第*7*章

## 第7章 用户画像

古语有云，千人千面。世界上的每个人都是唯一的个体，就像人不可能两次跨过同样的一条河一样，世界上也找不到同样的人。从出生起，人就开始有了自己独属的标签，姓甚名谁、父母何许人也、家在哪里、来到世界的时刻、即将开启的人生旅程......经历得越多人的独特性就愈发凸显，即使是一母同胞，总有各自分开的时候，差异也就开始有所产生。孤独是人一生永恒的主题，这是心理状态的一种描述，寻一灵魂伴侣可便可不再孤单，然而在孤字之外更在于有独一无二的独字。由于个体间巨大的差异性和个体的独特性使得分析和研究个体变得十分复杂，甚至没法做到，而且也不太必要。所以从心理学社会学上探讨较多的都是群体特征，偶有个体的行为异常分析也不过是基于事后结果去追根溯源罢了。

即使在互联网信息爆炸到号称大数据人工智能时代的今天，从每个人产生的数字记录中得到的信息依然只是生命中的一部分，没办法完全知道一个人在想什么，未来会做什么。当然也没有必要太过悲观，能记录到的信息如果能够利用好也是可以一定程度上了解每个人。因此基于某个层面的数据便可以做部分的个体画像，从特定角度去形容群体乃至个体的大体差异。说了这么多，这和本章要探讨的用户画像是什么关系呢？首先要讲明用户画像里的用户是谁，数据收集方往往是开发出来某件产品供人使用，这部分使用者便是数据收集方也即产品提供方的用户，产品提供方为了能够推广产品同时也持续维护和改善提供用户体验便需要对用户操作而产生的数据进行挖掘，以期从中发现群体乃至个体的行为偏好，形成数据层面上所谓的画像。更广泛一点，对任何群体都可以进行画像，比如一个地区、一个年代、一个社群等等。

智能手机普及以后，吸引了人们大部分闲暇时间的目光，各式各样的APP产生了大量的用户操作行为的记录数据，这就为用户画像打下了基础，因此基于用户画像的机器学习算法便有了众多的应用场景，此类场景的竞赛也是占据了一部分江山。本章将分为什么是用户画像、用户画像的八要素、常见的数据形式、用户画像的应用以及思考练习五个部分进行用户画像方面内容的介绍。

### 7.1 什么是用户画像

不可否认的是，当人们想到一个人时无论是明面上还是潜意识里其实对这个人都有一个大致印象，比如身材面相，社会属性，性格修养，兴趣爱好等等，而这个脑海里的大致印象也可以说成是一种具有相对主观意识的画像，这当然不是本节要讲到的用于商业分析和数据挖掘的用户画像。

在机器学习中所提到的用户画像，通常是基于给定的数据对用户的属性以及行为进行描述，提取用户的个性化指标，再以此分析可能存在的群体共性，落地应用到各种业务场景中。在互联网时代面向用户的产品多如牛毛，且数据的采集相对容易，由此推动了机器学习在面向用户方面的应用，用户画像便是其中最重要的一环。在各式各样的机器学习算法竞赛中，对于用户的数据挖掘始终占据有一席之地。因此用户画像时常出没在竞赛中，扮演着不可或缺的角色，接下来将向大家详细介绍用户画像的组成以及如何在竞赛中去使用。

### 7.2 用户画像的八要素

由于用户画像广泛的应用，有人给出了如下八要素的概念：

1.基本性：指的是用户角色获取时的应该遵循的基本步骤。是否经过情景访谈，是否有过数据统计和归纳，是否依照具体案例，是否遵循普遍规律等等;

基本性是指数据的获取和采集方式，有哪些信息，是否敏感，是否可信，乃至是否合法，数据是一切分析画像的基础，只有基本性得到良好保障的画像才是科学的。

2.同理性：指的是在设定用户角色画像时，是否经过基于“共情”的讨论，也就是营销人员是否换位在角色角度思考需求问题，画像设计人员是否具备这样的同理心等;

同理性是指画像设计人员能够一定程度地代入用户角色，思考种种行为的产生缘由，以及理解用户，从而能够以同理心进行画像，凸显出普遍规律。

3.真实性：是指被塑造的这个用户角色画像，是否符合真实世界客户的个性，是否可以是一个具备普遍规律的真实人物，是否是凭空捏造，是否脚本太偏离主流等等;

画像能够被得到认可的一切前提便是真实有效。

4.独特性：指的是针对特定产品的营销中，其用户画像是否将目标用户的特点进行有效梳理，是否具备由独特的个性化圈层而引发的限制范围，而不是大一统没有边界的普通自然人。

这里强调的独特性通常是介于大范围群体特性与个体鲜明个性之间，大范围的群体特性由于太过普遍不具有辨识度，而个体鲜明个性又过于细化，因此都不太能够做出具备一定功效的用户画像。独特性便是专注于刻画出某个具有代表性小群体的画像，既具有一定辨识度又不至于太过个性化，通常这样的画像可用性较强。

5.目标性：指的是该用户角色的画像是否包含与产品相关的高层次目标，是否包含关键词来描述该目标等;

