基于 APP 行为数据的用户预测问题分析

摘要

“互联网+”时代开辟了用户从事各项活动的新途径，人们依赖网络 APP 进行娱乐、消费、学习等多种活动。本文通过用户 APP 行为数据，建立了合理的预测用户行为的数学模型，并以其为基础建立了无人工干预的 AutoML 模型。

针对问题一，根据用户 APP 行为数据从而建立模型预测用户后期的消费行为。我们首先对所给数据做了数据清理、数据变换两种数据预处理；然后通过确立用户在该 APP 上的消费欲望、浏览时间及操作次数三个指标，分析了用户行为习惯与偏好。从消费欲望中我们可以看出用户对该消费的需求程度；浏览时间和操作次数这两个指标则反映了用户对其是否感兴趣。接着建立了 Logistic 回归模型，用柱状图、散点图清晰地反映出大部分用户在使用该 APP 消费时的心理与行为。由此我们可知，在越接近期望完成消费时间时用户消费欲望越高；浏览时间在 10 到 30 秒之间、操作次数在 50 到 100 次之间时用户消费的可能性最大。经过数学分析与建立模型，我们便可预测用户的消费行为。

对于问题二，我们针对问题一的建模过程，又考虑了应用数据场景变化和埋点类别变化等因素。首先使用 excel 和 SPSS 处理数据，接着运用 Python2.7 中的 Pandas 库、Graphlab Create 和 DataFrame 结构三个模块对数据进行进一步处理分析，最后处理完的数据变成 cvs 形式。通过以上过程设计出了适用于这种时序埋点数据的、无人工干预的 AutoML 模型。机器学习简单方便，可自动完成整个特征化，模型训练，模型选择/组合，超参数调优，部署上线等环节，高效率地解决算法问题。

综上所述，本文通过建立 Logistic 回归模型和 AutoML 模型，对用户 APP 行为数据进行了多方面的分析，用两种不同的方法预测用户未来的消费行为，这对于各大企业实时根据用户的行为采取相应措施，从而吸引更多客户，增大消费量具有重要的参考价值。

关键词：行为数据 Logisitic 回归模型 AuotoML 机器学习 行为预测

一、问题重述

1.1 背景资料与条件

随着“互联网+”时代的到来，人们愈发地依赖网络 APP 从事相关活动，其中包括了许多消费行为，例如购物消费、点评评价、信贷消费等。用户在 APP 上的活动体现了用户的不同行为习惯、行为偏好，同时不同用户从事相关活动的时间、顺序各不相同，同一用户在不同时间段从事同一活动的行为也有差异。

1.2 需要解决的问题

（1）试基于用户 APP 行为数据，建立合理的预测用户行为的数学模型，并分析评估其性能。

（2）针对问题一所建立的预测用户行为的模型，充分考虑数据应用场景变化、埋点类别变化等因素，设计出适用于这种时序埋点数据的、无人工干预的 AutoML 模型。

二、问题分析

2.1 问题一的分析

问题一要求基于用户 APP 行为数据，建立合理的预测用户行为的数学模型，我们可以主要根据用户在该 APP 上的消费欲望 D、浏览时间 T 及操作次数 F 三个指标来提取用户的行为特征。

针对用户消费欲望 D 这一指标，我们可利用用户实际完成消费时间（所给训练数据集中 apply_create_dt 列）和期望完成消费时间（所给训练数据集中 apply_expect_dt 列）来进行估计。两者时间的差值的大小代表着用户消费欲望程度。

针对用户在该 APP 上浏览时间 T 这一指标，我们可将埋点被抓获时间（所给训练数据集中 entry_time 列）进行时序排列，用最后一个时间减去第一个时间所得到的时间差，即为用户浏览所用时间。

针对用户在该 APP 上操作次数 F 这一指标，我们可将所给训练数据集中 apply_id 列进行去重处理，即可直接得到用户操作次数。

2.2 问题二的分析

问题二要求在问题一所建立的预测用户行为的模型基础上，充分考虑数据应用场景变化、埋点类别变化等因素，设计出适用于这种时序埋点数据的无人工干预的 AutoML 模型。对于 AUTOML 模型我们有如下看法：

（1）因为 AUTOML 模型是机器学习算法，需要数据的大量堆叠来产生合理曲线，所以第一步是如何构建数据特征，我们试图采用多个数据，从用户习惯、使用环境、欲望等特征来综合性的描述我们的算法，得到初等模型，并不断加以改造，提高模型的准确率。

（2）因为 AUTOML 模型学习方法简单，我们需要更加客观有效地评定数据的有效性，于是我们对于各种异常数据进行了删除，例如：如果张三的欲望显然的大于三倍标准差多倍以上，那么显然对欲望曲线的影响极大并且偏差极大，该数据应当予以删除，这个删除是机器学习的自动化的。

