系统工程 张云佳 1800900

声明：自己编程，没有掉包！

一、选题：第二题

二、编程语言：Python。共三个小程序，其中 train_data_execute.py 为训练数据处理程序，test_data_execute.py 为测试数据处理程序，logstic_r.py 为逻辑回归训练与预测程序。

三、分类器：逻辑回归（正则化）

四、识别率：0.838

五、问题陈述：ML_data2数据将人群的年收入分为高收入（>5万美金）和低收入（<5万美金）两类，包括32561个训练样本和16281个测试样本，每个样本有14个属性（6个数字属性，8个类别属性）可以用来分类，最后一个属性为分类标签。

六、问题分析：

1. 查看数据的情况：

① data_train.columns：

```
Index(['age', 'workClass', 'fnlwgt', 'education', 'education_num',
       'marital_status', 'occupation', 'relationship', 'race', 'sex',
       'capital_gain', 'capital_loss', 'hours_per_week', 'native_country',
       'salary'],
      dtype='object')
```

数据共 15 个字段，代表 15 个属性，最后一个属性'salary'为标签属性。

② data_train.info()：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
age              32561 non-null int64
workClass        30725 non-null object
fnlwgt           32561 non-null int64
education        32561 non-null object
education_num    32561 non-null int64
marital_status   32561 non-null object
occupation       30718 non-null object
relationship     32561 non-null object
race             32561 non-null object
sex              32561 non-null object
```

