BDCI2018 套餐的个性化匹配

赛题名称

面向电信行业存量用户的智能套餐个性化匹配模型

赛题背景

电信产业作为国家基础产业之一，覆盖广、用户多，在支撑国家建设和发展方面尤为重要。随着互联网技术的快速发展和普及，用户消耗的流量也成井喷态势，近年来，电信运营商推出大量的电信套餐用以满足用户的差异化需求，面对种类繁多的套餐，如何选择最合适的一款对于运营商和用户来说都至关重要，尤其是在电信市场增速放缓，存量用户争夺愈发激烈的大背景下。针对电信套餐的个性化推荐问题，通过数据挖掘技术构建了基于用户消费行为的电信套餐个性化推荐模型，根据用户业务行为画像结果，分析出用户消费习惯及偏好，匹配用户最合适的套餐，提升用户感知，带动用户需求，从而达到用户价值提升的目标。

套餐的个性化推荐，能够在信息过载的环境中帮助用户发现合适套餐，也能将合适套餐信息推送给用户。解决的问题有两个：信息过载问题和用户无目的搜索问题。各种套餐满足了用户有明确目的时的主动查找需求，而个性化推荐能够在用户没有明确目的的时候帮助他们发现感兴趣的新内容。

赛题任务

此题利用已有的用户属性(如个人基本信息、用户画像信息等)、终端属性(如终端品牌等)、业务属性、消费习惯及偏好匹配用户最合适的套餐，对用户进行推送，完成后续个性化服务。

赛题方案

1 数据探索与评价指标

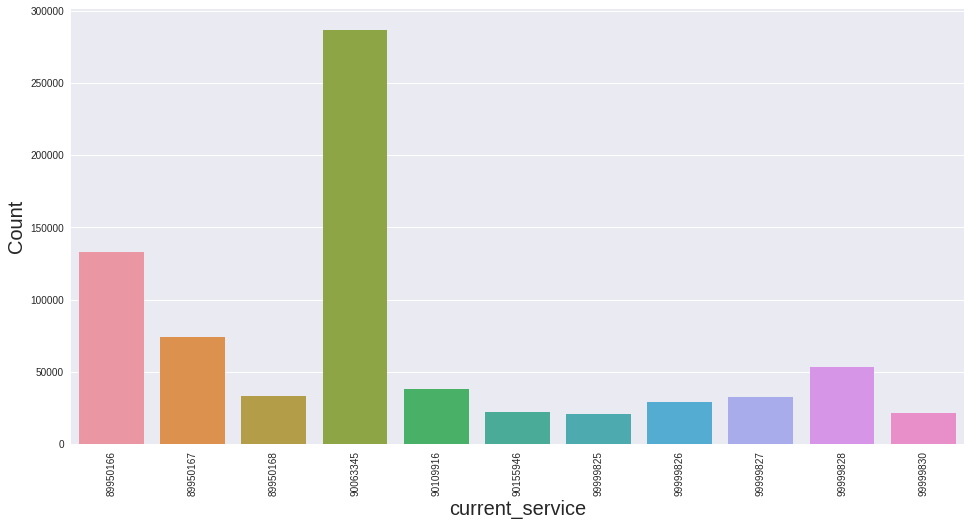


图1：类别和类别的计数观察

我们可以看到此问题是一个多分类问题，对应着有11种套餐，其中他们的分布较为不均匀。作为一个典型的多分类为题，其评价指标为：

首先针对每个套餐类别，统计分别统计（预测答案正确），（错将其他类预测为本类），（本类标签预测为其他类标）。）通过第一步的统计值计算每个类别下的precision和recall，计算公式如下：

计算结果计算每个类别下的,计算方式如下

然后通过第三步求得的各个类别下的 求均值，得到最后的评测结果，计算方式如下：

以下是关于原始特征和标签相关性的观察图表：

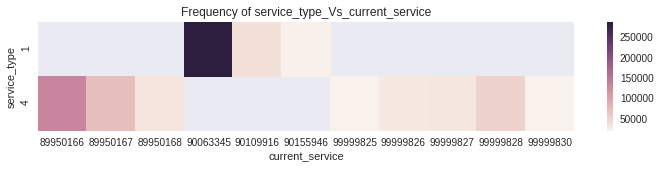


图2：service\_type和类别的相关性观察

我们可以得到一个明显的规律，如图2所示，service\_type可以讲套餐分为两个部分，这两部分是没有交叉的，其中一类有8个，另外一类有3个。这给我们比赛带来一个思路是，可以分模型预测，在最后的实验中我们也通过此取得了提升。