（3）模型构建方式多种，我们试图通过多次构建模型，寻找最合理拟合程度最高的模型，并且通过实际训练以及对分析结果的合理分析，最终取长补短，通过集中改造部分算法，提高最终的拟合程度。

三、模型假设

（1）假设用户在该 APP 上的浏览习惯即为其生活行为习惯。

（2）假设用户在该 APP 上的消费行为不受节日促销等时间因素的影响。

（3）假设用户下一次消费不受其之前的消费商品自身属性所影响。

四、符号说明

符号	说明
D	消费欲望
T	浏览时间

F	操作次数
α	实际完成消费时间
β	期望完成消费时间
t	埋点被抓获时间
t_1	埋点被抓获第一个时间点
t_2	埋点被抓获最后一个时间点

五、模型的建立与求解

5.1 问题一的模型建立与求解

5.1.1 指标确立

在用户使用该 APP 从事相关活动时，具有浏览和消费两种行为，基于用户的浏览和消费行为数据预测该用户是否要进行再一次消费行为，我们确立了用户在该 APP 上的消费欲望、浏览时间及操作次数三个指标。

（1）消费欲望D

消费欲望是指用户实际完成消费时间和期望完成消费时间的差值，公式表示如下：

消费欲望D =实际完成消费时间α（秒）-期望完成消费时间β（秒）

这一指标反映了用户对该消费行为的需求程度。若该指标越低，也就是实际完成消费时间越早于期望完成消费时间，说明用户对这种消费行为的需求越高，较可能地进行再次消费。反之，若该指标越高，实际完成消费时间越推迟于期望完成时间，说明该用户对这种消费犹豫不定，需求程度越低。

（2）浏览时间T

浏览时间的长短是评判用户对该消费是否感兴趣的重要指标之一，公式表示如下：

浏览时间T=埋点被抓获最后一个时间t_2（秒）-埋点被抓获第一个时间t_1（秒）

所谓埋点就是在应用服务器中的每一个页面中都嵌入一段 js 脚本，用户在访问页面时自动触发 js 收集用户访问行为日志，并提交到日志服务器。因此，我们可认为埋点被抓获的第一个时间即是用户打开应用程序的时间，埋点被抓获的最后一个时间即是用户关闭应用程序的时间。

（3）操作次数F

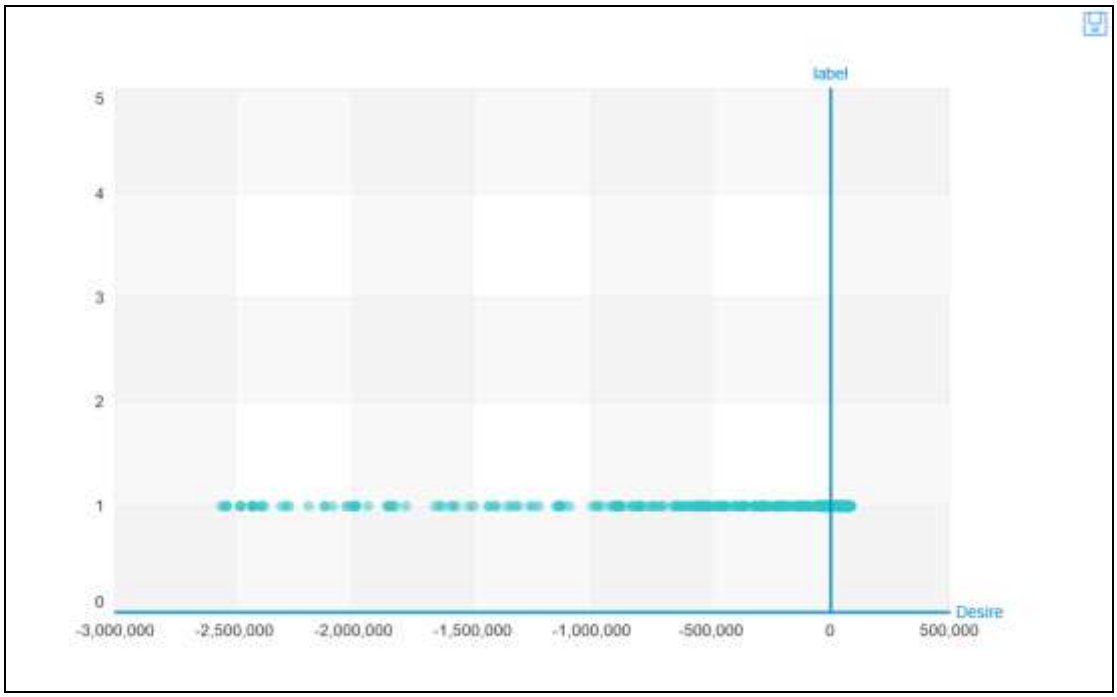
用户在该 APP 上的操作次数是评判用户对该消费是否感兴趣的另一重要指标。将所给训练数据集中 apply_id 列进行去重处理，即可直接得到用户操作次数。

