目录

1 场景数据简介 2
1.1 场景与数据简介 2
1.2 初识数据 2
1.3 简单描述性统计 3
2 用户行为分类 5
2.1 用户行为特征 5
2.2 K 均值聚类 7
2.3 定义用户类别 8
3 用户行为时序特征 9
3.1 使用时段集中特征 9
3.2 影响使用频率的重要因子 10
3.3 用户留存时序特征 11
4 应用使用偏好 12
4.1 应用特征 12
4.2 应用类别差异 14
4.3 用户偏好 15
5 构建初步推荐系统 17
5.1 ALS 推荐系统（分布式实现） 17
5.2 Recommenderlab 包推荐系统 24
6 推荐策略总结 27
参考文献 28
附录 29

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

1.1 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。

1.2 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。

i. 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

ii. 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。

iii. 所涉及到应用个数为 3694 个

iv. 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。

v. 变量 duration 中，最大值为 1.428e+09，是明显的异常值。

1.3 简单描述性统计

i. 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，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。

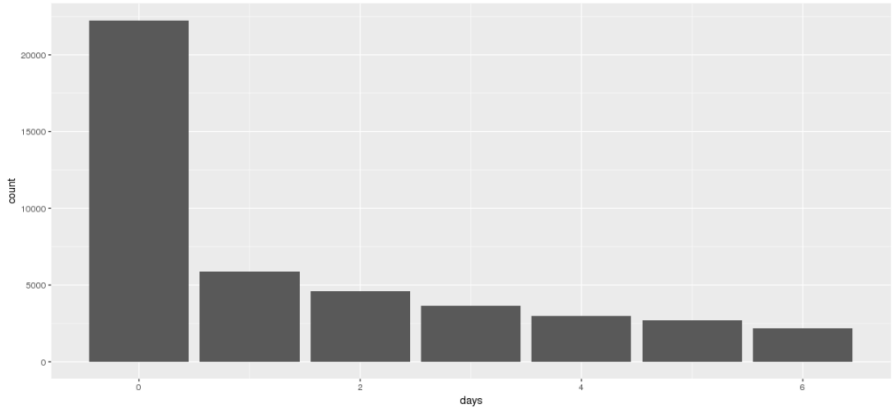


图 1-1 七天用户监测数据缺失的天数频数分布直方图

ii. 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。

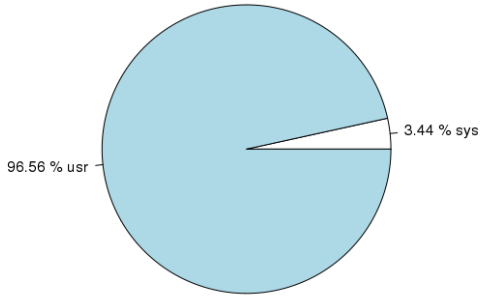


