基于用户使用偏好的手机应用推荐策略

苏锦华、刘洁、高聪

**摘要：**本文旨在通过对用户使用偏好进行数据分析，指导基础推荐系统的精细化推荐，构筑一套基于用户使用偏好的手机应用推荐策略。本文主要从用户角度分析智能手机用户监测数据，用户使用应用时间、时长、和应用类别等数据，对用户通过K均值方法聚类，定义用户类别。除了通过基础分组以及序列描述分析用户使用行为的时序特征，本文还运用分布式技术，用logit回归、随机森林和支持向量机等方法分析用户行为时序特征，并依据用户使用手机应用偏好，构建初步应用推荐系统，构建了Recommenderlab包推荐系统以及利用分布式实现ALS推荐系统系统，以此对聚类得到的用户群体进行推荐，并依据分析和推荐系统总结应用推荐策略。

**展示后后续进展：**由于小组展示使用分布式计算，展示内容比较朴素，后续通过讨论确定主题后进行了大量分组数据处理以及相关绘图（本文超过一半的图片为后续绘制）。

1.场景数据简介

* 1. 场景与数据简介

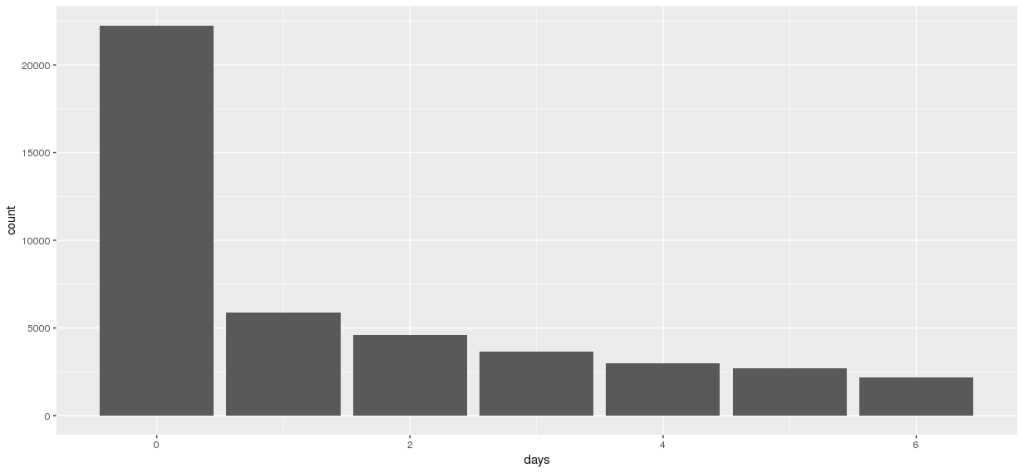
全部数据是某年连续30天的4万多智能手机用户的监测数据，记录了用户使用各款APP的起始时间、使用时长、上下行流量等，app\_class.csv数据集记录了APP所属类别，从用户和应用角度进行分析，并依据用户使用偏好构建手机应用推荐系统与策略。

* 1. 初识数据

由于时间以及能力有限，在单机读取全部数据时，时间过长，在简单描述性统计时以七天数据为代表，但是在其他模块均使用全体数据进行分析。对七天数据进行简单的总结及观察，得到以下需要注意的点。

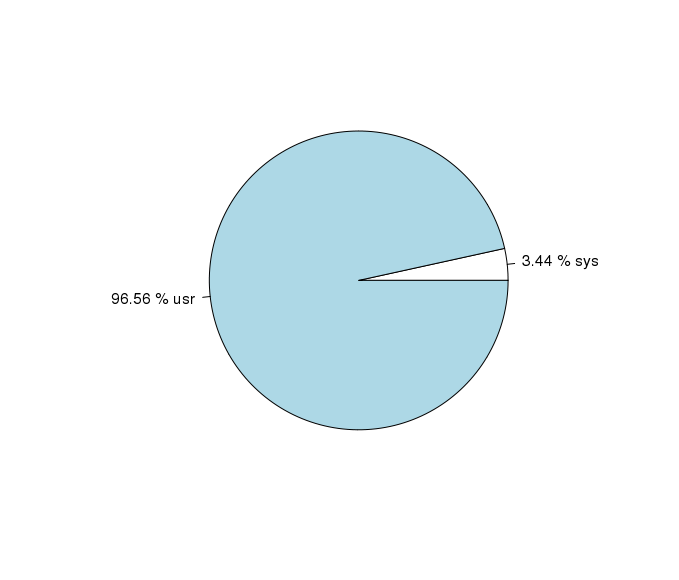
1. 七天数据文件长度分别为：5860124，5756942，5716771，5065923，4767333，4805679，4047786，每天数据长度不等
2. 七天被检测到的用户有44155名，有些天可能某些用户并没有监测到，因此需要注意。
3. 所涉及到应用个数为3694个
4. 变量app\_type的数值中，除了sys,usr之外，还有用户与预装，涉及到该变量时要注意进行处理。
5. 变量duration中，最大值为1.428e+09，是明显的异常值。
   1. 简单描述性统计
6. 缺失数据

有的用户在某一天没有任何应用的监测信息，说明是该用户在该天没有被监测到，对每个用户没有被监测的天数进行统计，得到七天用户监测数据缺失的天数频数分布直方图图，随着天数的增加，人数减少，没有被监测天数为0的用户占大部分，说明大部分用户没有缺失，少部分用户有缺失。



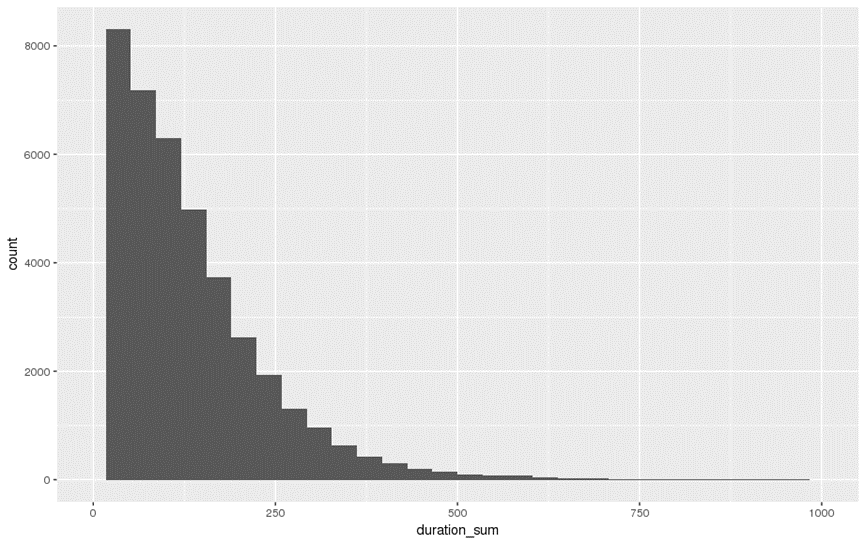
1. 系统与用户安装应用

统计所有涉及到的应用个数，得到其中有127个是系统安装应用，剩下3567个应用都是用户安装应用，下图为七天数据涉及到软件中，系统软件与用户安装软件占比饼图，可看出用户安装软件占大多数。