5.1.2 模型建立与分析

Logistic 回归模型是对二分类因变量（即 y=1 或 y=0）进行回归分析时应用最普遍的多元量化分析方法，因将目标概率进行 logit 变换而得以避免线性概率模型的结构缺陷。在估计模型时采用极大似然估计的迭代，找到系数的“最可能”的估计。【1】

因此将“是否消费”中的“否”编码为“0”，“是”编码为“1”，之前为均衡样本随机抽取的 1700 万条训练集数据导入 SPSS 进行 Logistic 多元回归分析。其中“消费”作为因变量，其它三种指标即消费欲望、浏览时间和操作次数作为自变量，选择“向前 LR 方法”，经过 31 个步骤的运算后模型稳定。

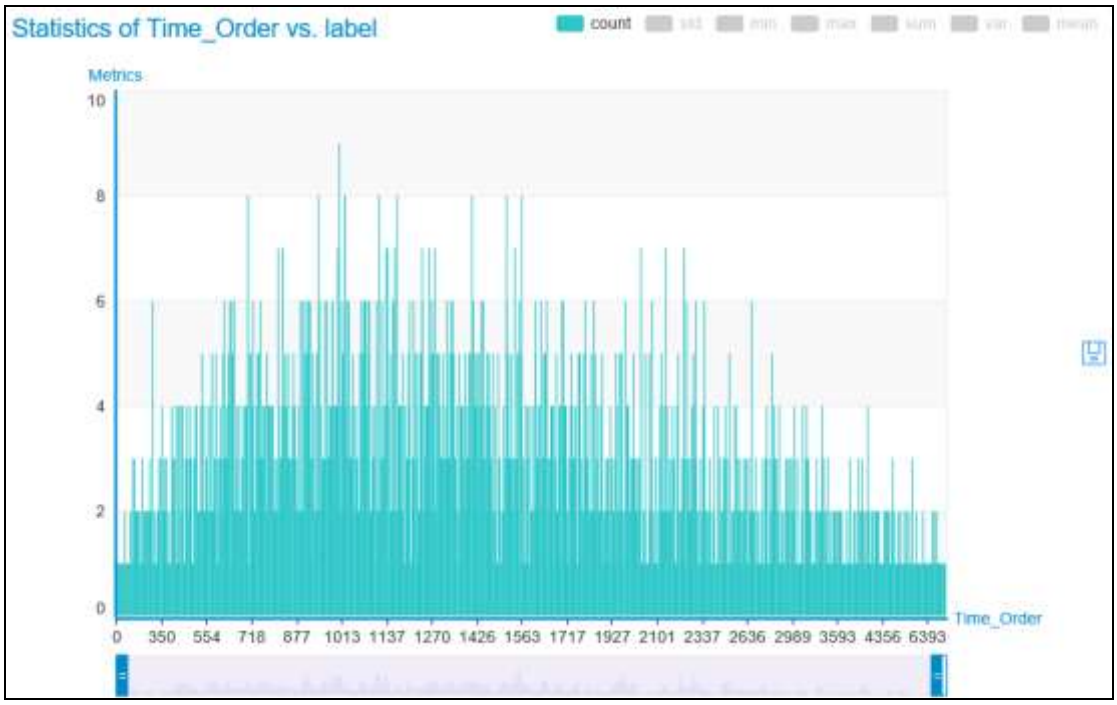
（1）消费欲望与消费行为



图一 消费欲望与消费行为

图一中横轴代表用户消费欲望大小，时间单位为秒，纵轴“1”代表用户消费。由此可知，大部分用户都是在期望完成消费时间之前完成了消费，并且越接近期望完成消费时间β，消费用户越多。通过计算可得用户在期望完成消费时间的前三天其消费欲望最高，符合生活习惯，同时也说明我们所建立的 Logistic 模型正确。