图 1-2 系统软件与用户安装软件占比饼图

iii. 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。

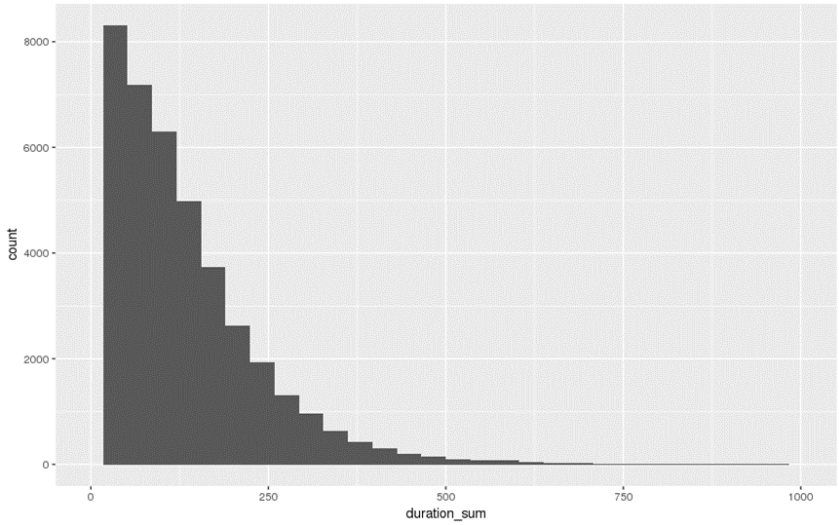


图 1-3 每位用户每天平均使用 APP 时长频数分布直方图

iv. 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。

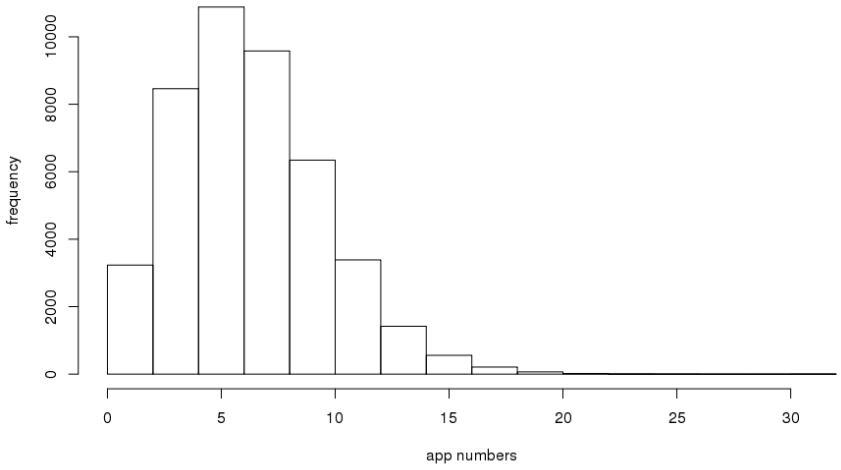


图 1-4 每位用户每天平均使用 APP 个数频数分布直方图

v. 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。

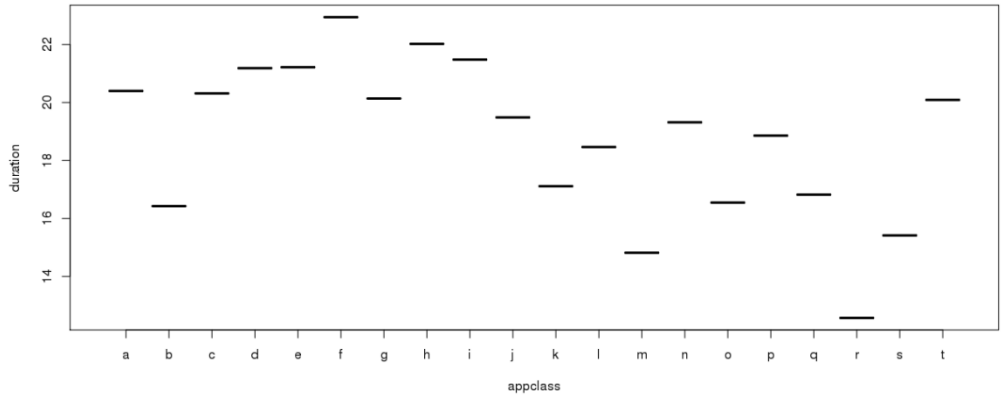


图 1-5 各类 APP 使用总时长

vi. 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。

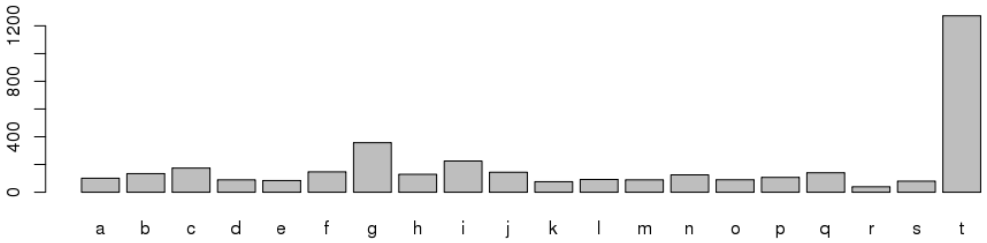


图 1-6 各类软件中的软件个数柱状图

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

	user1	user3	user4	user5
count	22244.000000	22244.000000	22244.000000	2.224400e+04
mean	14.373353	13.670653	170.709636	5.704573e+03
std	5.060598	3.155627	125.214709	2.691227e+05
min	1.000000	0.142857	1.857143	4.594736e+00
25%	11.000000	11.714286	86.285714	4.844824e+01
50%	14.142857	13.857143	142.857143	7.086160e+01
75%	17.571429	15.857143	223.285714	1.068525e+02
max	40.142857	22.571429	3403.714286	3.030741e+07

