用户流失画像

在这次实例中我们根据中国电信的脱敏的真实数据，通过数据分析和建立数学模型，来建立用户流失画像，判断怎样的用户流失的规律

概述

本次案例包括以下三个部分

- 数据导入及预处理
- 数据建模（采用决策树模型）
- 总结和思考

一、数据导入及预处理

1.数据导入：

文件为‘CustomerSurvival.csv‘，我们将数据导入到Python中：

```
# 导入分析用到的模块
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('ggplot')
import seaborn as sns
sns.set_style('darkgrid')
sns.set_palette('muted')
# 导入csv文件
df = pd.read_csv('F:/CustomerSurvival.csv',encoding='utf-8')
```

2.数据理解及变量解释

变量含义:

> ***id*** *– 用户的唯一标识*
> ***pack_type*** *– 用户的月套餐的金额，1为96元以下，2为96到225元，3为225元以上*
> ***extra_time*** *– 用户在使用期间的每月额外通话时长，这部分需要用户额外交费。数值是每月的额外通话时长的平均值,单位：分钟*
> ***extra_flow*** *– 用户在使用期间的每月额外流量，这部分需要用户额外交费。数值是每月的额外流量的平均值，单位：兆*
> ***pack_change*** *– 是否曾经改变过套餐金额，1=是，0=否*
> ***contract*** *– 用户是否与联通签订过服务合约，1=是，0=否*
> ***asso_pur*** *– 用户在使用联通移动服务过程中是否还同时办理其他业务，1=同时办理一项其他业务，2=同时办理两项其他业务，0=没有办理其他业务*
> ***group_use*** *– 用户办理的是否是集团业务，相比个人业务，集体办理的号码在集团内拨打有一定优惠。1=是，0=否*
> ***use_month*** *– 截止到观测期结束（2012.1-2014.1），用户使用联通服务的时间长短，单位：月*
> *loss – 在25个月的观测期内，用户是否已经流失。1=是，0=否*

3.自变量与因变量之间的关系：

对于extra_time和extra_flow绘制散点图观察：

```
plt.figure(figsize = (10,6))
df.plot.scatter(x='extra_time',y='loss')
df.plot.scatter(x='extra_flow',y='loss')
```

接着看其他自变量与流失的关系：

```
fig,axes = plt.subplots(nrows = 2,ncols = 3, figsize = (12,8))
sns.countplot(x = 'pack_type',hue = 'loss',data =df,ax = axes[0][0])
sns.countplot(x = 'pack_change',hue = 'loss',data =df,ax = axes[0][1])
sns.countplot(x = 'contract',hue = 'loss',data =df,ax = axes[0][2])
sns.countplot(x = 'asso_pur',hue = 'loss',data =df,ax = axes[1][0])
sns.countplot(x = 'group_user',hue = 'loss',data =df,ax = axes[1][1])
```

 在这里插入图片描述
初步得出以下结论： 1）.套餐金额越大，用户越不易流失，套餐金额大的用户忠诚度也高

2）.改过套餐的用户流失的概率变小

3）.签订过合约的流失比例较小，签订合约也意味着一段时间内（比如2年，3年）用户一般都不会更换运营商号码，可以说签订合约的用户比较稳定

4）.办理过其它套餐业务的用户因样本量太少，后续再研究

5）.集团用户的流失率相比个人用户低很多

二、数据建模

因为自变量大多数为分类型，所以用决策树的效果比较好，而且决策树对异常值的敏感度很低，生成的结果也有很好的解释性。

1.特征的预处理

根据前面的探索性分析，并基于业务理解，我们决定筛选这几个特征进入模型： extra_time，extra_flow，pack_type, pack_change, asso_pur contract以及group_use，这些特征都对是否流失有一定的影响。 对于extra_time，extra_flow这两个连续型变量我们作数据转换，变成二分类变量，这样所有特征都是统一的度量。

```
df['time_tranf'] = df.apply(lambda x:1 if x.extra_time>0 else 0,axis =1)
df['flow_tranf'] = df.apply(lambda x:1 if x.extra_flow>0 else 0,axis =1)
```

将没有超出套餐的通话时间和流量记为0，超出的记为1。

2. 建立自变量x, 因变量y的二维数组：

```
x = df.loc[:,['pack_type','time_tranf','flow_tranf','pack_change','contract','asso_pur','group_user']]
x = np.array(x)
y = df.loss
y = y[:, np.newaxis]
```

3. 拆分训练集和测试集，比例为7:3

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
```

4. 建立决策树模型并拟合训练：

```
from sklearn import tree
clf = tree.DecisionTreeClassifier(criterion='gini', //设置衡量的系数
                                  splitter='best', //选择分类的策略
                                  max_depth=5, //设置树的最大深度
                                  min_samples_split=10,//节点的最少样本数
                                  min_samples_leaf=5 //叶节点的最少样本数
                                  )
clf = clf.fit(x_train,y_train)  -- 拟合训练
```

这里我们采用决策树中CART算法，基于gini系数进行分类，设置树的最大深度为5，区分一个内部节点需要的最少的样本数为10，一个叶节点所需要的最小样本数为5。 决策树最大的缺点是容易出现过拟合，所以我们先看一下模型对于训练数据和测试数据两者的评分情况：

```
train_score = clf.score(x_train,y_train) # 训练集的评分
test_score = clf.score(x_test,y_test)   # 测试集的评分
'train_score:{0},test_score:{1}'.format(train_score,test_score)
```

可以看到针对训练集评分为0.867，针对测试集评分为0.880，两者近乎相等，说明模型较好的拟合训练集与测试集数据。

三、总结和思考

从数据的探索性分析中我们可以看出，运营商关注的重点应该放在那些高价值的用户上，比如使用通话和流量比较多的，套餐金额比较大的这些用户。应采取相应的运营策略预防其流失，并且可以分析用户流失的主要原因，是优惠福利不满意还是竞争对手在某些方面比自己有优势。在这个通讯行业激烈竞争的时代，且手机用户数量已基本饱和，维护老用户比获取新用户更容易。 在预测模型的优化上，还有很多的改进之处，比如调整决策树的参数，特征的精细化筛选，或者采用多种算法进行模型评估。而且这个数据还有很多东西值得分析挖掘，用户的分类，用户的生命周期分析，各变量之间的交叉分析等等，今后还需对这个项目进行多方面的改进。