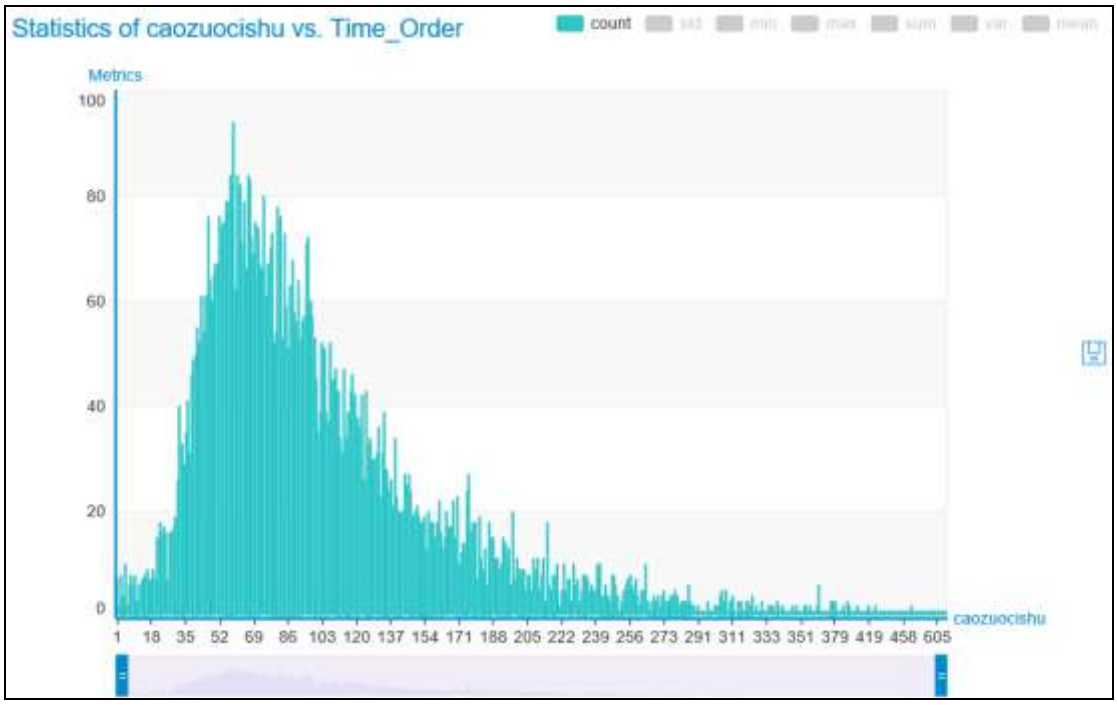
（2）浏览时间与消费行为



图二 浏览时间与消费行为

在有关浏览时间这一指标的分析上，我们所主要关注的对象是已经进行过二次购买的用户。如图二可见，横轴代表用户浏览某项消费行为所用时间，单位为秒，纵轴代表不同浏览时间的用户的数量所占权重，其中 718 秒到 1426 秒时所占权重较大。通过计算可得，在浏览 10~30 秒后产生消费行为的用户较多，我们可推测得，浏览时间过少的用户可能没有明确消费目标，广泛地随意浏览；相反的，浏览时间过长的用户可能仍因为某些因素犹豫不决。因此，我们可推断，浏览时间在 10~30 秒的用户将会消费的可能性更大。