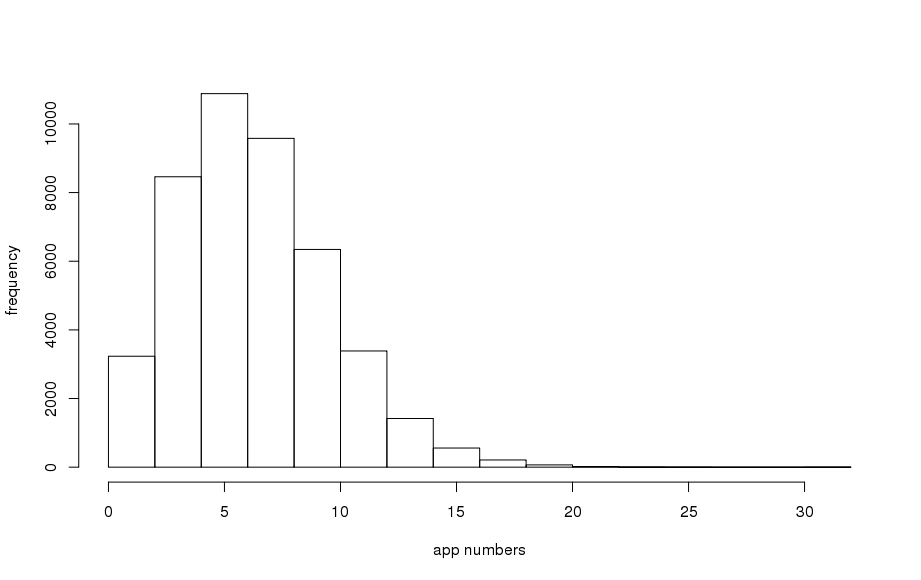
1. 每位用户每天平均使用APP时长（分钟）

在分析过程中，得到使用应用时长大多集中在1000分钟以内，但是有少数在1000之外，最大值为680万分钟左右，这和一开始介绍数据时提到的start\_day有异常值有关，例如在每天的数据集中，start\_day最小值均为-16524。因此作频数分布直方图时，限定x轴范围在1000以内。此外，由于有些用户有的天数并没有监测到，因此取平均值时，除以其在这七天内监测到的天数。由下图可知，每位用户每天平均使用APP时长频数分布直方图呈现右偏分布，大多数用户平均每天使用应用的时间在500分钟以内。



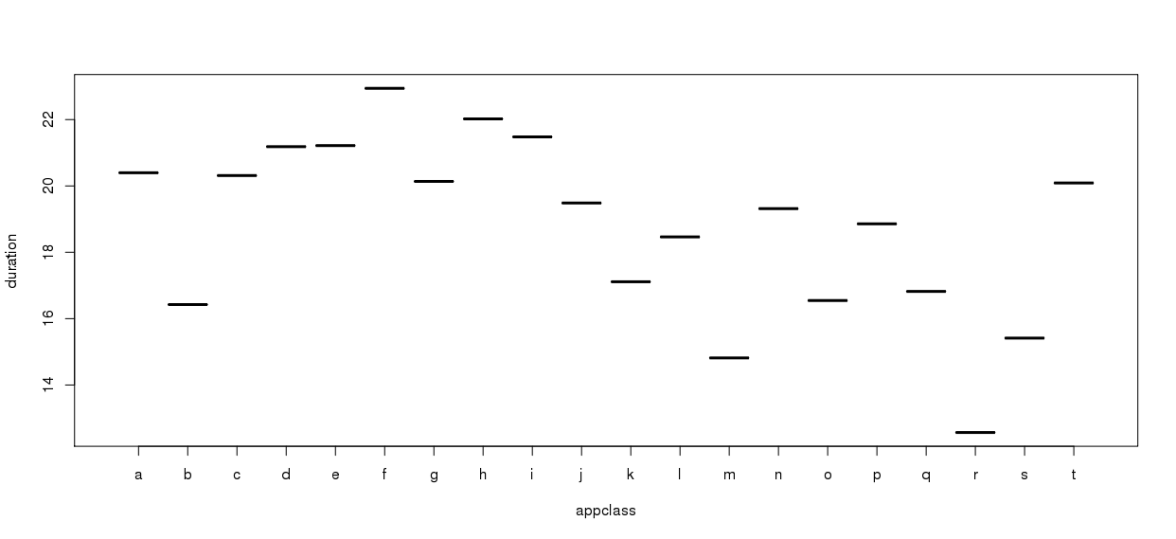
1. 每位用户每天平均使用APP个数频数分布直方图

每位用户在被监测到的每天内平均使用APP的个数如下图所示，由于一天中用户可能会多次使用某些应用，因此在处理时，首先要得出每天所使用不同应用的个数，求和，再除以有效监测天数，求其平均值，得到下图。由下图可知，每位用户每天平均使用APP个数大多在20以内，大多数用户每天平均使用应用个数在5个左右。



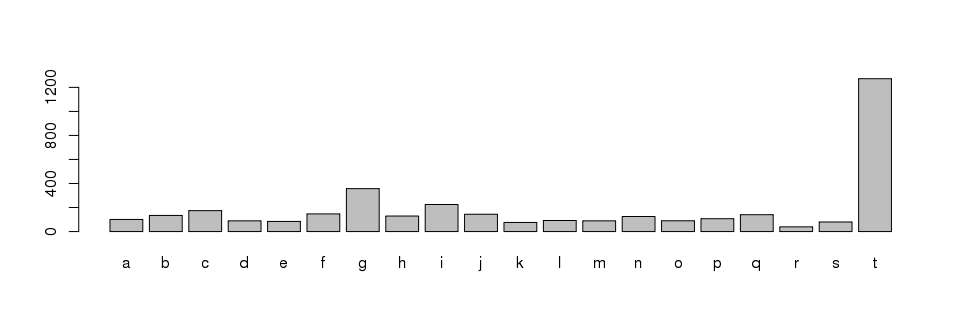
1. 各类APP使用总时长

为判断某类APP是否会被用户频繁使用，得到每类APP被使用总时长，由于数据中所用时长单位为秒，在此分析中将时长取对数，得到每类应用在这七天中被使用的的总时长，作出下图，由图可知f、h、i类应用使用的总时长高于其他类应用，其中f类应用被使用时长最大，r类应用使用总时长最短。



1. 各类软件中的软件个数

对软件类别中软件个数进行分析，得到七天数据涉及到的各类软件中的软件个数，如下图所示，可知各软件中t类软件最多，多于1000，其余类别软件个数均在500以下。



2.用户行为分类

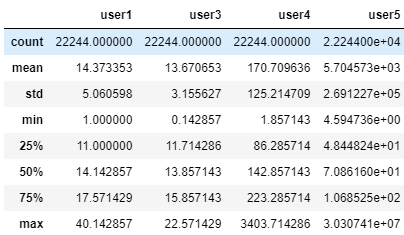
**2.1用户行为特征**

我们选取描述统计分析中的五个变量，分别为“平均每天使用APP个数”、“使用时长最长的APP种类”、“使用APP总时长最长的时间段”、“平均每天使用APP次数”、“平均每天使用APP时长”，来描述用户的行为特征，便于对用户进行分类。表2-1展示了五个变量的详细说明，表2-2展示了四个连续变量简单的分布特征。

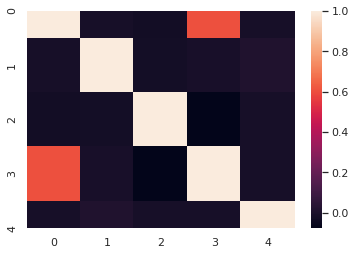
表2-1 五个变量说明

|  |  |  |  |
| --- | --- | --- | --- |
| **变量编号** | **变量含义** | **变量类型** | **变量说明** |
| 0 | 30天内平均每天使用APP个数 | 连续变量 |  |
| 1 | 30天内使用时长最长的APP种类 | 分类变量 |  |
| 2 | 30天内使用APP总时长最长的时间段 | 连续变量 | 1-24小时 |
| 3 | 30天内平均每天使用APP次数 | 连续变量 |  |
| 4 | 30天内平均每天使用APP时长 | 连续变量 | 单位：秒 |

表2-2 四个连续变量的简单分布特征



对五个变量进行相关性分析，图2-1通过颜色深浅反映了五个变量之间的相关程度，颜色越浅，代表相关程度越高。由图可知，总体来说，五个变量之间的线性相关关系不是很显著，其中仅“平均每天使用APP个数”与“平均每天使用APP次数”存在一定的相关性，但对于根据这五个变量进行聚类分析影响不大。



**图2-1 各个变量之间的相关程度**

**2-2. K均值聚类**

对五个变量进行对数化处理后，根据五个变量进行K均值聚类。其中，经过层次分类得到k值选取为3或4 时，误差平方和局部最小，因此我们将k值选取为4。四类中心点的值如表2-3所示，聚类结果如图2-2所示。

**表2-3 四类用户的中心点的值**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **平均每天使用APP个数** | **使用时长最长的APP种类** | **使用APP总时长最长的时间段** | **平均每天使用APP次数** | **平均每天使用APP总时长** |
| 聚类中心1 | -0.17846413 | 1.475509 | 0.037949 | -0.255078 | 0.008575 |
| 聚类中心2 | 1.056346 | -0.317623 | -0.217414 | 1.095905 | -0.019539 |
| 聚类中心3 | -0.493431 | -0.597071 | 0.101085 | -0.475190 | -0.011254 |
| 聚类中心4 | -0.567802 | 0.204807 | -0.778426 | -0.892165 | 94.191807 |