最成功的用户画像无疑是能够用产品相关的关键词进行刻画与描述，二者直接关联使得用户画像直击目标，具有较强的匹配度。

6.数量性：指的是用户角色画像的数量级是否能明确，以便营销团队能根据较为明确的数量级梳理相应的推广预算，或各个渠道的推广权重设置;

数量级的明确有助于评估用户画像的落地应用前景以及画像的具体提取方式，通常需要在计算资源成本、效率和精准度之间做出一些取舍，维持不错的平衡。

7.应用性：指的是这个用户角色的最终画像，是否具备在营销体系内对其进行相应的营销策略设计等的应用基础。方向的把控和筛选是否可以执行。

用户画像是个比较广泛的概念，因此有必要贴合实际应用进行有针对性的设计，既可以大大减少工作量，同时也能避免无关信息造成一定程度的干扰。

8.长久性：指向的是这个用户画像标签的保留时限问题，是即用即散型还是同产品的长久契合型。营销的整体概念和设计，需要围绕的是未来同产品长期契合的目标受众，而非临时性消费迷雾导致的过度乐观。

人是随着时间而不断变化的，除了用户的一些具有较长时间保质期的固有属性之外，很多用户画像的标签都可能随时发生变化，因此有必要区分出长短期行为，同时根据需要隔段时间便进行相应地更新以响应最新需求。总结来说，用户画像的八要素准确描述了有关用户标签的生成、采集、整理、迭代和应用的全流程场景特点，特别是在面向用户层面的AI应用场景当中，要保持对用户画像八要素的洞察，避免给建模带来干扰以及丢失重要的用户信息。

### 7.3 常见的数据形式

在各式各样的竞赛当中，数据的形态和格式也是多种多样，本节以用户画像为例将数据的有关字段大致分为数值型变量、类别型变量、多值型变量、以及文本型变量四种常见的数据形式，对每种变量都有对应的处理方式。需要强调的是这些变量都是针对于用户层面的，即所有样本数据是以用户为唯一主键进行区分，且每个用户只有一条记录，之所以这样去举例是因为通常来说基于用户画像的机器学习模型所需的数据是以用户池的形式进行呈现的，对用户的标签进行对应的特征学习。实际竞赛当中给到的数据可能十分复杂，甚至是以打点记录的方式描述了用户的行为，这时候往往还需要参赛者进行用户特征的构建提取，这就涉及到更深的应用技巧了。

#### 7.3.1 数值型变量

数值型变量是指具有数值大小含义的变量，如年龄、身高、体重等等，这是最常见的一种数值变量，其他如消费金额、流量累计等。

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 年龄 | 身高 | 体重 |
| 王贺 | 25 | 170 | 75 |
| 刘鹏 | 18 | 163 | 68 |
| 钱乾 | 26 | 172 | 72 |

#### 7.3.2 类别型变量

类别型变量则是指具有类别标识的变量，如性别、籍贯、所在城市等等，这类变量记录了用户的一个固有属性。

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 性别 | 籍贯 | 所在城市 |
| 王贺 | 男 | 陕西 西安 | 北京 |
| 刘鹏 | 男 | 四川 广安 | 东莞 |
| 钱乾 | 男 | 上海 普陀 | 上海 |

#### 7.3.3 多值型变量

多值型变量是指用户在某个维度具有多个取值的变量，比如兴趣爱好、穿衣风格、看过的电影等等，这类变量由于其特殊结构无法直接应用到模型中，需要特借助特别的数据结构如稀疏矩阵进行处理，举例如下，

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 兴趣爱好 | 穿衣风格 | 看过的电影 |
| 王贺 | 撸串 直播 秀恩爱 | 黑色 黄色 红色 卫衣 | 少年的你 海上钢琴师 |
| 刘鹏 | 健身 摄影 | 黑色 白色 灰色 短袖 | 独自等待 海蒂与爷爷 |
| 钱乾 | 做饭 骑行 篮球 | 衬衣 黑色 | 唐人街探案3 姜子牙 |

#### 7.3.4 文本型变量

文本型变量是指利用文本记录的变量，如用户对某件商品或者某次购物的评价等等，处理这类变量则需要用到自然语言处理的一些工具，如中文分词工具jieba等。

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 商品 | 打分星级 | 评论 |
| 王贺 | 双截棍 | 4 | 质地很精良，手感很好 |
| 刘鹏 | TNT天酷街舞中心成人体验课 | 5 | 非常棒的一次学舞体验，期待全新事物带来的新世界 |
| 钱乾 | 拳击手套 | 3 | 戴上去有点紧，表皮有股淡淡的油脂味儿 |

### 7.4 用户画像的应用

用户画像之所以这么值得去研究学习，是因为其广阔的应用场景，且互联网时代用户的行为会产生大量的可供分析建模的数据，这也为用户画像提供了良好的条件，本节将从精准营销、用户分析、战略规划、风险控制等四个主要层面进行用户画像应用方面的介绍。