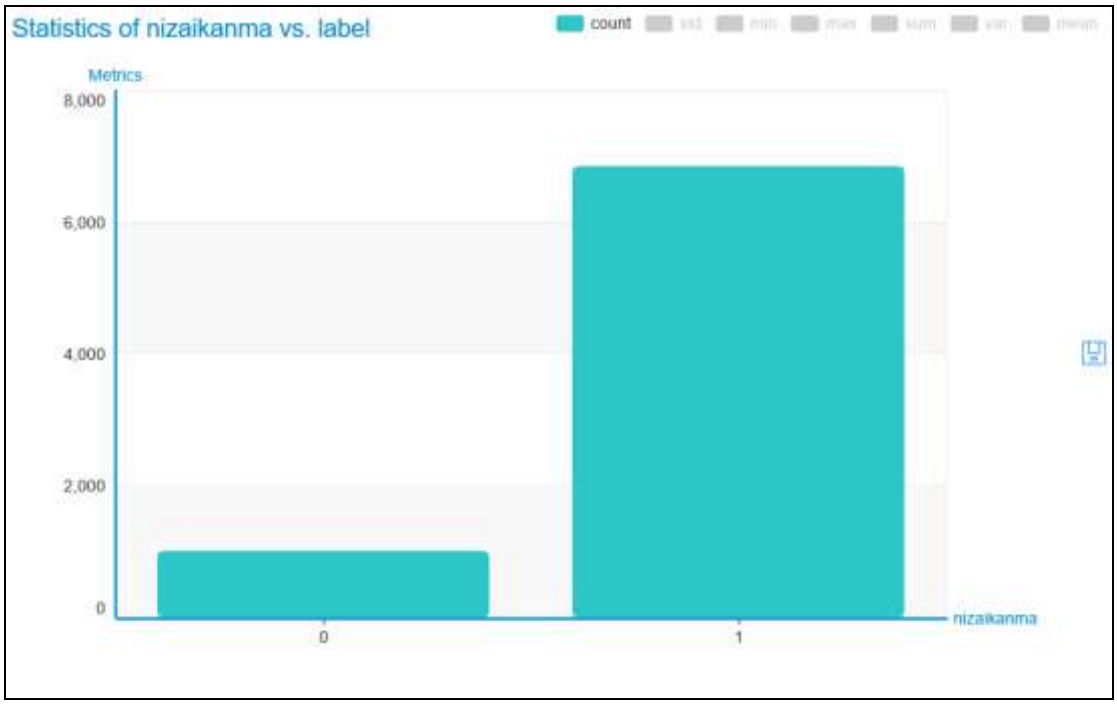
（3）操作次数与消费行为



图三 操作次数与消费行为

关于操作次数这一指标的分析，我们所主要关注的对象也是已经进行过二次购买的用户。在图三中，横轴代表用户在消费前关于此项消费行为所进行的操作次数，纵轴代表不同操作次数的用户的数量所占权重。由图我们可明显看出，操作次数和用户是否会进行再次消费呈一维拟合曲线。分析操作次数数值我们会发现，大多数用户在操作 50~100 次后会产生消费行为，并不是操作次数越多，用户越可能消费。用户在该 APP 上的操作次数是评判用户对该消费是否感兴趣的一重要指标，操作次数太少可能是用户对这种消费已经有所了解或者兴趣较低，相反的，操作次数过多则可能是用户因为某些原因犹豫不定而造成的。所以，我们可认为操作次数在 50~100 次的用户将会消费的可能性更大。

（4）用户在线率



图四 用户在线率

根据上文中所提到的用户浏览时间的长短代表着用户对某种消费行为的感兴趣程度，因此我们从另一个时间角度来考虑用户在此 APP 上的消费行为。我们选取上午十二点作为一个时间节点，根据埋点被抓获的时间我们可知用户是不是浏览使用该 APP。购买的前提是先浏览，如上图四我们可发现，在十二点大部分使用该 APP 的用户产生了消费行为。因此，我们可认为，使用该 APP 频率越高，即在线率越高的用户发生消费行为的可能性越大。

5.2 问题二的模型建立与求解

所谓机器学习 AutoML，简单来说是指用户给出可能的 raw 数据（但是经过标注），机器自动完成整个特征化，模型训练，模型选择/组合，超参数调优，部署上线等环节，高效率地解决算法问题。

5.1.1 变化因素分析

（1）埋点类别变化

不同的用户有不同的行为习惯、行为偏好，同时不同用户从事相关活动的时间、顺序各不相同，同一用户在不同时间段从事同一活动的行为也有差异。用户

使用 APP 时会产生不同操作，就会有不同的埋点被抓获，我们所知道的共有 f1-f122，122 个埋点类型。基于用户在一个周期（十分钟）内的不同操作，我们确定了用户在该 APP 上的开始操作和退出操作。

Out[41]:

f5	5433
f60	1941
f75	1791
f2	393
f29	204
f14	161
f31	104
f6	86
f66	84
f7	67
f13	52
f3	46
f120	38
f44	19
f10	18
f82	18
f62	14
f1	13
f28	12
f8	9
f17	9
f38	9
f63	6
f32	6
f16	6
f49	5
f22	4
f33	4
f34	4
f26	3
f78	3
f72	2
f117	2
f57	2
f51	2
f85	1
f105	1
f84	1
f47	1

表一 各埋点被抓获次数

Out[42]:

f9	6710
f1	1644
f2	575
f42	408
f6	201
f13	116
f3	101
f11	82
f87	79
f4	63
f8	54
f31	40
f91	37
f29	35
f10	31
f117	25
f18	23
f72	22
f40	20
f24	18
f77	17
f44	16
f62	16
f49	15
f76	15
f64	14
f25	13
f55	13
f28	13
f56	12
...	
f96	3
f19	2
f115	2
f66	2
f90	2
f83	2
f41	2
f103	2

表二 各埋点被抓获次数

我们通过对时序埋点数据的正序排序并删除重复项后的数据表示每个 apply_id 所对应的用户第一次进行的操作。从表一我们可以看出 f5 所占比例很高且达到 50%以上，所以我们可假设埋点类型 f5 为开始的操作。

然后通过对时序埋点数据的倒序排序并删除重复项后的数据表示每个apply_id所对应的用户最后一次进行的操作。从表二我们可以看出f9所占比例很高达到60%以上，所以我们亦可假设埋点类型f9为退出的操作。