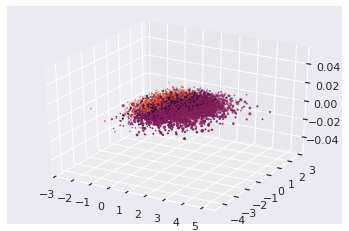


图2-2 聚类结果

其中第一类用户数为5608，约占总用户数的25.3%；第二类用户数5941，约占总用户数的26.7%；三类用户数为10693，约占总用户数的48.1%；第四类用户数为2，远少于其他三类用户数，可初步判断为异常值，同时，该类用户聚类中心的“平均每天使用APP总时长”变量值的94.191807，表现为明显的异常，因为将第四类用户判定为异常用户。

**2.3 定义用户类别**

对四类用户的聚类中心在四个连续变量的值进行数据处理后，分别观察前七天的数据变化情况，图2-3至图2-6分别反映了四类用户的聚类中心在“平均每天使用APP个数”、“使用APP总时长最长的时间段”、“平均每天使用APP次数”、“平均每天使用APP总时长”四个连续变量上的值。

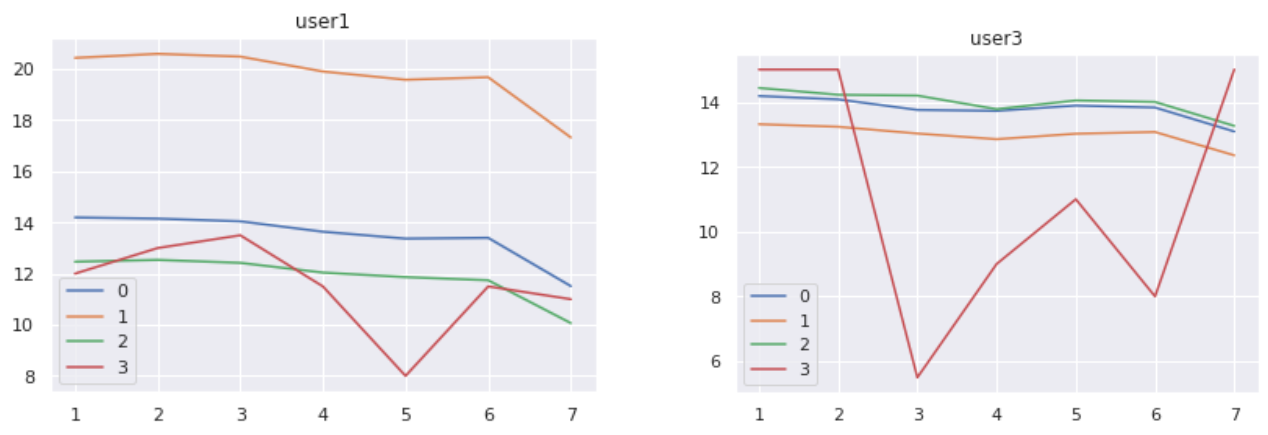


图2-3 图2-4

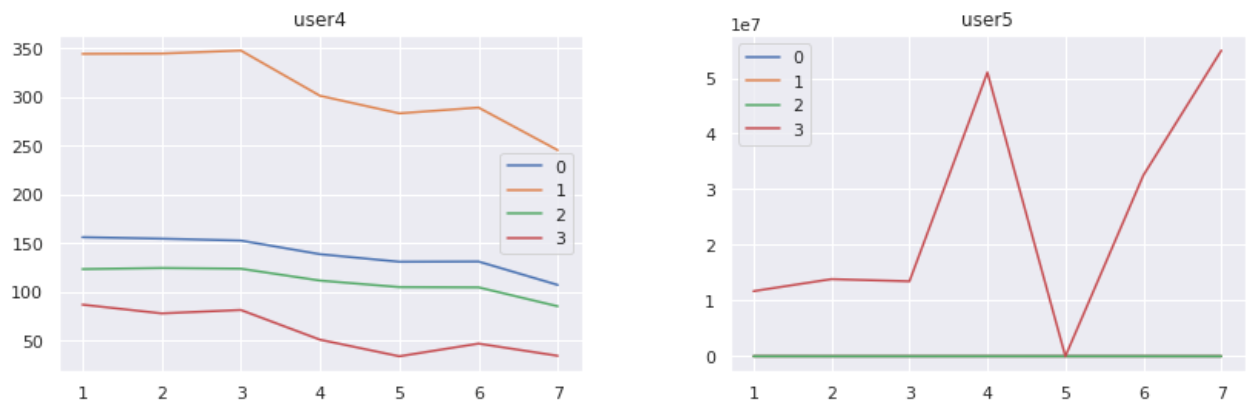


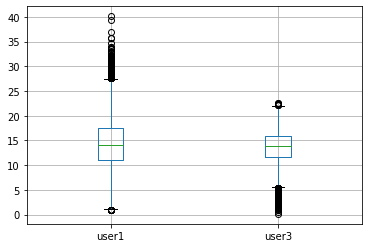
图2-5 图2-6

由图可知，前三类用户使用APP时长最长的时间段均集中在12时-14时，无明显差别。其中，第一类用户平均每天使用APP个数和次数都远大于第二类和第三类用户，因此认为第一类用户为“手机依赖症”重度患者；第二类用户平均每天使用APP个数和次数较多，认为第二类用户仅在特定时间段或特定情况下内使用手机；第三类用户平均每天使用APP个数和次数又低于第二类用户，认为第三类用户为仅在有需要时使用手机，通过手机满足需求的用户群体。而第四类用户，其聚类中心四个连续变量的值均表现出异常，因此认为第四类用户为数据记录异常用户。

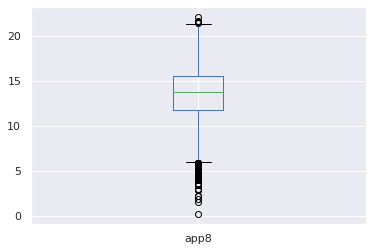
1. 用户行为时序特征

**3.1 使用时段集中特征**

由下图可知，用户使用app时长最长的时段分布在中午两点左右，0-6点使用人数较少，但存在24时段使用手机app最多的人。

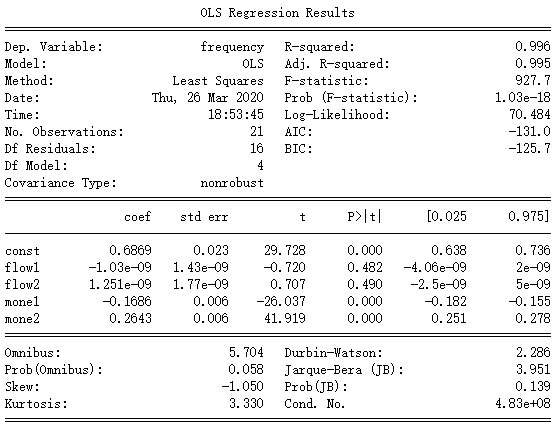


上图以用户视角描述了用户使用时长的分布，下图从应用被使用的视角分析了应用被使用时长最长的时间段的分布，发现虽然最活跃的时段依旧是中午两点，但是没有app在午夜24时被使用最多，也就是说虽然存在夜猫子个体，但是一个应用的所有受众不可能都是夜猫个体，所以最热门时段自然不存在大于22时的值。



**3.2影响使用频率的重要因子**

进一步探究用户使用强度的差异，将上下行平均流量和近期和长期两种使用强度对用户使用频率进行拟合。由下图可知，用户使用流量的多少并不是影响使用频率的重要因素，近期和长期的强度都是显著因子，而近期、长期强度拟合系数分别为负和正可以发现近期过度使用反而会负向影响使用频率。这种“过犹不及”的现象引起了我们的注意，这种情况可能会解释一些过度营销反而影响用户留存。



**3.2用户留存时序特征**

基于3.2得出的结论我们将以用户留存的视角对这一现象进行分析，我们构建以下自变量和因变量，自变量的选取主要借鉴了商业分析中RFM的概念，其中在强度这一特征上构建了多个不同时段的变量，基于这些变量的回归结果我们可以分析用户留存与不同期使用强度的联系。