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。

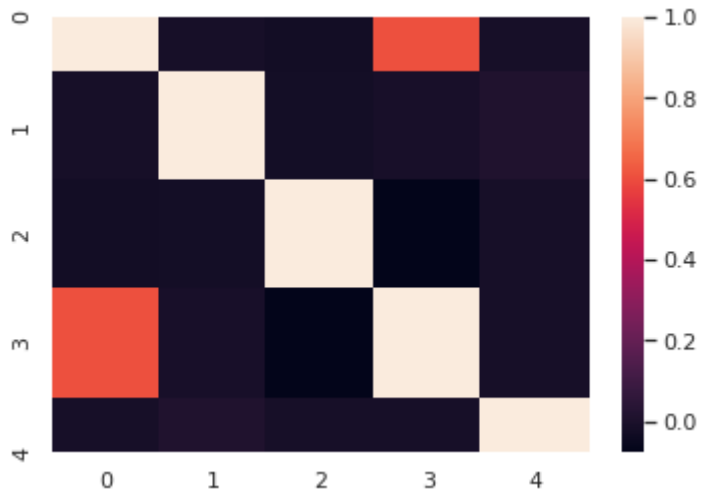


图 2-1 各个变量之间的相关程度

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

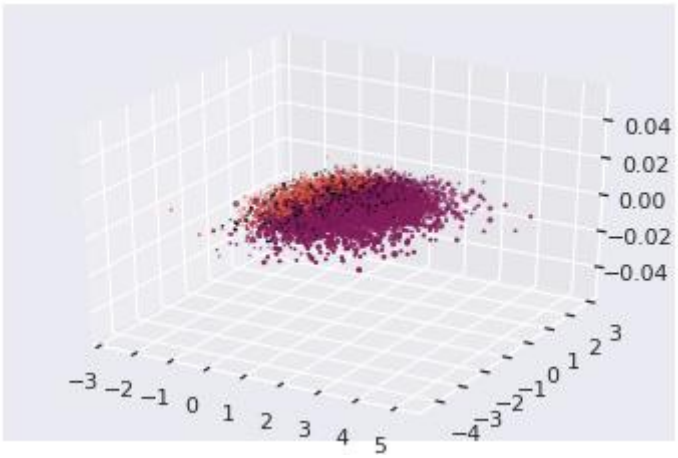


图 2-2 聚类结果

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。

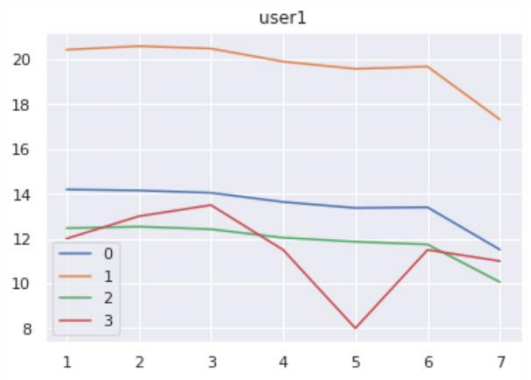


图 2-3

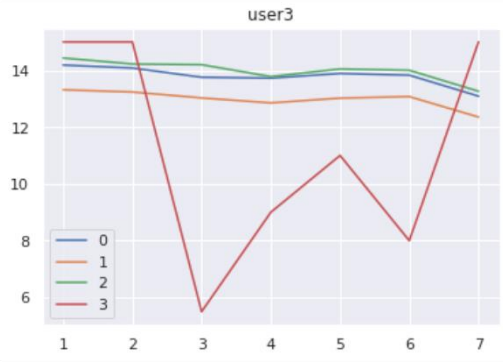


图 2-4

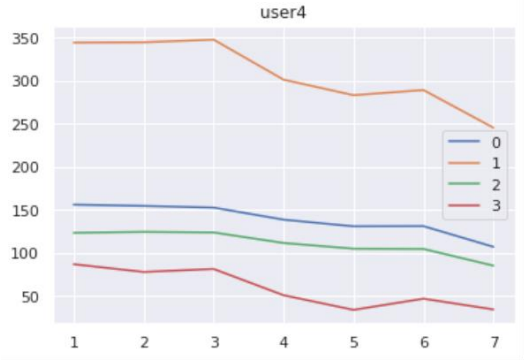


图 2-5　　图 2-6

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。

3 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。

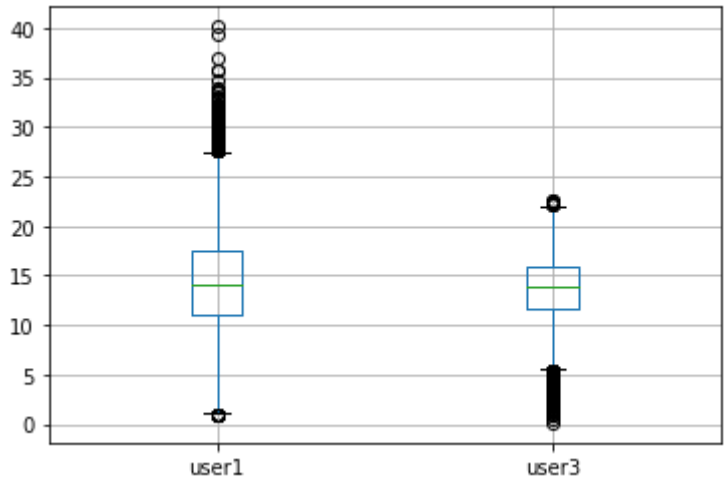


图 3-1 用户使用 app 时长最长的时段

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。

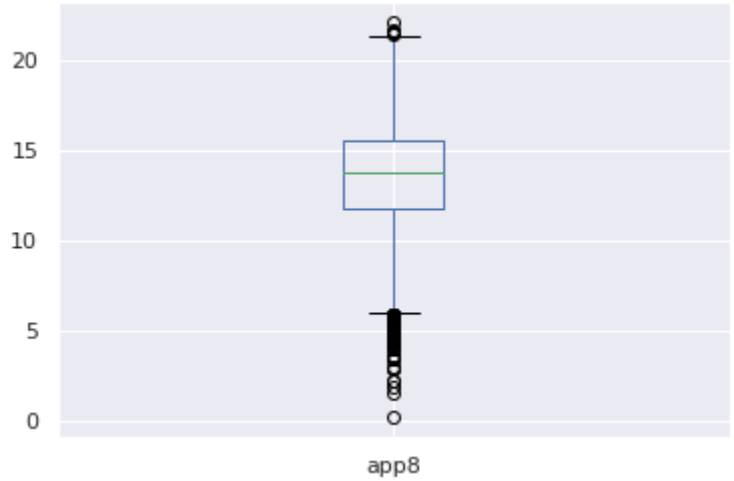


图 3-2 应用被使用时长最长的时间段

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。

OLS Regression Results

Dep. Variable:	frequency	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	927.7
Date:	Thu, 26 Mar 2020	Prob (F-statistic):	1.03e-18
Time:	18:53:45	Log-Likelihood:	70.484
No. Observations:	21	AIC:	-131.0
Df Residuals:	16	BIC:	-125.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6869	0.023	29.728	0.000	0.638	0.736
flow1	-1.03e-09	1.43e-09	-0.720	0.482	-4.06e-09	2e-09
flow2	1.251e-09	1.77e-09	0.707	0.490	-2.5e-09	5e-09
mone1	-0.1686	0.006	-26.037	0.000	-0.182	-0.155
mone2	0.2643	0.006	41.919	0.000	0.251	0.278