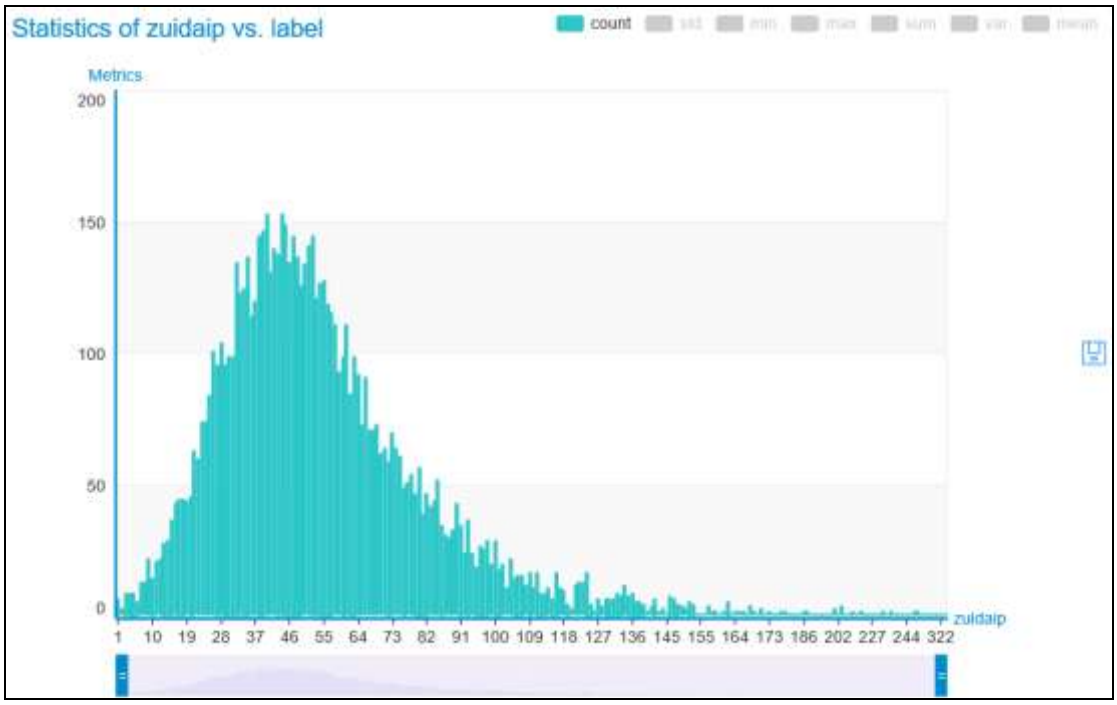
我们可以认为用户特征是可以用用户使用按键的频数之差来表示的，我们在上文中已经提到，用户 F5 以及 F9 的节点数指标性的显示了用户的使用行为，也可以体现用户的某些逻辑特性。

我们认为F5按键使用多于F9按键的用户为正在使用APP的用户，并且通过数值表分析，我们显然可以得出在正在使用 APP 的用户中，用户再次购买的概率约为 90%，而对于大数据分析，用户再次购买的概率约为 75%左右，显然用户的这一行为特征，积极影响了用户的购买可能。

同时，用户是否正在使用也体现了用户使用 APP 频率的强弱，它可以近似的指代用户使用模型数据中习惯数据的一部分，也间接性的缓和了单纯频率曲线的粗糙程度，增加了拟合度，提高了用户行为分析的可靠性，实现了用户行为的客观分析以及数据的理论表示。

（2）数据应用场景变化

所谓的数据应用场景也就是指用户在不同的 IP 地址下使用数据流量在网上从事相关活动。在日常生活中我们走在哪里都离不开网络，换一个新的环境即有一个新的IP地址。在所给用户行为数据中，IP地址的变化体现了用户不同的行为操作。



图五 最大 IP

随着用户使用 IP 位置的变化，我们认为对于相对 IP 使用较专一的用户，他的 IP 行为更加具有参考性，IP 行为不仅体现了用户的使用习惯，也体现了用户使用中环境的变化率，增加了环境因素对用户购买欲望和购买需求的影响偏差，缓和了因为单纯购买欲望产生的实际偏差与使用环境之间的客观偏差，使一阶线性方程中的一维线性化为一维曲线，提高了模型的容错率，增加了模型的使用范性，延展了模型的生活属性，这符合我们对生活的预期。

5.2.2 模型准备

针对问题一的建模过程，我们加以改进，把大量数据更多地交给机器做处理分析。主要用 excel 将数据进行分类处理，再用 SPSS 进行统计分析。SPSS 采用类似 EXCEL 表格的方式输入与管理数据，数据接口较为通用，能方便的从其他数据库中读入数据。然后利用 Python 的三个模块进行数据的处理与分析。