图2-2 预测任务变量解释

我们采取了多类预测模型进行预测，发现Logit回归与随机森林和支持向量机的差距并不大，说明线性模型能较好地拟合自变量与因变量。Logit模型训练时间短，同时系数具有一定解释力，所以下面针对其拟合系数做进一步分析。

表2-2 模型预测效果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 预测准确率 | PR曲线面积 | ROC曲线面积 |
| Logit回归 | 0.71 | 0.83 | 0.71 |
| 随机森林 | 0.82 | 0.85 | 0.83 |
| 支持向量机 | 0.76 | 0.83 | 0.79 |

表2-2 Logit模型拟合系数

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 系数估计值 | -6.6 | -851 | 0.22 | -7.4 | 7531 | 2630 | 4009 | 6208 |

结果显示使用间隔越大，用户越可能流失，使用频次越大，用户越可能继续活跃。在不同的使用强度中，近期使用强度越大，用户反而越可能流失，长期使用强度越大，用户越不可能流失。这说明如果通过过度营销仅仅刺激用户的短期使用需求，而不注重用户长期使用习惯的话有可能造成用户的厌恶心理导致用户流失。

1. 应用使用偏好

不同应用满足用户的不同需求，不同类别的用户有不同的需求，自然对不同类别的应用有使用偏好上的差异。通过分析不同应用的特征差异，进一步交叉分析用户类别与应用类别的关联。

**4.1应用特征**

表2-1 8个变量说明

|  |  |  |
| --- | --- | --- |
| **变量编号** | **变量说明** | **变量类型** |
| **1** | **APP平均使用频度** | **连续变量** |
| **2** | **APP平均使用强度** | **连续变量** |
| **4** | **APP平均使用流量** | **连续变量** |
| **6** | **7天内使用用户数量** | **连续变量** |
| **7** | **APP所属类别下同类APP数量** | **连续变量** |
| **8** | **7天内APP被使用时长最多的时间段(1-24)** | **分类变量** |

类似于上文中针对每一位用户构造的重要变量，构造从每一款app为出发点的重要变量。由于构造上述变量时产生了较多高度相关的变量，所以仅保留6个变量作为我们的重要研究变量。

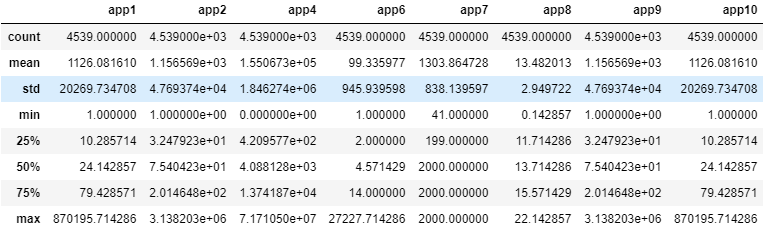


图2-2 用户变量描述性统计

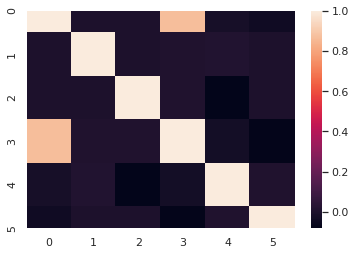
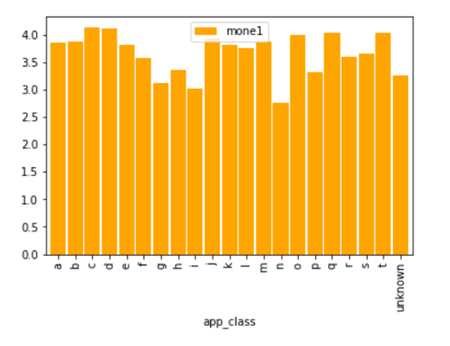
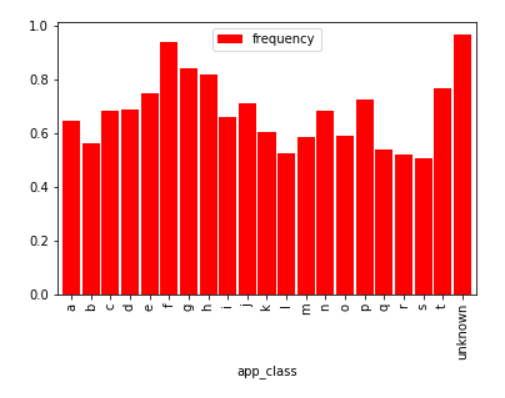
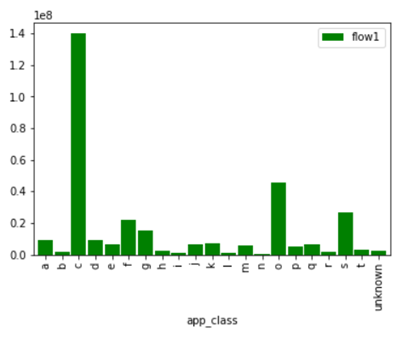
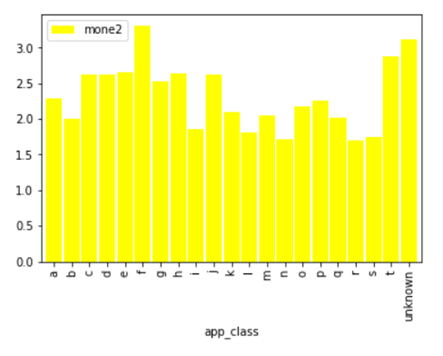


图2-2 变量间的相关性热力图

APP平均使用频度和7天内使用用户数量呈高度相关，说明越多用户使用的app和用户频繁使用的app往往高度重合。

**4.2应用类别差异**





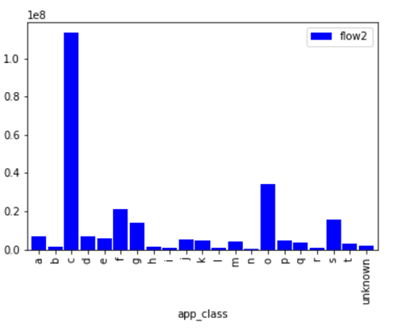


图2-2 不同变量在app中的差异

由上图可知，不同app使用频率和长期使用强度呈现高度相似，但短期使用强度则不存在比较明显的类别差异。而上下行流量不同app差异巨大，但上行和下行流量常常高度线性比例。

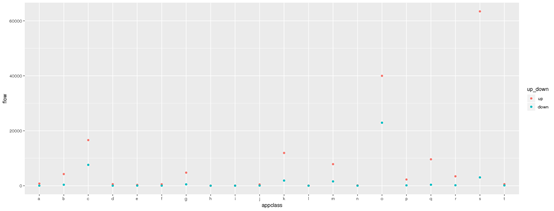


图2-2 上下行流量在app中的差异

由上图可知，每类应用平均上行流量大多大于下行流量，o类应用平均上行流量和下行流量都比较大，s类应用平均上行流量最大。

**4.3用户偏好**

对0,1,2三类用户分别计数，发现第0类用户使用的app种类较少，而且仅对t类与unknown类app使用较多。第1类用户爱好广泛，unknown使用较少，主要使用f和c类。第2类用户同样爱好广泛，但重度热爱的app种类也较多：依次有f、c、g、d、e。

**表4- 各类别用户使用不同类型app的交叉计数**

|  |  |  |
| --- | --- | --- |
| user\_type | app\_type | count |
| **0** | k | 39 |
| l | 5 |
| m | 27 |
| o | 121 |
| p | 233 |
| q | 107 |
| s | 43 |
| t | 2407 |
| unknown | 2626 |
| **1** | a | 188 |
| b | 40 |
| c | 1002 |
| d | 161 |
| e | 443 |
| f | 2026 |
| g | 671 |
| h | 254 |
| i | 73 |
| j | 203 |
| k | 101 |
| l | 5 |
| m | 4 |
| n | 23 |
| o | 58 |
| p | 101 |
| q | 40 |
| s | 18 |
| t | 341 |
| unknown | 189 |
| **2** | a | 364 |
| b | 60 |
| c | 2095 |
| d | 1025 |
| e | 1005 |
| f | 3706 |
| g | 1488 |
| h | 237 |
| i | 141 |
| j | 358 |
| k | 144 |
| l | 3 |
| m | 4 |
| n | 62 |
| r | 1 |

