# 为何她/他收入高？ 1. 题目

如题：地球上，总有些人很无聊，想看看别人的收入如何，今天我们就来看个够。



请根据非常可信的统计局发布的信息（包括，性别、年龄、教育状况等），对某地居民是否属于高收入进行预测（可以输出0（非高收入）1（高收入）或者概率值，概率值越大越代表其是高收入者）。

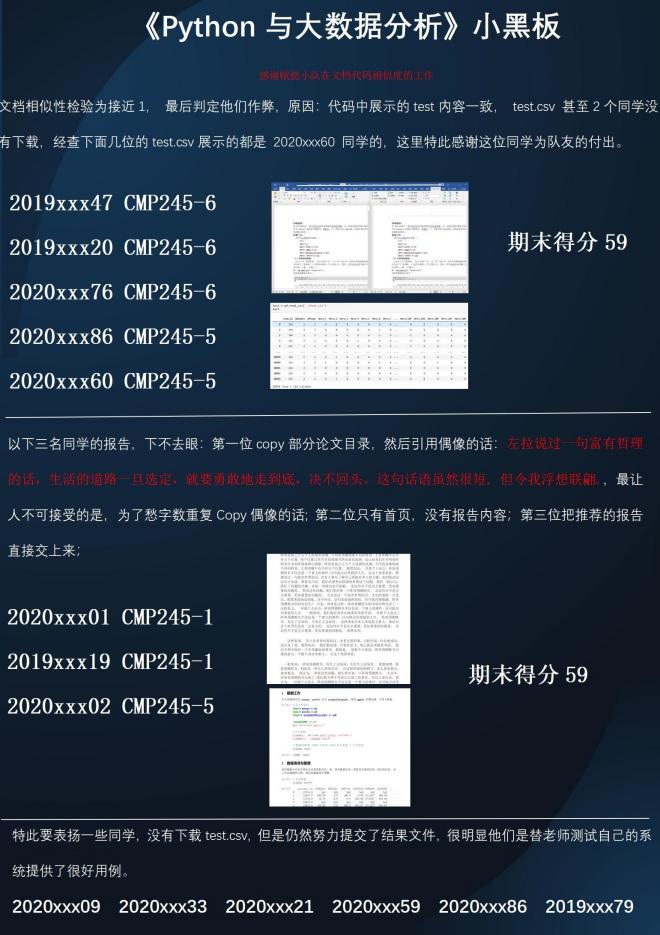
请在论文中详细描述①你是如何建立模型的；②她/他是高收入，有哪些特点，或者收入欠佳，特点有哪些，如果能分析一下为何最好不过。

各位需要提交：

1. 论文首页和论文（两个Word文档 即：1-论文首页-务必填写.doc 和 2-论文格式-务必参考.doc）
2. 代码（Python代码，可以是ipynb 也可以是py ）
3. 结果文件（可提交3组）：Results\_1.csv Results\_2.csv Results\_3.csv（一定要这样命名）

注意 1 不会做，可以看压缩包中：参考案例文件夹，里面有非常漂亮的代码供你欣赏，**实在抢救不了的**，可以运行Py4Submit.py，随机生成 Results\_1.csv 供你提交，切勿copy。

2不要尝试考验老师分析同学们copy嫌疑的能力，如果两个同学，被证明论文或者代码有copy嫌疑，两个人的分数都会是59.99分。如下图，展示于微信群中：



# 2. 数据介绍

# 下面详细介绍数据：

Train.csv 38842行，14列 最后一列Y 中 1：为高收入 0：为非高收入

['年龄','工作情况','教育','教育时间', '婚姻状况','职业类型', '社会关系', '种族', '性别', '投资收入', '投资损失', '工作天数', '省份', 'Y' ]

Test.csv 10000行，13列 无列Y，其他列信息与Train列信息一直

['年龄','工作情况','教育','教育时间', '婚姻状况','职业类型', '社会关系', '种族', '性别', '投资收入', '投资损失', '工作天数', '省份']

# 3. 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10000行×1列。每一行为0或者1（1代表是高收入），或者[0-1]的一个概率值（即该居民是高收入者的概率）。

注意：

**1 结果文件中的sample顺序与test.csv文件中sample顺序严格一致；**

**2 结果须为** 10000**行，请不要添加列标题或增减行数。**

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(0,1,10000).reshape(10000,1)) # 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?fr=aladdin>

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)

# 5. 参考代码

见压缩包中 参考资料 文件夹 和 Py4Submit.py