# 亲，你到底买不买啊？

## ——请在论文中描述过程并分析结果

### 序言

从前，东哥的购物网站，遇到一个问题（其实现在也一样）：大量人只看不购买（接近80%的访问），请帮我们甄别一下哪些是潜在的购买人。

本次《Python与大数据分析》大作业的题目为：基于已有的服务器所提供的用户访问数据（例如：用户之前浏览了哪些页面、浏览了多长时间、页面的一些属性、是否是节假日等等），预测某些访问会不会产生购买行为，详细情况下一节数据介绍。

注意：

**1 需要提交：1-结果CSV(测评)（提交时间12月12日13:00至18:00）、2-首页、3-正文、4-代码；**

**2 论文部分包括《首页》（不装订）和《正文》（左侧装订）都需要打印提交；**

**3 不要尝试考验老师对Copy的厌恶程度和检测能力。如果两个同学，被证明有copy嫌疑（包括代码、论文、结果文件），两个人的最终成绩都会是59.99分。PS. 已有往届生被查出Copy嫌疑。**

**4 不要把别人跑的结果csv来提交，或者拿别人的test.csv文件来跑结果，自己的事情自己做，否则会爆炸。**

### **数据介绍**

本次大作业题目的train数据位于（6-数据集）文件夹中，test数据会在指定时间（暂定为12月12日）开放3小时，请在这段时间内提交结果，其他时间建议在训练集上构建模型：

train.csv：训练集，8632行×18列，其中第一行是属性名称，其他每行为一位用户访问留下来的相关信息，共8631行。最后一列为该次访问是否产生了购买行为（1-购买，0-未购买）；

test.csv：测试集（点击下图下载），3700行×17列。其中第一行是属性名称，其他每行为一位用户访问留下来的相关信息，共3699行。你需要预测每条访问是否会产生购买行为（1-购买，0-未购买，或者购买的概率，越接近1越会购买）。



数据属性包括：10个数值属性和8个类别属性，每列变量表示的含义如下：

“用户相关页面”、“用户页面访问持续时间”、“信息相关页面”、“信息相关页面访问持续时间”、“产品相关页面”和“产品相关页面访问持续时间”表示访问者在该会话中访问的不同类型页面的数量以及在这些页面类别中花费的总时间。这些特性的值来自于用户访问的页面URL Session信息， Session记录了用户采取的操作（例如从一个页面移动到另一个页面）的实时更新信息。

“跳出率”、“退出率”和“页面值”特性代表了由Google Analytics为电子商务网站中的每个页面计算出来的指标。跳出率是指仅阅读了一个页面就离开的用户占一组页面或一个页面拜访次数的百分比。 跳出次数是指拜访者不拜访您网站的其他任何一页便从进入页退出的次数。 所以跳出率的算法就是：阅读了一个页面就离开网站的次数/进入网站的次数= 跳出率。退出率指：某个时间段内，离开网页的次数占该网页总浏览次数的比例。 比如，今天开心推首页综合浏览量是1000次，从这个页面离开本站的次数是40次，则首页的退出率是4%。“网页价值”是用户在进入目标网页或完成电子商务交易（或两者）之前访问过的网页的平均价值。

“节假日”属性表示网站访问时间接近某个特定的日子（如母亲节、情人节），在这一天附近会更有可能最终完成交易。

数据集还包括“操作系统类型”、“浏览器类型”、“区域”、“流量类型”、“访客类型”（老访客或新访客）、“周末”。

训练集train.csv最后一列“购买”属性为类别标签， 1为购买，0为未购买。

### **所提交的结果**

你需要提交的结果文件为一个.csv格式文件，总计3699行×1列。每一行为0或者1（1代表购买），或者购买的概率值（越接近1越会购买）。

注意：

**1 再次警告，请用自己的test.csv文件生成结果，否则会爆炸；**

**2 结果须为 3699行，请不要添加列标题或增减行数；**

**3 实在不行，请运行下面代码提交一个随机结果。**

参考保存csv代码如下：

#注意 predictY是pandas DataFrame 类型

import pandas as pd

import numpy as np

predictY = pd.DataFrame(np.random.uniform(0,1,3699).reshape(3699,1)) # 产生随机数

predictY.to\_csv('Results\_1.csv', encoding = 'utf-8', index=False , header=False)

### **评价标准——AUC**

详细介绍请见B开头网站：

https://baike.baidu.com/item/AUC/19282953

使用示例

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

import sklearn

Score\_AUC = sklearn.metrics.auc(y\_true, y\_predict)

例如：

>>> import numpy as np

>>> from sklearn import metrics

>>> y = np.array([1, 1, 2, 2])

>>> pred = np.array([0.1, 0.4, 0.35, 0.8])

>>> fpr, tpr, thresholds = metrics.roc\_curve(y, pred, pos\_label=2)

>>> metrics.auc(fpr, tpr)

### **参考代码**

见压缩包 5-参考案例-强力推荐 文件夹