1. 构建初步推荐系统

**5.1 ALS推荐系统（分布式实现）**

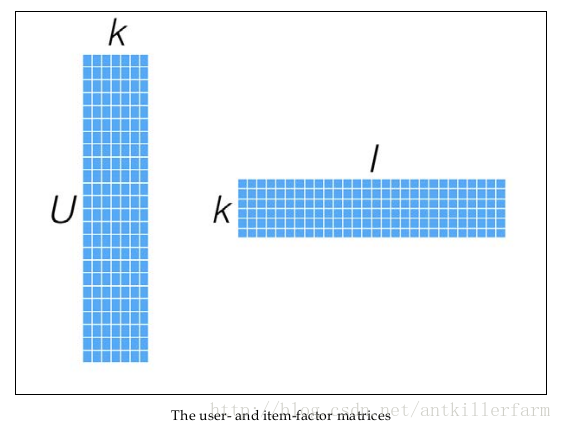
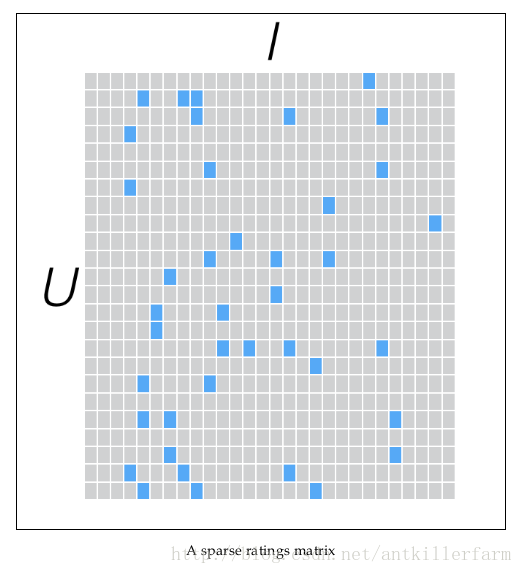


图5- ALS协同过滤模型示意图

通过spark的ML库自带的推荐系统，选取k=50构建了以下模型，rating使用日均使用时长（对大于4小时的直接取4）。为了验证推荐的差异性，对0、1、2三类用户中随机各选取了一个实力，给出前30的推荐结果以及其实际使用的app及rating。

**表4- 第0类用户推荐实例（uid=289899198）**

|  |  |  |  |
| --- | --- | --- | --- |
| appid | usage\_average\_hours | app\_type | data type |
| 17735 | 1.040475118 | unknown | recommend |
| 53723 | 1.010016572 | unknown | recommend |
| 1105 | 1.000622618 | unknown | recommend |
| 27690 | 0.983041486 | unknown | recommend |
| 936 | 0.981290367 | unknown | recommend |
| 15304 | 0.951119067 | unknown | recommend |
| 7005 | 0.943596919 | unknown | recommend |
| 18000 | 0.938854503 | p | recommend |
| 16424 | 0.926294471 | unknown | recommend |
| 11877 | 0.916901539 | t | recommend |
| 16065 | 0.892368282 | unknown | recommend |
| 8823 | 0.888022325 | unknown | recommend |
| 16258 | 0.878370742 | unknown | recommend |
| 10438 | 0.870159724 | unknown | recommend |
| 19350 | 0.862226563 | unknown | recommend |
| 10600 | 0.854875869 | unknown | recommend |
| 3309 | 0.834896556 | f | recommend |
| 27463 | 0.830922515 | unknown | recommend |
| 5608 | 0.810433328 | unknown | recommend |
| 10759 | 0.804366165 | unknown | recommend |
| 17818 | 0.803082166 | unknown | recommend |
| 27546 | 0.773087631 | unknown | recommend |
| 11738 | 0.762764456 | unknown | recommend |
| 22309 | 0.736345305 | unknown | recommend |
| 11951 | 0.728006899 | t | recommend |
| 10990 | 0.719243025 | unknown | recommend |
| 19136 | 0.708399674 | unknown | recommend |
| 16719 | 0.700600192 | unknown | recommend |
| 20619 | 0.683958824 | unknown | recommend |
| 22146 | 0.678908403 | j | recommend |
| 18000 | 1.184777778 | p | meta |
| 3309 | 0.888259259 | f | meta |
| 17442 | 0.182481481 | e | meta |
| 13667 | 0.160546296 | unknown | meta |
| 13043 | 0.101472222 | unknown | meta |
| 2548 | 0.089314815 | unknown | meta |
| 11088 | 0.039638889 | unknown | meta |
| 16759 | 0.028055556 | f | meta |
| 17355 | 0.026546296 | d | meta |
| 8832 | 0.021907407 | unknown | meta |
| 8165 | 0.015490741 | c | meta |
| 20720 | 0.014398148 | unknown | meta |
| 17938 | 0.010046296 | s | meta |
| 6367 | 0.007333333 | unknown | meta |
| 2859 | 0.006962963 | unknown | meta |
| 23310 | 0.006166667 | unknown | meta |
| 10686 | 0.005722222 | unknown | meta |
| 12795 | 0.004851852 | n | meta |
| 5205 | 0.004194444 | unknown | meta |
| 5657 | 0.003166667 | g | meta |
| 20124 | 0.002296296 | unknown | meta |
| 22509 | 0.002157407 | k | meta |
| 1723 | 0.001805556 | unknown | meta |
| 9628 | 0.001287037 | l | meta |
| 22687 | 0.001268519 | unknown | meta |
| 13162 | 0.000296296 | unknown | meta |
| 22003 | 0.000148148 | unknown | meta |
| 15595 | 0.000101852 | unknown | meta |
| 6588 | 9.26E-05 | unknown | meta |
| 3204 | 8.33E-05 | unknown | meta |
| 3427 | 8.33E-05 | unknown | meta |
| 7541 | 7.41E-05 | unknown | meta |
| 8646 | 6.48E-05 | n | meta |
| 22017 | 3.70E-05 | q | meta |
| 19014 | 2.78E-05 | unknown | meta |
| 21149 | 1.85E-05 | g | meta |
| 23544 | 1.85E-05 | unknown | meta |
| 2635 | 1.85E-05 | unknown | meta |
| 10055 | 1.85E-05 | unknown | meta |
| 11995 | 1.85E-05 | unknown | meta |
| 5250 | 9.26E-06 | unknown | meta |

第0类用户重点使用unknown和t类应用，推荐系统推荐的也主要基于此，发现推荐的3309应用给出的rating为0.835，而实际的rating为0.89，精度相当可观。