#### 7.4.1 精准营销

智能手机下载量前十的APP中总少不了电商的身影，网购的方便快捷以及商品的琳琅满目极大增强了用户的购物体验。利用用户的历史消费行为进行画像，彰显出用户的消费偏好，使得在用户需要添置物品的时候电商平台能够迅速准确地响应用户的需求，同时商家也能容易地找到自家的种子用户，除了电商搜索外，广告投放也属于精准营销的范畴。在机器学习的各大竞赛平台上，有关面向用户精准营销的比赛层出不穷，以下对用户画像在精准营销方面的算法竞赛实例进行简单介绍。

本书王、刘二位作者竞赛生涯的首冠，即DC竞赛平台上举行的《2018科大讯飞AI营销算法大赛》。主办方是讯飞AI营销云，它隶属于科大讯飞股份有限公司，基于深耕多年的人工智能技术和大数据积累，赋予营销智慧创新的大脑，以健全的产品矩阵和全方位的服务，帮助广告主用AI实现营销效能的全面提升，打造数字营销新生态。在讯飞AI营销云在高速发展的同时，积累了海量的广告数据和用户数据，如何有效利用这些数据去预测用户的广告点击概率，是大数据应用在精准营销中的关键问题，也是所有智能营销平台必须具备的核心技术。这次大赛提供了讯飞AI营销云的海量广告投放数据，参赛选手需要通过人工智能技术构建预测模型预估用户的广告点击概率，即给定广告点击相关的广告、媒体、用户、上下文内容等信息的条件下预测广告点击概率。虽然这场竞赛也属于广告投放领域，但是二位作者通过对用户画像的挖掘，准确地将用户与广告联系在一起，取得了不错的成绩。

#### 7.4.2 用户分析

日活、月活是时常会听到的两个词语，通常是形容每天或者每月的在线活跃人数，是互联网产品的重要评价指标。虽然每个产品在设计之初都或多或少有定位目标用户群，但上线之后总会面临各种各样的挑战，拉新、促活、留存每样都是难题，因此需要不停地做用户画像分析，提炼人群特点，继而不断优化产品的功能、UI交互。

#### 7.4.3 战略规划

战略规划这词听起来怪唬人的，其实就是想凭借不同年龄段用户的特点，以及时代的科技、政策发展背景进行一些策略的规划与调整，避免与时代脱节。特别在互联网领域，各大网络平台如电商、直播、视频等门户，尤其关注其现有用户的使用特点，以及潜在用户的画像素描，以保证在维持现有用户量的情况下，吸引更多用户使用并产生更多商业价值。

#### 7.4.4 风险控制

除了前面三节讲到的用户画像之外，还有一类特殊的应用场景，那就是通常被用在金融领域的风险控制。这类场景主要关注于用户的经济状况，结合征信等维度评估用户的偿贷能力。在移动支付盛行的时代，电子流水取代了现金流通，也使得人们的各种消费行为记录能够被自动记录，这就产生了大量交易流水，为做消费者的消费能力和信用评估提供了支撑。以DF竞赛平台的《消费者人群画像-信用智能评分》竞赛为例，中国移动福建公司提供2018年x月份的样本数据（脱敏），包括客户的各类通信支出、欠费情况、出行情况、消费场所、社交、个人兴趣等丰富的多维度数据，希望参赛者通过分析建模，运用机器学习和深度学习算法，准确评估用户消费信用分值。

关于这场比赛，主办方提供的赛题背景如下：随着社会信用体系建设的深入推进, 社会信用标准建设飞速发展，相关的标准相继发布，包括信用服务标准、信用数据釆集和服务标准、信用修复标准、城市信用标准、行业信用标准等在内的多层次标准体系亟待出台，社会信用标准体系有望快速推进。社会各行业信用服务机构深度参与广告、政务、涉金融、共享单车、旅游、重大投资项目、教育、环保以及社会信用体系建设，社会信用体系建设是个系统工程，通讯运营商作为社会企业中不可缺少的部分 同样需要打造企业信用评分体系，助推整个社会的信用体系升级。同时国家也鼓励推进第三方信用服务机构与政府数据交换，以增强政府公共信用信息中心的核心竞争力。传统的信用评分主要以客户消费能力等少数的维度来衡量，难以全面、客观、及时的反映客户的信用。中国移动作为通信运营商拥有海量、广泛、高质量、高时效的数据，如何基于丰富的大数据对客户进行智能评分是中国移动和新大陆科技集团目前攻关的难题。运营商信用智能评分体系的建立不仅能完善社会信用体系，同时也中国移动内部提供了丰富的应用价值，包括全球通客户服务品质的提升、客户欠费额度的信用控制、根据信用等级享受各类业务优惠等，希望通过本次建模比赛，征集优秀的模型体系，准确评估用户信用分值。

由此可见，对于用户消费行为的画像，能够帮助对用户的信用进行打分评估，从而支撑政府、企业和其他社会机构进行相关决策，具有重要意义。

### 7.5 思考练习

1、请就数值型变量、类别型变量、多值型变量以及文本型变量分别举五个常见的例子。

2、你觉得用户画像是想体现用户的共性还是个性，为什么？

3、就你日常所使用的APP，思考其算法与运营团队会如何给你画像呢？