Omnibus:	5.704	Durbin-Watson:	2.286
Prob(Omnibus):	0.058	Jarque-Bera (JB):	3.951
Skew:	-1.050	Prob(JB):	0.139
Kurtosis:	3.330	Cond. No.	4.83e+08

图 3-3 回归结果

3.3 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。

表 10-3 因变量和自变量说明

变量符号	变量名称的含义	类型	单位/说明
y	第 24~30 天是否使用该类 APP	分类变量	*y*=1 使用; *y*=0 未使用
*x*1	第 24 天前最后一次使用该类 APP 的日期距离第 24 天的天数(市场营销领域中的“近度”(Recency)变量)	连续变量	天
*x*2	第 24 天前最后一次使用那天的使用强度	连续变量	秒
*x*3	前 23 天使用总天数除以有效观测天数(市场营销领域中的“频度”(Frequency)变量)	连续变量	无
*x*4	前 23 天使用天数当中的平均使用强度(市场营销领域中的“强度”(Monetary)变量)	连续变量	秒
*x*5	前 23 天有效观测天数当中的平均使用强度	连续变量	秒
*x*6	第 24 天前一天(第 23 天)的使用强度	连续变量	秒
*x*7	第 24 天前一周内(第 17~23 天)的有效观测天数当中的平均使用强度	连续变量	秒
*x*8	第 24 天前一周外(第 1~16 天)的有效观测天数当中的平均使用强度	连续变量	秒

图 3-4 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 3-1 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 3-2 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。

4 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 4-1 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。

	app1	app2	app4	app6	app7	app8	app9	app10
count	4539.000000	4.539000e+03	4.539000e+03	4539.000000	4539.000000	4539.000000	4.539000e+03	4539.000000
mean	1126.081610	1.156569e+03	1.550673e+05	99.335977	1303.864728	13.482013	1.156569e+03	1126.081610
std	20269.734708	4.769374e+04	1.846274e+06	945.939598	838.139597	2.949722	4.769374e+04	20269.734708
min	1.000000	1.000000e+00	0.000000e+00	1.000000	41.000000	0.142857	1.000000e+00	1.000000
25%	10.285714	3.247923e+01	4.209577e+02	2.000000	199.000000	11.714286	3.247923e+01	10.285714
50%	24.142857	7.540423e+01	4.088128e+03	4.571429	2000.000000	13.714286	7.540423e+01	24.142857
75%	79.428571	2.014648e+02	1.374187e+04	14.000000	2000.000000	15.571429	2.014648e+02	79.428571
max	870195.714286	3.138203e+06	7.171050e+07	27227.714286	2000.000000	22.142857	3.138203e+06	870195.714286