**表4- 第1类用户推荐实例（uid=1012024248）**

|  |  |  |  |
| --- | --- | --- | --- |
| appid | usage\_average\_hours | app\_type | data type |
| 7005 | 0.819997683 | unknown | recommend |
| 11399 | 0.66100014 | h | recommend |
| 15668 | 0.651768967 | unknown | recommend |
| 21749 | 0.648412344 | h | recommend |
| 13252 | 0.5933882 | unknown | recommend |
| 1446 | 0.557624082 | unknown | recommend |
| 507 | 0.54940568 | unknown | recommend |
| 17829 | 0.546536716 | unknown | recommend |
| 476 | 0.531855028 | unknown | recommend |
| 16924 | 0.515872554 | unknown | recommend |
| 3906 | 0.506763835 | unknown | recommend |
| 9464 | 0.501139561 | unknown | recommend |
| 21975 | 0.499201185 | unknown | recommend |
| 21699 | 0.492960424 | unknown | recommend |
| 13619 | 0.491923721 | h | recommend |
| 1413 | 0.472544709 | unknown | recommend |
| 14676 | 0.460350093 | unknown | recommend |
| 57324 | 0.452224246 | unknown | recommend |
| 14799 | 0.450808475 | unknown | recommend |
| 6710 | 0.448932593 | unknown | recommend |
| 19265 | 0.443168637 | unknown | recommend |
| 4984 | 0.436548827 | unknown | recommend |
| 16222 | 0.430061985 | unknown | recommend |
| 10505 | 0.428767278 | t | recommend |
| 1397 | 0.424514355 | h | recommend |
| 20271 | 0.423495474 | t | recommend |
| 6597 | 0.4221417 | e | recommend |
| 13314 | 0.420994976 | unknown | recommend |
| 17976 | 0.415754993 | unknown | recommend |
| 978 | 0.403730406 | j | recommend |
| 6597 | 0.610842593 | e | meta |
| 13043 | 0.506027778 | unknown | meta |
| 18136 | 0.336546296 | i | meta |
| 13667 | 0.260018519 | unknown | meta |
| 4803 | 0.186907407 | a | meta |
| 16759 | 0.174027778 | f | meta |
| 17355 | 0.162722222 | d | meta |
| 6367 | 0.112472222 | unknown | meta |
| 6192 | 0.097814815 | unknown | meta |
| 21776 | 0.090148148 | g | meta |
| 16061 | 0.080111111 | g | meta |
| 10686 | 0.074203704 | unknown | meta |
| 3309 | 0.057722222 | f | meta |
| 2859 | 0.043333333 | unknown | meta |
| 949 | 0.030703704 | i | meta |
| 3482 | 0.029157407 | g | meta |
| 3034 | 0.024740741 | e | meta |
| 14772 | 0.024287037 | n | meta |
| 8165 | 0.018231481 | c | meta |
| 11088 | 0.016407407 | unknown | meta |
| 20124 | 0.016148148 | unknown | meta |
| 20720 | 0.014916667 | unknown | meta |
| 21106 | 0.013898148 | g | meta |
| 23650 | 0.011018519 | g | meta |
| 3427 | 0.010638889 | unknown | meta |
| 776 | 0.010425926 | g | meta |
| 17937 | 0.008490741 | s | meta |
| 10055 | 0.007685185 | unknown | meta |
| 12388 | 0.007314815 | i | meta |
| 11995 | 0.006361111 | unknown | meta |
| 1093 | 0.005907407 | a | meta |
| 17142 | 0.005444444 | a | meta |
| 17247 | 0.004768519 | g | meta |
| 11320 | 0.003314815 | g | meta |
| 13162 | 0.003092593 | unknown | meta |
| 23310 | 0.002851852 | unknown | meta |
| 2820 | 0.002842593 | f | meta |
| 16745 | 0.002611111 | g | meta |
| 22687 | 0.002092593 | unknown | meta |
| 6841 | 0.001898148 | g | meta |
| 19816 | 0.001268519 | i | meta |
| 15151 | 0.001046296 | g | meta |
| 19014 | 0.001009259 | unknown | meta |
| 21079 | 0.000888889 | g | meta |
| 13528 | 0.000611111 | i | meta |
| 8832 | 0.000287037 | unknown | meta |
| 23078 | 0.000222222 | i | meta |
| 13990 | 0.000185185 | a | meta |
| 9112 | 0.000185185 | unknown | meta |
| 4664 | 0.000175926 | unknown | meta |
| 21553 | 0.000175926 | unknown | meta |
| 4724 | 0.00012963 | q | meta |
| 16452 | 0.00012037 | f | meta |
| 17069 | 0.000101852 | unknown | meta |
| 13108 | 9.26E-05 | unknown | meta |
| 22442 | 9.26E-05 | g | meta |
| 56554 | 8.33E-05 | unknown | meta |
| 23547 | 5.56E-05 | k | meta |
| 8400 | 4.63E-05 | c | meta |
| 16078 | 3.70E-05 | unknown | meta |
| 18153 | 2.78E-05 | e | meta |
| 2974 | 9.26E-06 | unknown | meta |

虽然第1类用户特点主要是较少使用unknown较多使用f，但是不是所有第1类用户均是这种特征，发现推荐的unknown应用较多，实际上该用户使用的unknown应用也较多。

**表4- 第2类用户推荐实例（uid=1824575646）**

|  |  |  |  |
| --- | --- | --- | --- |
| appid | usage\_average\_hours | app\_type | data type |
| 16065 | 2.052123231 | unknown | recommend |
| 10062 | 1.571899781 | unknown | recommend |
| 19011 | 1.516709273 | t | recommend |
| 3309 | 1.478687036 | f | recommend |
| 27463 | 1.4494365 | unknown | recommend |
| 8471 | 1.425099439 | unknown | recommend |
| 13614 | 1.358917441 | t | recommend |
| 15304 | 1.33512486 | unknown | recommend |
| 19031 | 1.316729363 | unknown | recommend |
| 19350 | 1.282445526 | unknown | recommend |
| 16258 | 1.275055423 | unknown | recommend |
| 602 | 1.270300546 | unknown | recommend |
| 16719 | 1.269081513 | unknown | recommend |
| 10438 | 1.255255719 | unknown | recommend |
| 1105 | 1.22307127 | unknown | recommend |
| 21669 | 1.216847323 | unknown | recommend |
| 17818 | 1.193145973 | unknown | recommend |
| 20501 | 1.192709721 | unknown | recommend |
| 10099 | 1.170808437 | unknown | recommend |
| 13747 | 1.155817308 | unknown | recommend |
| 15475 | 1.151740458 | t | recommend |
| 20619 | 1.146145875 | unknown | recommend |
| 13152 | 1.135951439 | unknown | recommend |
| 10637 | 1.124057328 | unknown | recommend |
| 16424 | 1.123501006 | unknown | recommend |
| 9685 | 1.106060361 | t | recommend |
| 15814 | 1.090866082 | t | recommend |
| 19042 | 1.09036874 | unknown | recommend |
| 17543 | 1.079553939 | t | recommend |
| 1768 | 1.075097683 | unknown | recommend |
| 3309 | 1.611638889 | f | meta |
| 11332 | 0.946388889 | t | meta |
| 7280 | 0.915518519 | c | meta |
| 9098 | 0.337462963 | c | meta |
| 4803 | 0.164601852 | a | meta |
| 11088 | 0.043314815 | unknown | meta |
| 10686 | 0.028685185 | unknown | meta |
| 13043 | 0.025083333 | unknown | meta |
| 13667 | 0.024648148 | unknown | meta |
| 22760 | 0.024009259 | unknown | meta |
| 2292 | 0.022314815 | q | meta |
| 6192 | 0.021833333 | unknown | meta |
| 62 | 0.01712963 | b | meta |
| 17355 | 0.015944444 | d | meta |
| 17030 | 0.00837963 | t | meta |
| 7540 | 0.006740741 | i | meta |
| 2859 | 0.006212963 | unknown | meta |
| 3427 | 0.005759259 | unknown | meta |
| 6764 | 0.005527778 | q | meta |
| 14540 | 0.004731481 | o | meta |
| 18552 | 0.004333333 | e | meta |
| 18547 | 0.004268519 | t | meta |
| 3962 | 0.003814815 | g | meta |
| 9795 | 0.003259259 | n | meta |
| 16745 | 0.00262963 | g | meta |
| 5205 | 0.002166667 | unknown | meta |
| 8832 | 0.001805556 | unknown | meta |
| 20720 | 0.001777778 | unknown | meta |
| 13528 | 0.000944444 | i | meta |
| 23410 | 0.000666667 | p | meta |
| 5651 | 0.000314815 | f | meta |
| 3022 | 0.000296296 | unknown | meta |
| 4 | 0.000287037 | g | meta |
| 13162 | 0.00025 | unknown | meta |
| 2974 | 0.000175926 | unknown | meta |
| 9112 | 0.000101852 | unknown | meta |
| 857 | 7.41E-05 | unknown | meta |
| 21079 | 6.48E-05 | g | meta |
| 12228 | 5.56E-05 | unknown | meta |
| 3443 | 4.63E-05 | unknown | meta |
| 3523 | 4.63E-05 | unknown | meta |
| 4664 | 2.78E-05 | unknown | meta |
| 7525 | 2.78E-05 | s | meta |
| 21526 | 2.78E-05 | unknown | meta |
| 1591 | 1.85E-05 | c | meta |
| 20124 | 9.26E-06 | unknown | meta |

第2类用户特点主要较多使用f，t，推荐中给予了多个

T类应用推荐，同时给予了大量unknown推荐，实际上该用户也确实使用大量unknown应用。

5.2 Recommenderlab包推荐系统

5.1.1 Recommenderlab简介

recommenderlab是一个用于做推荐的R包，提供了多种推荐算法，包括UBCF(基于用户的协同过滤)，IBCF(基于物品的协同过滤), Popular, Random，SVD，PCA，AR等。我们使用UBCF和IBCF两种算法构建推荐系统并对结果进行比较。

recommenderlab的核心数据结构是RatingMatrix（评分矩阵），该矩阵记录了user对item的评分。以RatingMatrix为输入，利用上述算法估计用户对尚未评价item的评分，并依此为依据进行推荐；针对此次作业，user即用户，item即应用，以30天APP用户对某一应用使用的总时长（单位为秒）为评分。

