使用人工智能（AI）挖掘数字金矿极简版案例

目 录

一、人工智能 AI 应用极简版案例：预测流失用户

关于“流失用户”的定义

有一些“流失用户”，或其它分类的用户，是可以确凿定义的，譬如贷款客户中的逾期未归还贷款客户，只需明确在合约期的最后一天是否已经归还贷款就可以定义了。

但是，快消零售行业或者电信用户这一类，就不方便归纳哪些用户是已经流失的。通常对于没有明确依据断定是否流失的情况，可以定义持续 n 个周期没有发生消费行为的用户，就认定为流失用户。

按照这个定义，以某快消品牌公司数据为例，我们假设超过 90 天没有发生消费行为的会员，认定为流失用户。则数据见下图：

CSTMSEQ	tracou	CSTMGRADECD	SEX	Age	tar
TW10075763	3	TWL0004	F	36	0
TW10114219	3	TWS0012	F	35	0
TW10321472	3	TWL0004	F	44	0
TW10123649	3	TWS0013	F	39	0
TW10355274	3	TWS0012	F	27	0
TW10317839	6	TWL0003	F	28	0
TW10358193	5	TWL0004	F	40	0
TW10682224	3	TWL0004	F	38	1
TW10357367	8	TWS0013	F	40	0
TW10668502	5	TWL0004	F	27	0

elect * from · 只读 查询时间: 66.922s 第 1 条记录 (共 25465 条)

备注：

1、 第一列 CSTMSEQ 是会员顺序号

2、 第二列 tracou 是交易次数（含领取权益物交易金额为零的次数）

3、 第三列 CSTMGRADECD 是会员等级

4、 第四列 SEX 是性别（因为几乎值全部为 F，所以在机器学习中舍弃）

5、 第五列 Age 是会员年龄

6、 第六列 tar 是标记列，即当持续未消费时间超过 90 天，值为 1（代表已流失），未超过 90 天则值为 0（代表未流失）

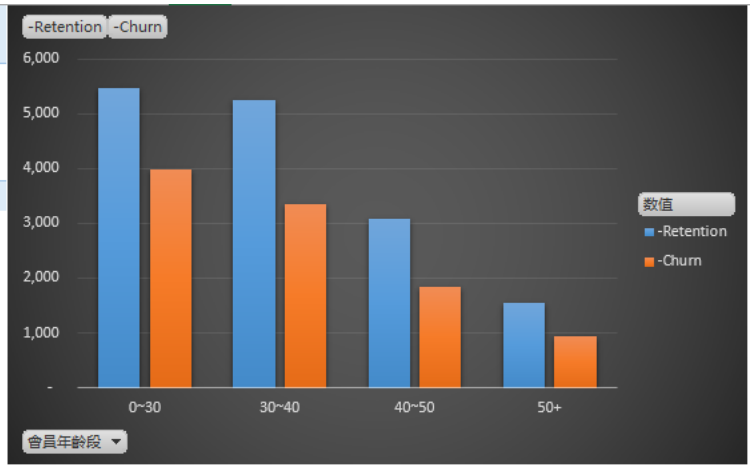
7、 该快消品牌公司的数据表中的大量字段数据为空，所以用于机器学习的特征值比较缺乏，目前只能采集到上述字段的数据。

二、传统方法观测用户流失概况

传统方法对（潜在）流失用户进行分析，一般会使用统计分类方法，譬如下图的分类计算：

按年龄段划分

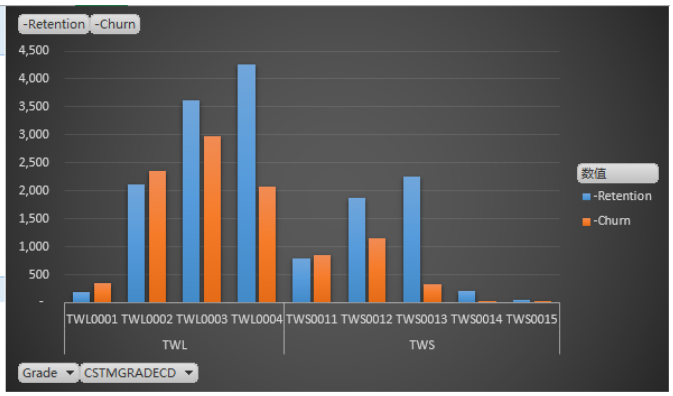
值		
會員年齡段	-Retention	-Churn
0~30	5,456	3,973
30~40	5,244	3,340
40~50	3,090	1,852
50+	1,558	952
总计	**15,348**	**10,117**



上图将年龄按照分组区段划分，划分了四个分组：0~30、30~40、40~50、50+。从图中发现各年龄段的留存用户与流失用户的比例大体相当。

按会员等级划分

Grade	CSTMGRADECD	值 -Retention	-Churn
TWL	TWL0001	180	345
	TWL0002	2,123	2,357
	TWL0003	3,609	2,978
	TWL0004	4,259	2,066
TWS	TWS0011	800	846
	TWS0012	1,865	1,153
	TWS0013	2,263	326
	TWS0014	201	12
	TWS0015	48	34
总计		15,348	10,117