图 3-5 用户变量描述性统计

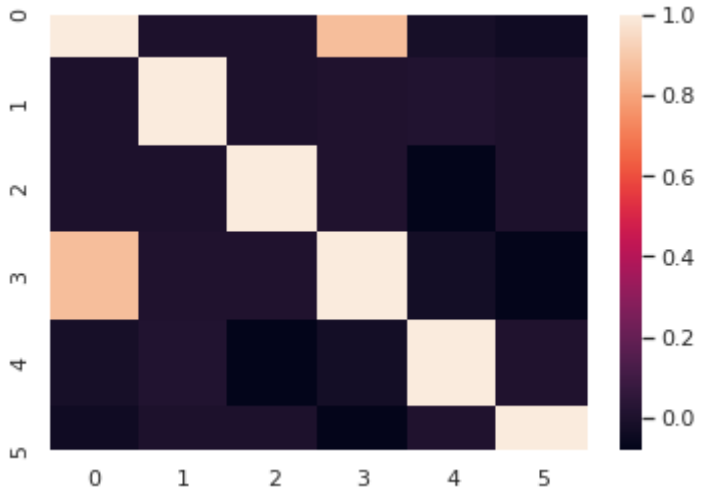


图 3-6 变量间的相关性热力图

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

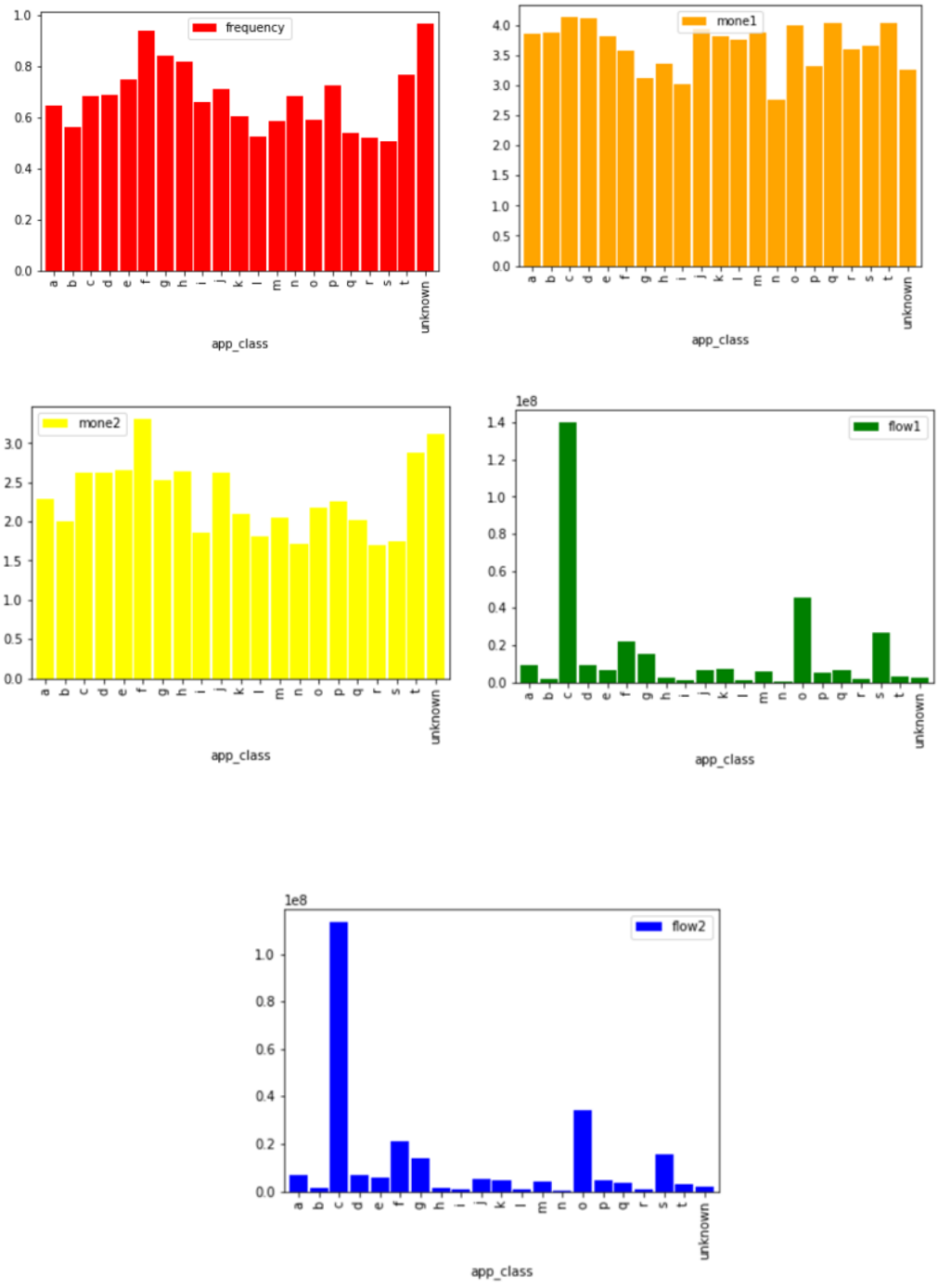


图 4-1 不同变量在 app 中的差异

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。

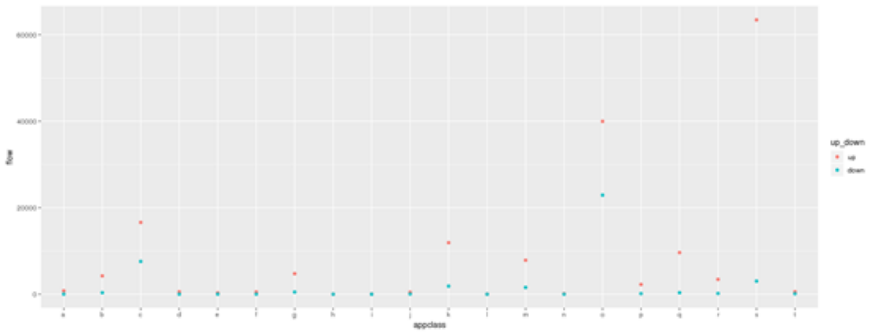


图 4-2 上下行流量在 app 中的差异

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 4-2 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