5.2.3 模型建立

在模型建立方面我们主要使用了 Python 2.7 中的三个模块：Pandas 库、Graphlab Create 和 DataFrame 结构。

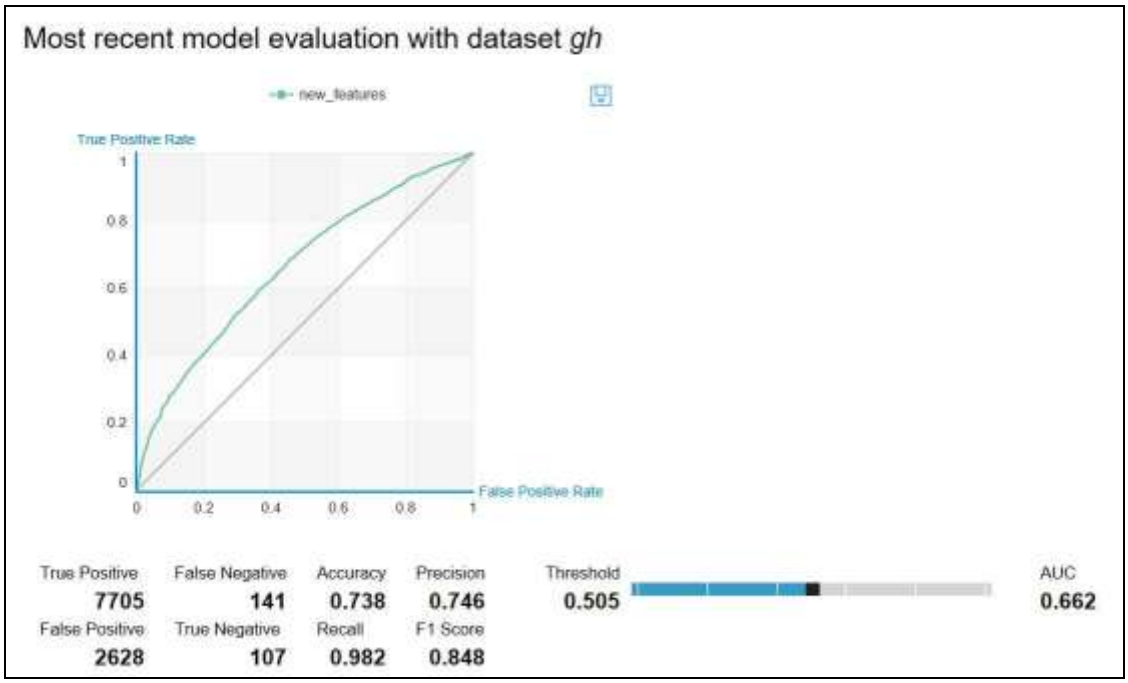
经过 excel 和 SPSS 处理过的数据文件，首先将其导入 Pandas 库，在 Pandas 库里进行数据预处理，清理坏值；接着处理过的数据以 Data Frame 结构导入 Graphlab Create 中，Graphlab Create 会利用线性回归分析的方法再对数据进行处理得到 Gh 对象，形成 list 列表；然后 list 列表中的数据再以 Data Frame 结构导入 Pandas 库中；最终处理完的数据变成 csv 形式。

5.2.4 模型检验

对于现有模型，我们试图建立一种 ROC 曲线以及线性回归之间的对比，所以我们调用了 Python 库中的 pandas 构建了 DataFrame 架构。同时，我们试图可视化我们已有的数据结构方式，于是我们使用了 GraphLab Create，它可以有效地处理传入的大量数据，同时我们使用 Pandas 库中自带的算法，优化了数据，将偏差过大的数据予以删除，将偏差修正到合理范围。