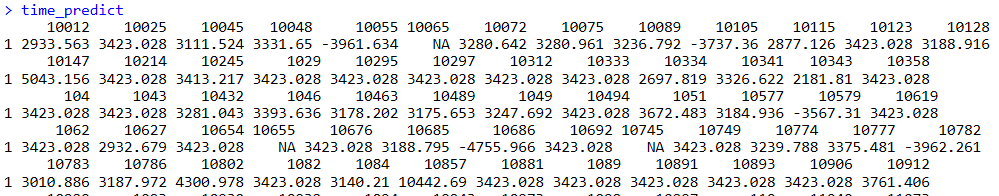
5.1.2数据说明及处理

将三十天每个用户使用单个APP的总时长以数据框形式存储。在处理过程中，发现appid中有两个特殊的数值，“河南省郑州市二七区康复后街”，“河南省郑州市二七区大学北路25号”，选取appid不是这两个的数据。得到总共有42675个不同的APP，48179位不同的用户。由于数据过多，选择被监测到次数大于100的用户且对应的APP被监测到次数大于20的数据，最后得到8637行数据，用来构建模型。

利用reshape包中的cast函数将数据转换数据，每个用户为一行，列为appid，数值为相应的总时间，如果没有使用某一APP，那么该app总时间就为NA，再转换为RealRatingMatrix型。

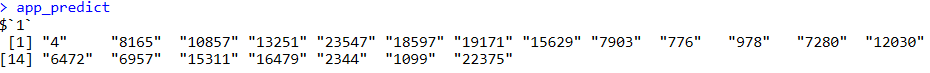
5.1.3模型构建

调用Recommender函数，参数method分别为UBCF（基于用户协同过滤算法）和IBCF（基于物品协同过滤算法），得到模型。再调用函predict进行预测，参数type为ratings得到评分矩阵，设置n则得到推荐的前n个应用。得到用户003A8891295BA3EA3A92E0F4B8F9CE42的评分矩阵部分如下（UBCF），得到该用户对各APP使用时间预测值，已经使用过的APP对应值为NA。

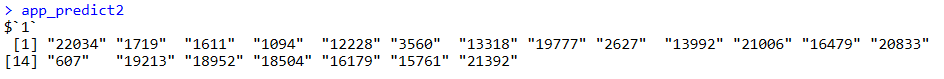


5.1.4预测与评估

通过UBCF模型预测推荐的前20个应用，如下：

****

通过IBCF模型预测推荐的前20个应用，如下所示：

****

进行对比知道，根据UBCF和IBCF模型进行推荐预测得到的结果相差较大，在UBCF中排第一位的并没有在IBCF模型推荐的前20个应用中出现，同样在IBCF推荐应用中排第一位的没有出现在UBCF推荐的前20个应用中。接下来进行评估两个模型。

调用evaluationScheme函数划分训练集和测试集进行预测，调用calcPredictionAccuracy函数对评分预测模型进行评估，得到UBCF和IBCF两个模型的RMSE、MSE、MAE值如下表：

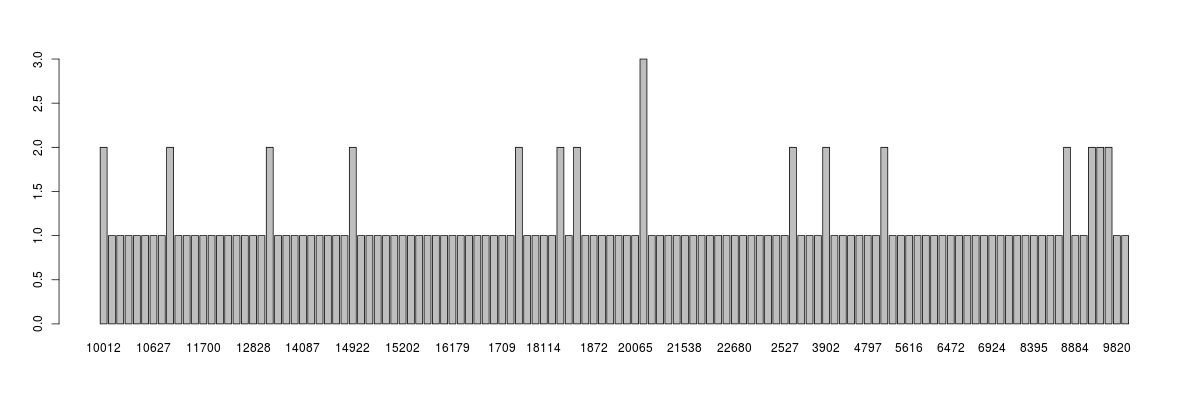
****

对比可知IBCF的三个值均小于UBCF模型，因此IBCF模型预测更好一些。

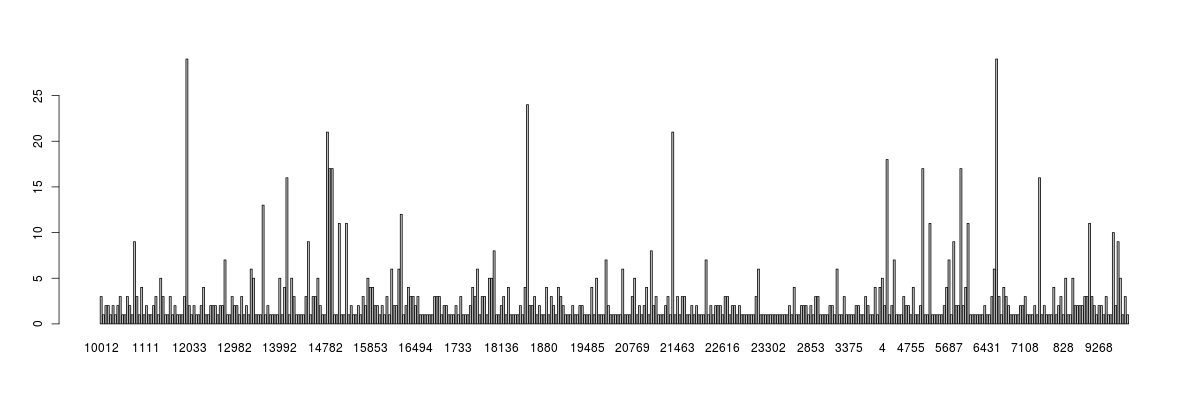
5.1.5对不同类别用户人群进行预测

由上面评估得到IBCF模型推荐更好些，因此根据IBCF模型针对K均值聚类得到的三个类别的用户人群进行应用推荐。构建模型涉及到的三类人数分别为0类7人，1类59人，2类5人。