5 构建初步推荐系统

5.1 ALS 推荐系统（分布式实现）

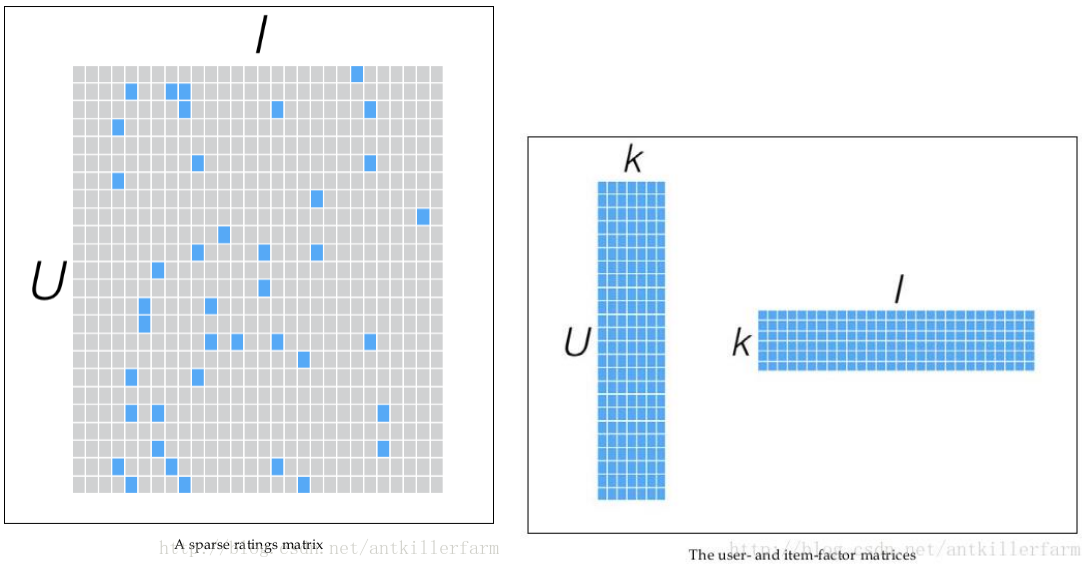


图 5-1 ALS 协同过滤模型示意图

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 5-1 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 5-2 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 5-3 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

5.2.1 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。

5.2.2 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。

5.2.3 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。

```
     10012    10025    10045    10048     10055 10065    10072    10075    10089    10105    10115    10123    10128
1 2933.563 3423.028 3111.524 3331.65 -3961.634    NA 3280.642 3280.961 3236.792 -3737.36 2877.126 3423.028 3188.916
     10147    10214    10245     1029    10295    10297    10312    10333    10334    10341    10343    10358
1 5043.156 3423.028 3413.217 3423.028 3423.028 3423.028 3423.028 3423.028 2697.819 3326.622 2181.81 3423.028
       104     1043    10432     1046    10463    10489     1049    10494     1051    10577    10579    10619
1 3423.028 3423.028 3281.043 3393.636 3178.202 3175.653 3247.692 3423.028 3672.483 3184.936 -3567.31 3423.028
      1062    10627    10654 10655    10676    10685     10686    10692 10745    10749    10774    10777     10782
1 3423.028 2932.679 3423.028    NA 3423.028 3188.795 -4755.966 3423.028    NA 3423.028 3239.788 3375.481 -3962.261
     10783    10786    10802     1082     1084    10857    10881     1089    10891    10893    10906    10912
1 3010.886 3187.972 4300.978 3423.028 3140.21 10442.69 3423.028 3423.028 3423.028 3423.028 3423.028 3761.406
```

5.2.4 预测与评估

通过 UBCF 模型预测推荐的前 20 个应用，如下：

```
 [1] "4"     "8165"  "10857" "13251" "23547" "18597" "19171" "15629" "7903"  "776"   "978"   "7280"  "12030"
[14] "6472"  "6957"  "15311" "16479" "2344"  "1099"  "22375"
```

通过 IBCF 模型预测推荐的前 20 个应用，如下所示：

```
 [1] "22034" "1719"  "1611"  "1094"  "12228" "3560"  "13318" "19777" "2627"  "13992" "21006" "16479" "20833"
[14] "607"   "19213" "18952" "18504" "16179" "15761" "21392"
```