10550	5.3E+07	0.54619	0	0	7165	669	543	2172
10551	5.3E+07	0.85688	1	1	7166	669	543	2172
10552	5.3E+07	0.62261	0	1	7166	669	543	2173
10553	5.3E+07	0.55499	1	0	7166	670	543	2173
10554	5.3E+07	0.82751	0	1	7166	670	543	2174
10555	5.3E+07	0.65811	1	1	7167	670	543	2174
10556	5.3E+07	0.75577	1	1	7168	670	543	2174
10557	5.3E+07	0.62583	0	1	7168	670	543	2175
10558	5.3E+07	0.65742	1	1	7169	670	543	2175
10559	5.3E+07	0.55802	0	0	7169	670	544	2175
10560	5.3E+07	0.70719	0	1	7169	670	544	2176
10561	5.3E+07	0.52768	0	0	7169	670	545	2176
10562	5.3E+07	0.79899	0	1	7169	670	545	2177
10563	5.3E+07	0.59414	0	0	7169	670	546	2177
10564	5.3E+07	0.64044	1	1	7170	670	546	2177
10565	5.3E+07	0.83678	1	1	7171	670	546	2177
10566	5.3E+07	0.77891	0	1	7171	670	546	2178
10567	5.3E+07	0.80579	1	1	7172	670	546	2178
10568	5.3E+07	0.73281	0	1	7172	670	546	2179
10569	5.3E+07	0.77084	1	1	7173	670	546	2179
10570	5.3E+07	0.69737	0	1	7173	670	546	2180
10571	5.3E+07	0.71459	1	1	7174	670	546	2180
10572	5.3E+07	0.8726	1	1	7175	670	546	2180
10573	5.3E+07	0.58821	0	0	7175	670	547	2180
10574	5.3E+07	0.78327	1	1	7176	670	547	2180
10575	5.3E+07	0.80857	1	1	7177	670	547	2180
10576	5.3E+07	0.75629	1	1	7178	670	547	2180
10577	5.3E+07	0.71501	1	1	7179	670	547	2180
10578	5.3E+07	0.6354	0	1	7179	670	547	2181
10579	5.3E+07	0.748	0	1	7179	670	547	2182
10580	5.3E+07	0.52018	1	0	7179	671	547	2182
10581	5.3E+07	0.79058	1	1	7180	671	547	2182
10582	5.3E+07	0.48702	1	0	7180	672	547	2182

图六 部分数据截取



图七 ROC 曲线

Highest Positive Coefficients

(intercept)	0.738
nizaikanma	0.107
caozuocishu	9.000e-4
f9	6.000e-4
Desire	0

Lowest Negative Coefficients

lppinlv	-0.198
f5	-0.003
Time_Order	0

表三 各指标拟合评估

我们先进行了线性回归分析，得到了相应的回归方程分析，但是通过图表观测，我们显然发现，逻辑回归分析更加可视化，准确率也相对提高，于是经过对比，我们采用了逻辑回归。

通过 Graphlab 逻辑回归分析，我们得到了对应的拟合曲线，对于拟合曲线的拟合度，我们使用了 ROC 图的展示形式，同时展示了模型的多个指标，包括准确率等.

六、模型的评价

6.1 模型的优点

（1）该模型充分考虑了不同用户的行为习惯、行为偏好，以及数据应用场景变化、埋点类别变化等因素。

（2）将庞大的数据进行分类预处理，得到了具体数据结果，有较强说服力，较好的解决了数据繁杂的问题。

6.2 模型的缺点

由于真实数据难以搜集全面，数据缺乏的问题使得我们无法对模型进行强有力的支撑与验证，只能通过程序的模拟间接处理。并且用户可能存在操作不规范的情况，导致数据有误。

七、模型的改进与推广

7.1 模型的改进

该模型没有对用户所浏览的商品属性进行过多的考虑，因此在有关商品数据较为充足的情况下可以适当考虑商品属性对用户是否选择消费的影响。

7.2 模型的推广

越来越多的企业已开始挖掘用户行为数据的商业价值，利用行为数据进行精准有效的数字营销。以科技金融行业为例，某知名企业的数据表明：用户行为数据的效力是金融数据的 4 倍，本文则是根据用户 APP 行为数据来预测用户后续的行为，我们可以将其推广至更多的创新型营销企业，并加以改进，吸引更多用户消费，具有很强的现实意义。

八、参考文献

【1】张鹏翼 王丹雪 焦祎凡 陈秀雨 王军，基于用户浏览日志的移动购买预测研究，北京大学信息管理系，2018

【2】吴国华 潘德惠，顾客购买行为影响因素分析及重购概率，东北大学工商管理学院，2005

【3】携程酒店浏览客户流失概率预测 https://cloud.tencent.com/developer/article/1063197

【4】Kaggle WSDM:音乐网站用户流失预测比赛 https://zhuanlan.zhihu.com/p/29598241

【5】https://blog.csdn.net/allwefantasy/article/details/81039505

【6】https://blog.csdn.net/zwqjoy/article/details/85049411

附录

部分程序代码：

```
import graphlab
import pandas as pd
train_Data = graphlab.SFrame(‘C:\User\Administrator\Documents/100w_down.csv’)
my_feature = [‘Desire’ , ‘f5’ , ‘f9’ , ‘nizaikanma’ , ‘Time_order’ , ‘caozuocishu’ , ‘Ippinlv’]
my_feature_model = graphlab.linear_regession.create (Train_Data, target = ‘label’ , features = my_features)
Test_Data = graphlab.SFrame(‘C:\User\Administrator\Documents/dadada.csv’)
Label = my_features_model.perdict (Test_Data)
Label
Labellist = list (Label)
From pandas.core.frame import DataFrame
LABEL = DataFrame (Labellist)
LABEL
LABEL.to_csv (‘over.csv’)
```