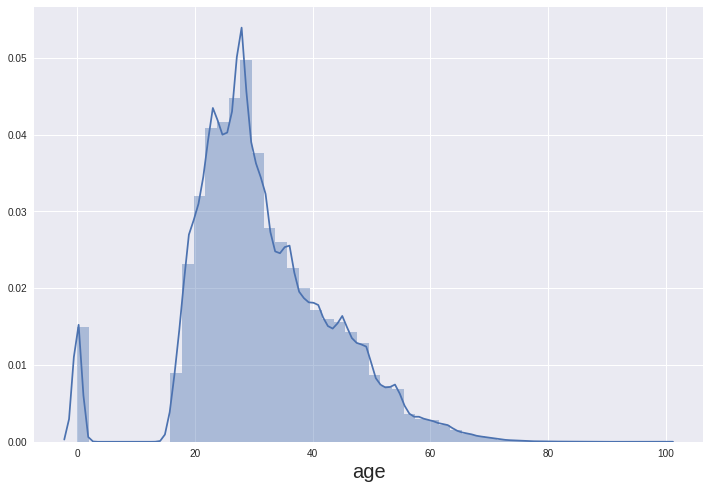


图3：年龄的分布

我们可以看到对于年龄来讲，基本上符合电信用户群体的分布，但是有很多0岁的异常值，对于异常值，我们尝试了替换service\_typpe字段对应的均值和使用原始值两种方法。最终选取了原始值，我们认为默认年龄在不同套餐中的转换率呈现了分布差异。

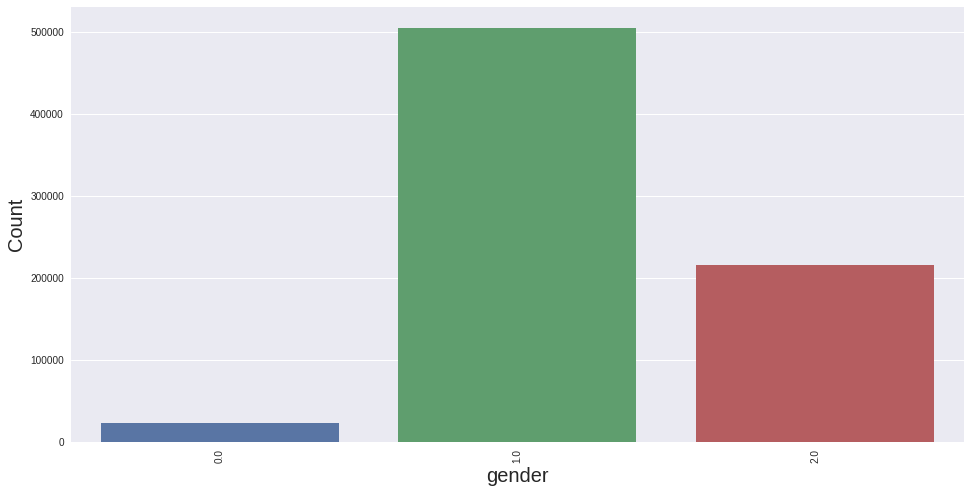


图3：性别的分布

我们观察到性别中有0 的缺省值，对于这部分，我们使用了两种方法处理，一种是填充service\_type对应字段的众数，和原始值。最终我们选取了原始值，我们认为默性别在不同套餐中的转换率呈现了分布差异。

