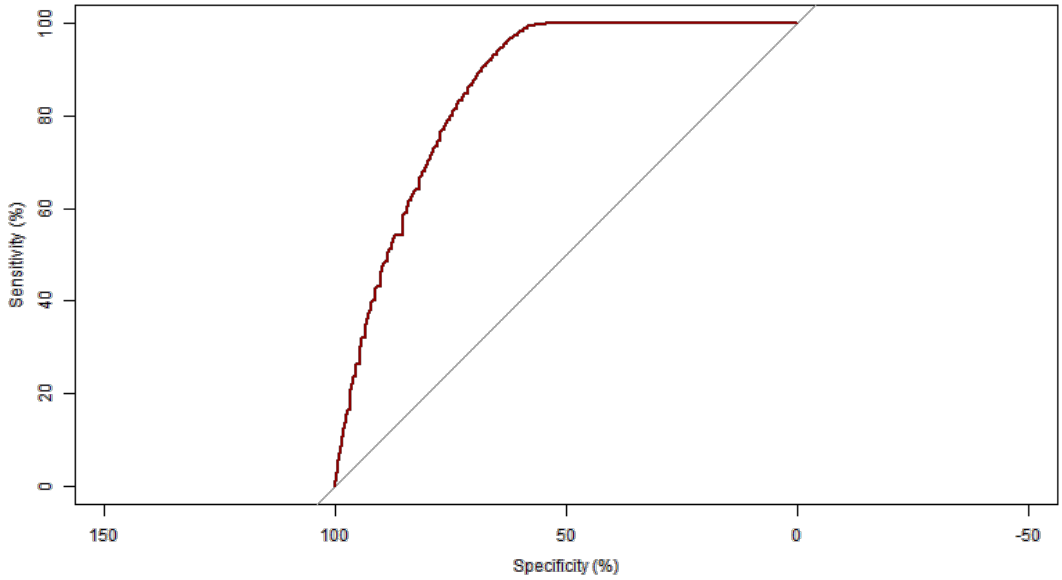


图 18 移动端普通访客流失模型 ROC 曲线

表 15lasso+logit 回归的模型评估汇总

cluster_type	lambda	AUC	Lift_10	precise
珍贵常客	0.00126	87.28539	5.77401	0.79651
频访群体	0.00079	91.70131	5.02008	0.81931
PC 端普通访客	0.00133	91.51142	11.93180	0.79737
移动端普通访客	0.00095	85.80244	5.35729	0.71556

上表中除了移动端普通访客的预测准确度为 71.56%外，其他模型均处在 75%以上合理的范围中。此外，AUC 值均在 85%以上，前 10%分位数提升度均在 5 倍以上。最后重新对各分群的所有数据集进行计算，生成相应的流失概率（目标值等于 0 的概率）和购票/留存概率（目标值等于 1 的概率）名单（其中低频访客群体无购票/留存概率和流失概率数值，用 NA 表示）。

3.3.6 新数据验证

导入 2017 年 2 月-2017 年 4 月的数据，对下一期的数据进验证。模型基于 1 月-3 月作为观测时间窗口进行训练，预测输入变量时间窗口为 2017 年 2 月-2017 年 4 月，输出变量时间窗口为 2017 年 5 月，预测结果和真实值如下表 16 所示。在珍贵常客、频访群体、PC 端普通访客、移动端普通访客四个分群中预测准确度分别为 79.70%、77.89%、76.83%、77.51%。

表 16 各分群特征的数据量情况（新数据验证）

珍贵常客		实际值		频访群体	实际值	
		0	1		0	1
预测值	0	52160	13168	0	143058	45676
	1	498	1505	1	1718	23859

移动端普通访客		0	1	PC 端普通访客	0	1
预测值	0	459603	135761	0	506221	153492
	1	3404	20134	1	797	5318

第四章相关文档附件

4.1 数据挖掘模型数据预处理脚本

见附件：dataclean20170503.sql

4.2 数据挖掘模型脚本

4.2.1 价值模型（全量数据脚本）

见附件：RFM_1_ModelSetup

4.2.2 聚类模型（全量数据脚本）

见附件：UserCluster_1_DataClean&ModelSetup.R

4.2.3 流失模型（全量数据脚本）

见附件：LostModel&PurchaseModel.R

4.3 客户价值、分群、流失概率购票概率名单（部分名单）

4.3.1 合并名单脚本

见附件：Output_Merge.R

4.3.2 最终名单

见附件：final_list.csv