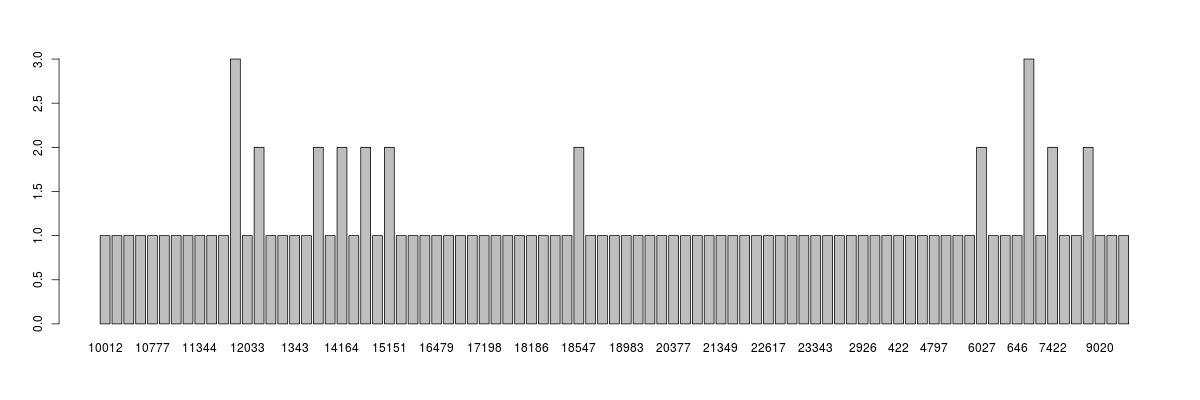
得到对0类推荐的应用频数分布如下图。



得到对1类人群推荐的应用频数分布如下图。

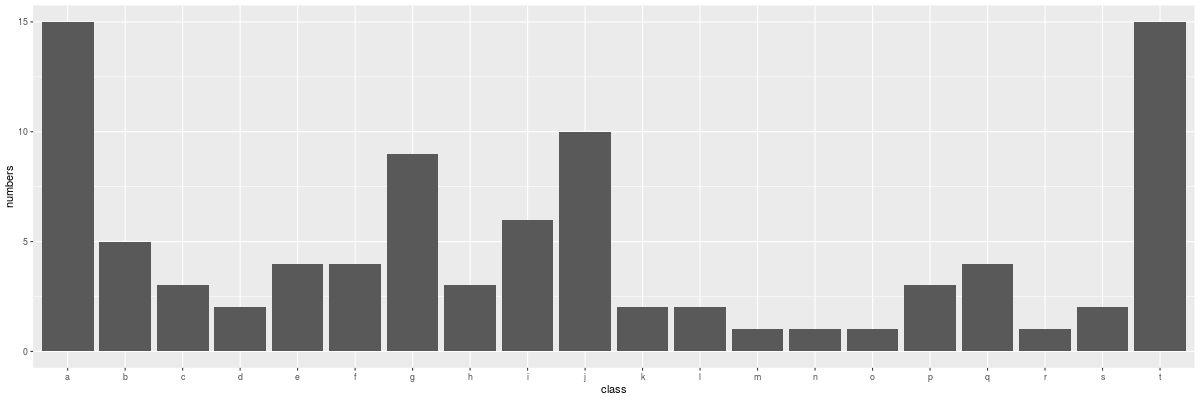


得到对2类人群推荐的应用频数分布如下图。

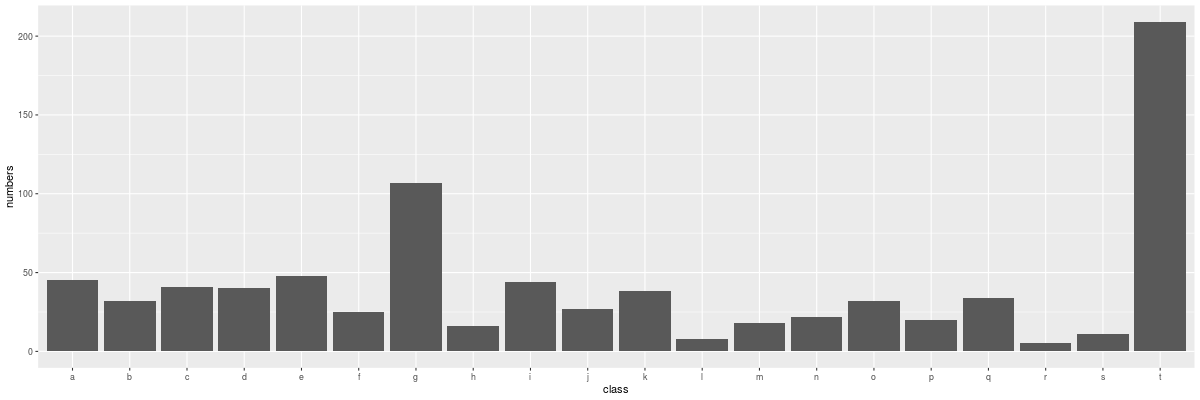


由上三图总结，由于0类用户人数和2类用户人群个数均较少，1类用户人数较多，因此对1类用户推荐的应用频数普遍高于为0类和2类用户推荐的应用频数，由对1类用户推荐的应用频数有较高数值，说明该推荐模型的推荐较为准确，因为对该用户群每个人推荐的应用重合度较高，因此说明推荐较为准确。

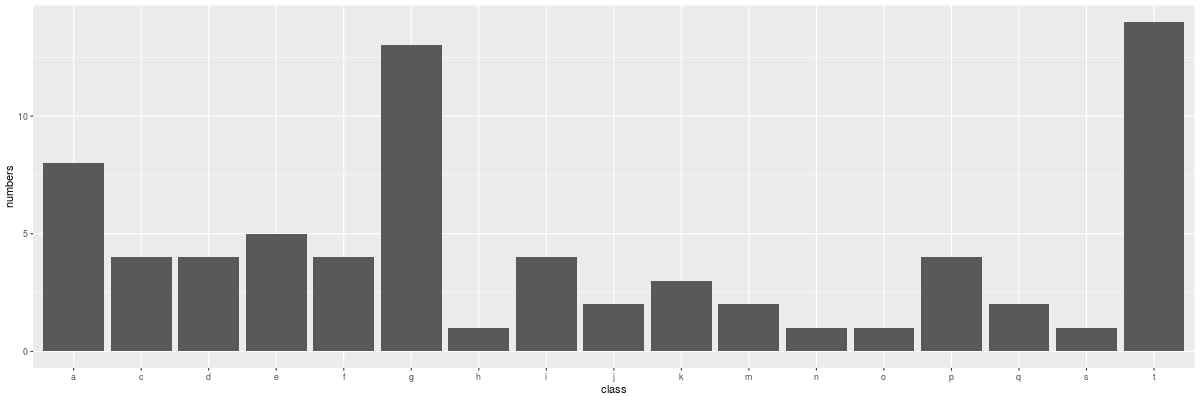
接下来分析对每类用户群推荐的应用类别分布，得到对三类用户群推荐的应用类别频数分布图如下所示。



对0类用户群推荐的应用类别频数分布图



对1类用户群推荐的应用类别频数分布图



对2类用户群推荐的应用类别频数分布图

由上图可知，对0类用户群，推荐的应用t类和a类均为15，是最多的应用类别，g类和j类也较多。对1类用户群，推荐的应用t类最多，其次为g类。对2类用户群，推荐的应用类最多的也是t类，其次为g类、a类。对三类用户群体推荐均是t类最多，a类和g类也较多。但是也有差别。这与应用的整体分布也有关系，在应用总体分布中，t类APP最多，其次就是g类。

同时，与三类用户群本身的特点进行比较，第0类用户使用的app种类较少，而且仅对t类app使用较多。第1类用户爱好广泛，主要使用f和c类。第2类用户同样爱好广泛经常使用的有f、c、g、d、e类应用。推荐结果与实际有些差别，也是说明推荐系统仍需改进，可能与建模选择的数据较少有关。

1. 推荐策略总结

总体上看,当前推荐策略仅仅依照现有模型在实际应用中是粗略的，我们将结合我们在数据分析过程中获取的经验与策略，指导初步推荐系统进一步筛选更为合适的推荐列表。下面给出在用户分类、用户行为时序特征、用户偏好几个板块中能对推荐列表筛选具有指导性的结论：

* 用户分为轻度、中度、重度用户，不同用户使用的app丰富度与偏好类别具有极大的不同，同时使用时段与使用强度均具有较大不同，针对不同类别的用户应当在推荐风格上做出差异，对于轻度用户应当更注重推荐工具类手机应用，使其逐步感到手机的方便，在对手机需求增大后，轻度用户可能转变为中度与重度后，才能进行大量推荐以满足其对app新鲜功能的需求。
* 用户在使用时段上存在差异，不同类型app也有不同的主流使用时段，不同app功能不同，在不同时段下用户优化存在差异，在同类且评分相近的app中应当考虑时段的匹配程度，选择时段最匹配的app达到用户体验最佳的效果。
* 针对当前大量app处于unknown类别以及新应用不断出现的时代，可以尝试用推荐系统比较不同的产品的相似度，以给出一个最相似的分类类别，用于辅助分析。
* 针对时序上存在的短期过强使用反而导致用户流失的现象实际上是普遍的，很多应用为了快速打开市场而采用比较激进的营销策略，但这些营销策略过后，用户反而可能流失。所以应用应当增强长期作战能力，在学习借鉴同类app的设计理念时，应当重点关注那些长期耐打的app。
* 基于协同过滤算法，利用用户的历史喜好信息，计算用户之间的距离，利用目标用户的最近邻居用户对物品评价的评价值来预测目标用户对特定物品的喜好程度，根据这一喜好程度对目标用户进行推荐。比如利用Recommenderlab包构建的推荐系统，通过UBCF（基于用户协同过滤算法）和IBCF（基于物品协同过滤算法）方法进行推荐。在本案例中，可以以用户使用应用的次数或者时长作为评分，进行推荐，既可以推荐具体的某一应用，又可以推荐某一类型的应用。