同时我们注意观察训练集和测试集是否同分布：

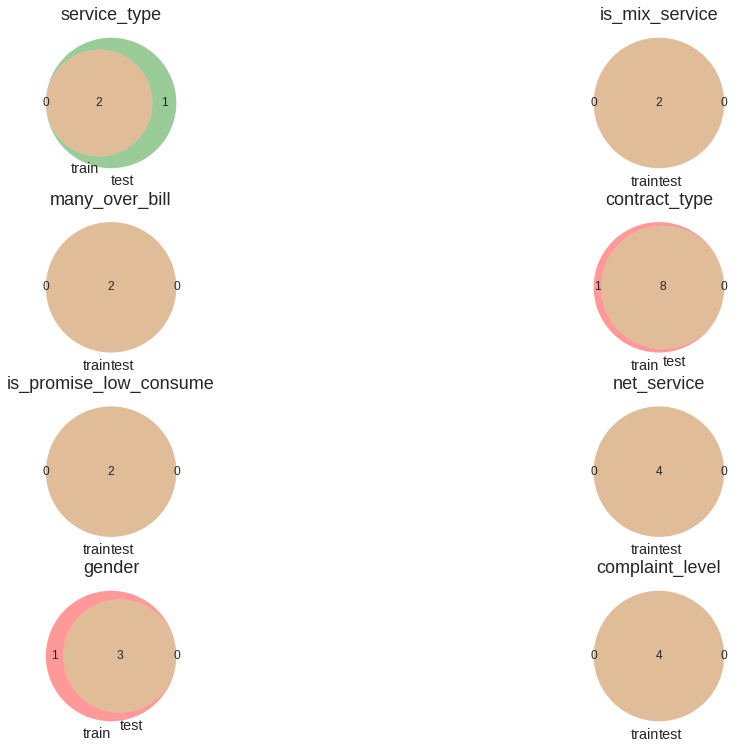


图4：同分布观察

我们观察到gender字段出现了不同分布的情况的，我们采取了归一化处理的方法，将因为数据清洗质量造成的不同分布消除。

2 特征工程

1 关联规则

关联规则是形如的蕴涵式，其中， 和分别称为关联规则的先导(或, )和后继(或, 。其中，关联规则，存在支持度和信任度。

有一个注明的笑话就是，在一家超市里，有一个有趣的现象：尿布和啤酒赫然摆在一起出售。但是这个奇怪的举措却使尿布和啤酒的销量双双增加了。这不是一个笑话，而是发生在美国沃尔玛连锁店超市的真实案例，并一直为商家所津津乐道。受max-encoding的方法的启发，我们对有关联关系的话费字段，1\_total\_fee, 2\_total\_fee, 3\_total\_fee, 4\_total\_fee四个字段的构建关联规则，这样可以利用整体信息，对话费字段这个强特征进行降维编码。使得其数据更加具有代表性。这种特征对结果也有很强的正反馈。我们提取了话费的最频繁二项集，把一月的话费对应的话费作为特征使用。

其实这里有更好的方式，对于对偶形式的数据也更通用，大家可以结合DeepWalk，nove2vec，Line，SDNE的思路想一下。

2 业务特征

业务特征的部分，我们深入研究了联通的套餐消费场景，从比赛的一开始，首先就通过联通官网以及消费论坛认真的开始调研工作，深入了解了联通的各种套餐特点和用户群差别。通过熟悉套餐的特性，我们可以为各种特定用户群推荐适宜他们的套餐，比如腾讯天王卡玩腾讯游戏看腾讯视频不花钱是深度腾讯用户的福音，蚂蚁大宝卡则可以赠送2g无差别流量给高流量消费者。联通传统套餐的各种优惠活动，比如预充值冲100返流量和话费，适合平时那些流量和话费不够用的用户，充值返话费则适合那些薅羊毛的用户。对此我们针对用户的流量和通话做了一系列特征，比例、差值，求和等，力求尽可能的描绘出一幅用户画像。

提出了以下几种针对业务的特征：

1. 话费减去16元是否是整数
2. 流量的有效数字是否是27的整数倍
3. 话费的有效数字能否被15乘除
4. 话费是否是整数（用户可能未超套餐）
5. 连续两个月套餐的差值能否被5,10,15,27，30等计费单元整除
6. 四个月话费的最小值
7. 计算流量的平均单价
8. 计算通话时间的平均单价
9. 等等…

3 引导学习

在机器学习中，针对不同阶段的同源数据，通常有迁移学习和直接拼接数据的方法来利用数据，可以利用不同分布的数据提升结果。迁移学习直接对预测概率编码为被迁移模型的特征，可以保证同分布，但是不能充分利用数据；拼接充分利用数据，但是出来的结果不能同分布；为此，我们提出一种称为“**引导学习**”的方法。

结构化数据的迁移学习方法：用不同分布的数据训练得到模型，预测出来的概率作为新任务的特征编码。与一般迁移学习不同的是，嫁接学习强调利用因时间演化或者采样造成的分布产生变化的同源数据。因为被一名植物分类学家在比赛中最先使用，因此也被戏称为“**嫁接学习**”。

在比赛过程中我们对各种数据迁移方法进行了比较，首先是全部初赛数据的迁移学习，但是效果并不显著。于是我们考虑增加复赛数据一起训练作为**贝叶斯正则项**，效果非常好。我们更深入的研究复赛数据的添加方式。如下:

1.直接将初赛和复赛数据拼接

2.拼接后将初复赛数据增加一列特征标注是否为复赛数据

3.拼接后将网络套餐和传统套餐分别训练。

通过训练的结果可以看到：直接得到的结果里第2项最好，因为它通过特征标注保证了在这个模型里初赛和复赛数据有各自的分布；将输出的结果用于迁移则第1项最好，因为这个模型更加关注数据本身的影响而不考虑一同训练两组数据不同。实验结果如下表所示：

表2：各种迁移学习方法对比

|  |  |
| --- | --- |
| 迁移方法 | Leaderboard分数 |
| 原始拼接 | 0.8327 |
| 标注拼接 | 0.8397 |
| 拼接分类训练 | 0.8385 |

其中验证集只选用与目标同源的数据，这也是**贝叶斯正则项**的另一来源。我们把加了**贝叶斯正则项**的方法称为**引导学习**，在比赛结果中效果显著。