从图中发现 TWL0001、TWL0002 和 TWS0011 这三个会员等级的流失用户，超过了本等级的留存用户。并且 TWL0003 的流失用户比较接近留存用户。

三、机器学习（深度学习）预测工具与结论

机器学习方法对上述数据进行流失用户预测，可以使用数据分析工具 Python、机器学习包 sklearn 和深度学习包 tflearn。

1、使用 sklearn 进行预测结果如下：

```
1  from sklearn.metrics import classification_report
2  print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.74	0.64	0.68	3100
1	0.54	0.59	0.56	1993
micro avg	0.65	0.62	0.63	5093
macro avg	0.64	0.61	0.62	5093
weighted avg	0.66	0.62	0.64	5093
samples avg	0.62	0.62	0.62	5093

解读：

- ✓ 若按百分比考察，将 100 个用户预测为未流失用户，其中只有 74 个用户预测正确，另外 26 个用户被预测为未流失，但实际上流失了；
- ✓ 若按百分比考察，将 100 个用户预测为流失用户，其中只有 54 个用户预测正确，另外 46 个用户被预测为流失，但实际上并未流失；
- ✓ 实际一共有 3100 个用户未流失，正确预测为未流失的比率为 64%；
- ✓ 实际一共有 1993 个用户流失，正确越策为流失的比率为 59%；
- ✓ 结果表明预测结果不理想，需要进一步调参优化。

2、使用 tflearn 进行深度学习后预测结果如下：

```
score = model.evaluate(X_test, y_test)
print('Test accuarcy: %0.4f%%' % (score[0] * 100))
```

```
Test accuarcy: 65.7177%
```

表示在测试集上的准确性从之前的 54%提到到了 65.72%，但还是有较大的提升空间。

上述预测的精确率偏低，主要原因在于提供的特征较少，意味着我们需要获取更多会员特征的信息。

四、机器学习（深度学习）预测原理与步骤

机器学习是指让机器通过自动学习数据的规律，并生成一个模型，然后根据这个模型做出判断（预测）。譬如，若希望让机器识别出“狗”，于是就给机器阅读大量的狗的图片信息，并且告诉机器，这个图片中的样子的就是狗。看的图片多了，机器就会知道图片中的模样的东西就代表狗。当再给机器看一张类似的图片的时候，不用告诉机器这个是狗，机器也可以自己做出判断：这就是狗。

机器是根据什么来把握规律的呢？是依靠特征。这里的特征譬如是有两个椭圆形的形状（耳朵）、有四个小柱子（四条腿）、长长的一块区域（身躯）、长长区域的一边是一根线条（尾巴），另一边是一小块不规则形状的区域（头部）……

所以，第一，机器学习最大需求就是需要大量的数据（上述例子就是图片）用来训练。训练的数据越多，判断就越准确；

第二，特征不能过于明确。譬如原本是希望机器能预测出“狗”，但给出的图片全部都是“二哈”的图片（也就是“特征”非常明确地锁定在二哈身上），当给出一只“金毛”的图片时，机器看着不像原来看到的（二哈）的样子，所以就会做出判断：这不是狗。

第三，特征也不能太少。譬如用户流失预测中，假如提供的特征只有性别，机器就只能完全依靠“性别”来判断这个用户会不会流失，这样肯定就会不准确了。所以增加特征，譬如客户的年龄、收入、家庭子女数、距离门店远近、兴趣爱好、工作类别、经常参加什么运动等等。这样才会越来越准确。

机器学习的步骤，首先是原始数据处理。包括对每个字段（属性）做判断，是否属于“结论（狗）”的特征，是就保留，不是就舍弃。每个特征之间需要保持独立性，比如身高和体重之间，就缺乏独立性，需要处理；

第二步，数据标识。对所有数据给出标记，就好像是每一副图片做标记：是狗（标记为 1），不是狗（标记为 0），这样机器在看到图片以及标记之后，就会归纳出结论什么样

子是狗，什么样子不是狗。

第三步，将数据划分为训练数据和测试数据。就好像对待学生一样，训练数据就是先做习题，测试数据就是考试。然后针对测试数据，将机器做出的判断和测试数据原本的标记进行比对，来计算器器预测的精确率。

第四步，如果精确率满足需求，使用模型对需要预测的数据进行预测，比如精确到 98%地预测出哪些用户将会流失。

五、后记

在使用大数据进行机器学习和深度学习过程中，大量的顾客属性数据和交易数据是必不可少的。在这方面，支付宝的做法可以借鉴：在支付宝 APP 中，用户补充哪些数据信息，可以分别获得多少积分；用户上传了哪些证件，又可以分别获得多少积分……等等，就是用于鼓励用户尽可能填写真实的可验证的信息。

涉足 AI 服务，既可以为客户提供更多高附加值服务，又可以在未来若是有融资需求的时候，突出“数字资产”。数字资产是一项非常有价值的资产，有助于提高企业估值。