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：

	RMSE	MSE	MAE
UBCF	37704.98	1421665648	10309.418
IBCF	15493.56	240050533	5866.105

图 5-2 评估结果

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。

5.2.5 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。选择推荐应用的前 20 个。

得到对 0 类推荐的应用频数分布如下图。

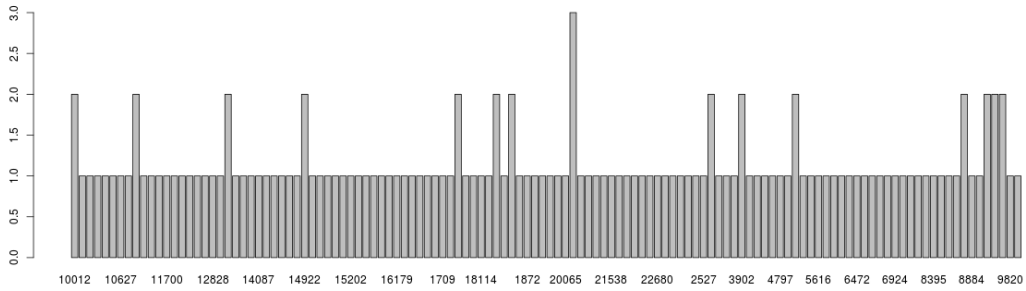


图 5-3 对 0 类用户推荐的应用频数分布图

得到对 1 类人群推荐的应用频数分布如下图。

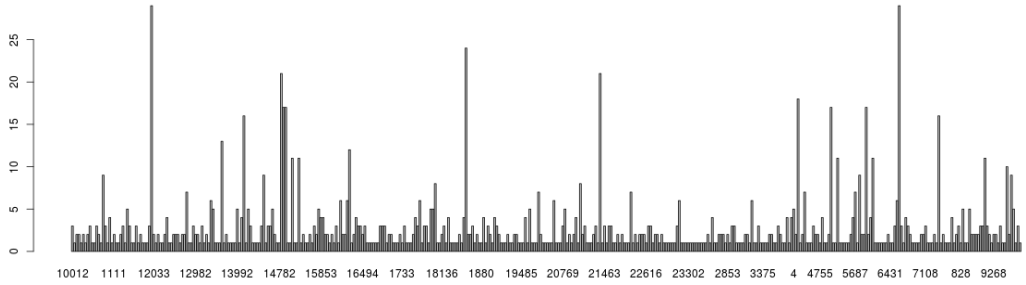


图 5-4 对 1 类用户推荐的应用频数分布图

得到对 2 类人群推荐的应用频数分布如下图。

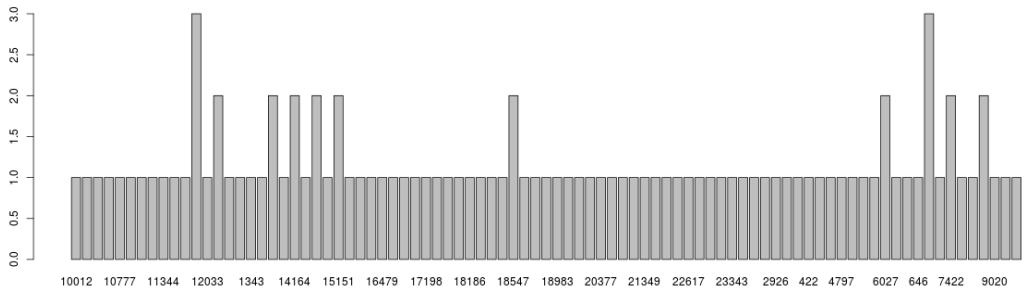


图 5-5 对 2 类用户推荐的应用频数分布图

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。

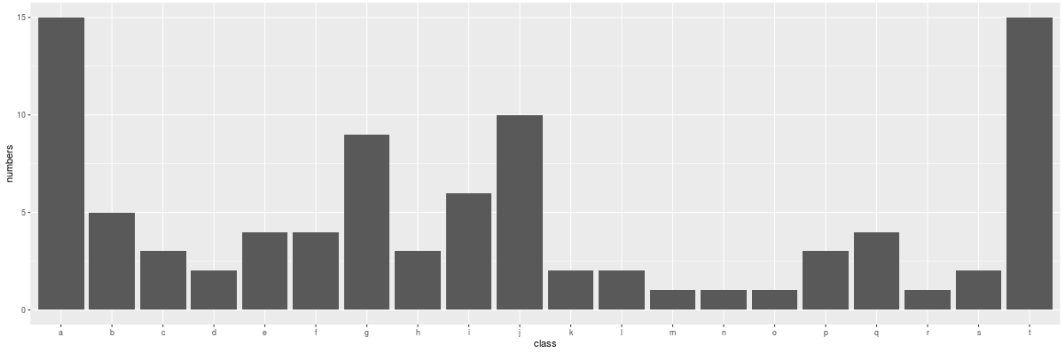


图 5-6 对 0 类用户推荐的应用类别频数分布图

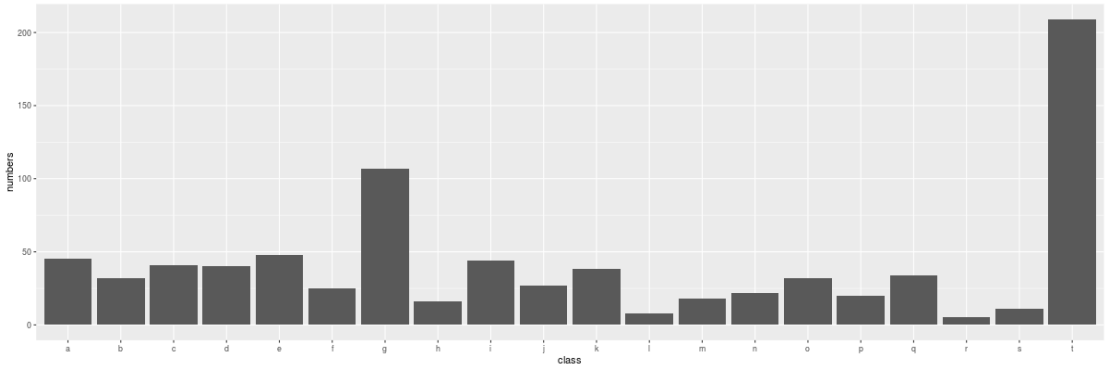


图 5-7 对 1 类用户推荐的应用类别频数分布图

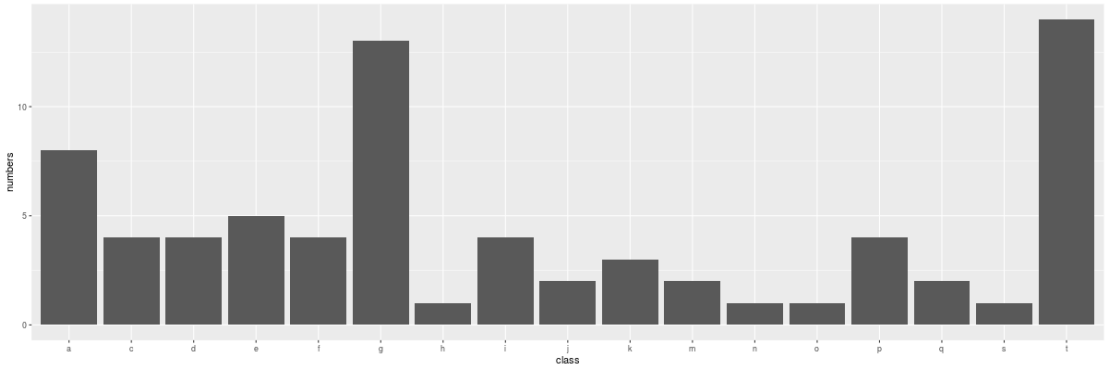


图 5-8 对 2 类用户推荐的应用类别频数分布图

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。

6 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：

i. 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。

ii. 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。

iii. 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。

iv. 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。

v. 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

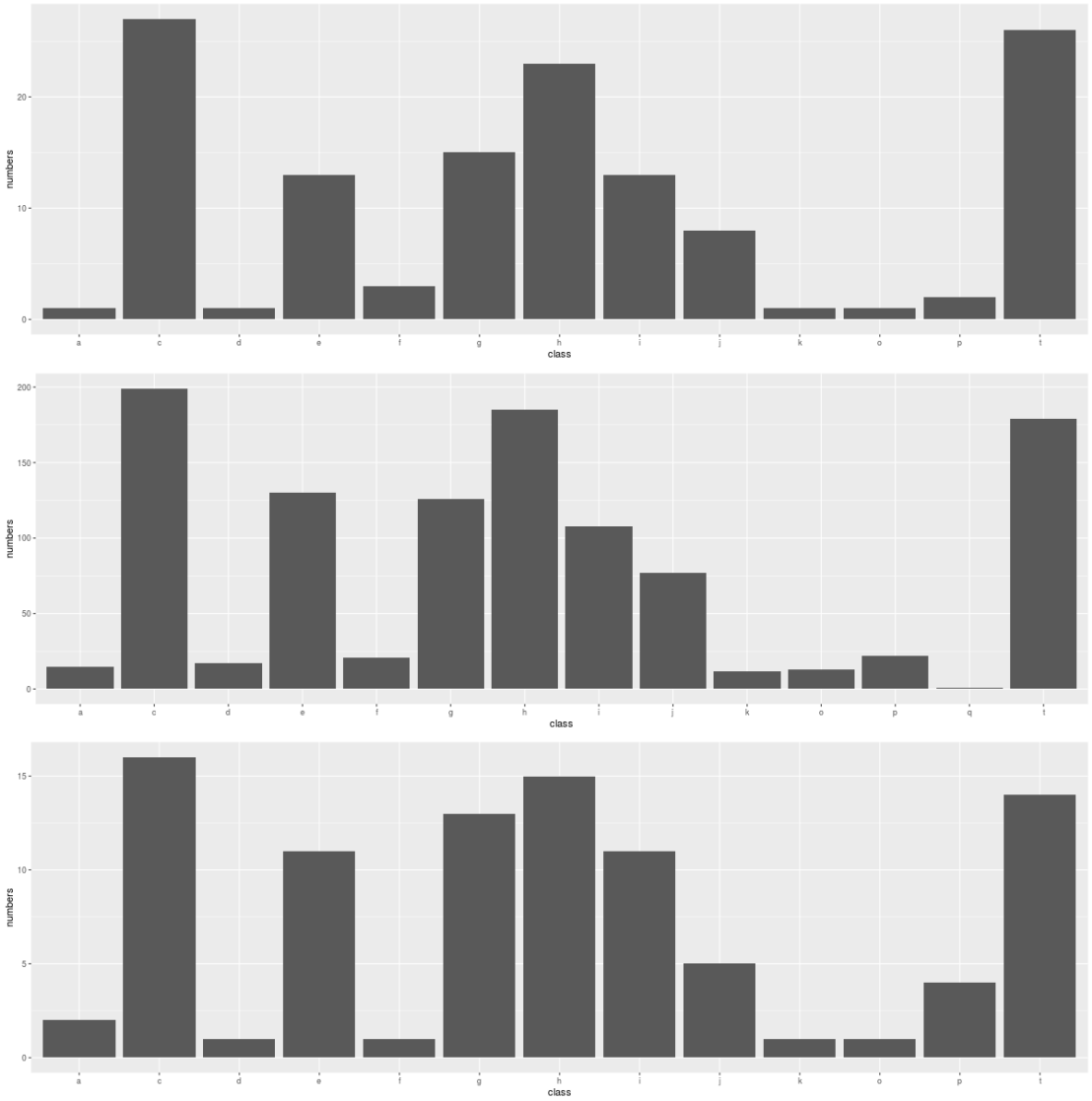
好程度对目标用户进行推荐。比如利用 Recommenderlab 包构建的推荐系统，通过 UBCF（基于用户协同过滤算法）和 IBCF（基于物品协同过滤算法）方法进行推荐。在本案例中，可以以用户使用应用的次数或者时长作为评分，进行推荐，既可以推荐具体的某一应用，又可以推荐某一类型的应用。

参考文献

[1]王国霞，刘贺平．个性化推荐系统综述[J]．计算机工程与应用，2012，48(7):66-76.
[2]杨清智．基于大数据技术的手机应用推荐系统的设计与实现[D]．哈尔滨工业大学.
[3]赵海燕，张健，曹健．基于主题分组与随机游走的 App 推荐算法[J]．计算机应用研究，v.35；No.322(8):43-46.
[4] 丁晨．基于混合推荐的手机阅读推荐系统的研究与实现[D]．重庆大学.

附录

UBCF 方法得到的三类用户推荐应用类别频数分布表分别如下图。



UBCF 方法得到的三类用户推荐的应用频数分布表分别如下图。