```
capital_gain      32561 non-null int64
capital_loss      32561 non-null int64
hours_per_week    32561 non-null int64
native_country    31978 non-null object
salary            32561 non-null int64
dtypes: int64(7), object(8)
memory usage: 3.7+ MB
```

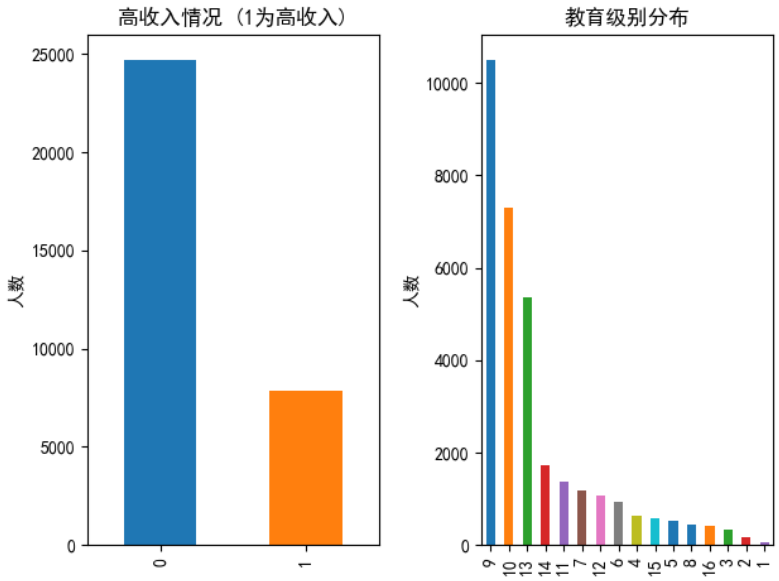
训练数据中总共有 32561 个样本，但是有的数据不全，比如 workClass 只有 30725 个记录；occupation 只有 30718 个记录；native_country 只有 31978 个记录。

③ data_train.describe()：

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week	salary
count	32561.00	32561.00	32561.00	32561.00	32561.00	32561.00	32561.00
mean	38.58	189778.37	10.08	1077.65	87.30	40.44	0.24
std	13.64	105549.98	2.57	7385.29	402.96	12.35	0.43
min	17.00	12285.00	1.00	0.00	0.00	1.00	0.00
25%	28.00	117827.00	9.00	0.00	0.00	40.00	0.00
50%	37.00	178356.00	10.00	0.00	0.00	40.00	0.00
75%	48.00	237051.00	12.00	0.00	0.00	45.00	0.00
max	90.00	1484705.00	16.00	99999.00	4356.00	99.00	1.00

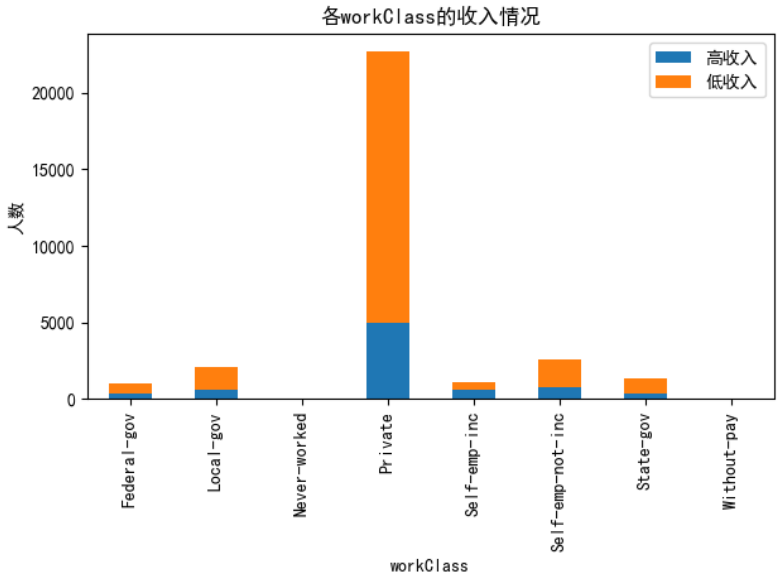
从 mean 字段可以看出，大概有 0.24 的人属于高收入群体，平均年龄在 38.58 岁等等。

④ 看看收入情况和教育分布情况：



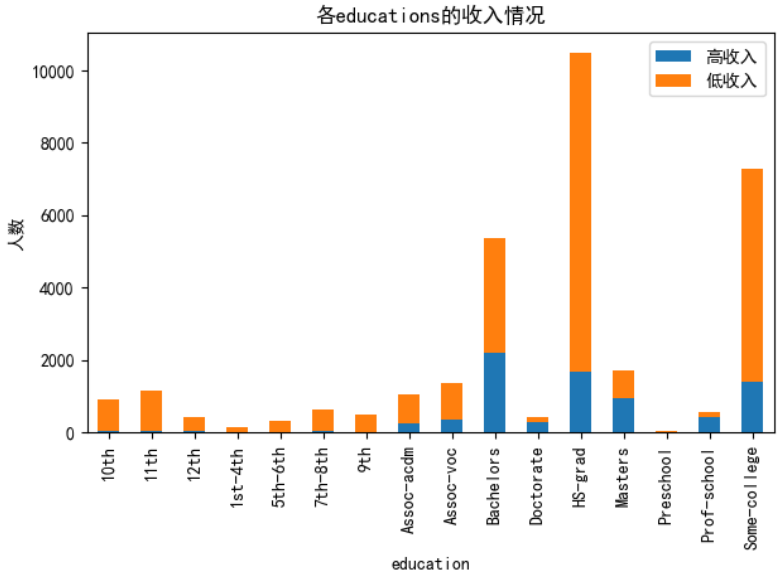
有 7800 多人为高收入，教育级别集中在 9,10,13 级。

⑤ 各 workclass 的收入情况：



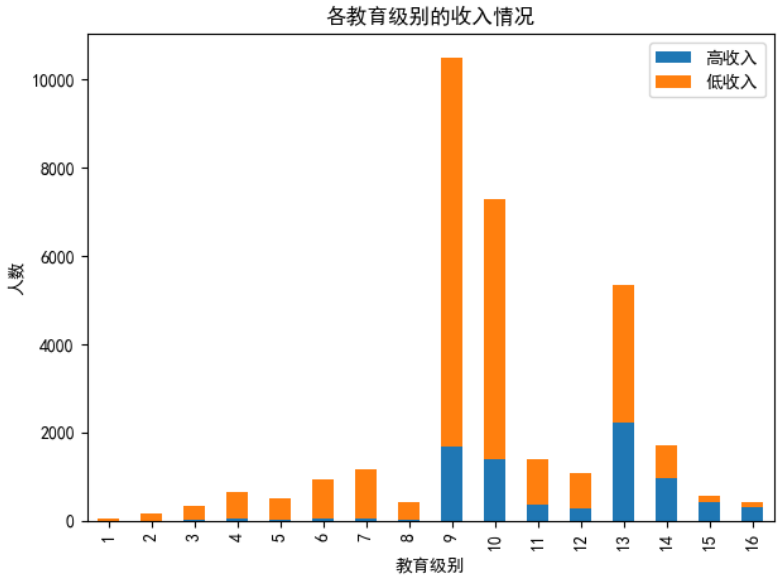
Workclass 为 private 时低收入率偏大。

⑥ 各 education 的收入情况：



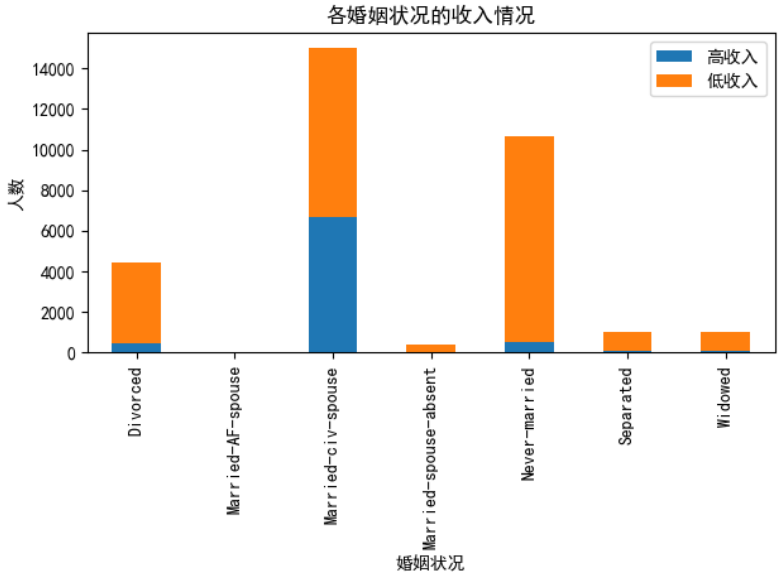
Education 为 Bachelors 时高收入人群比例偏高

⑦ 各教育级别的收入情况：



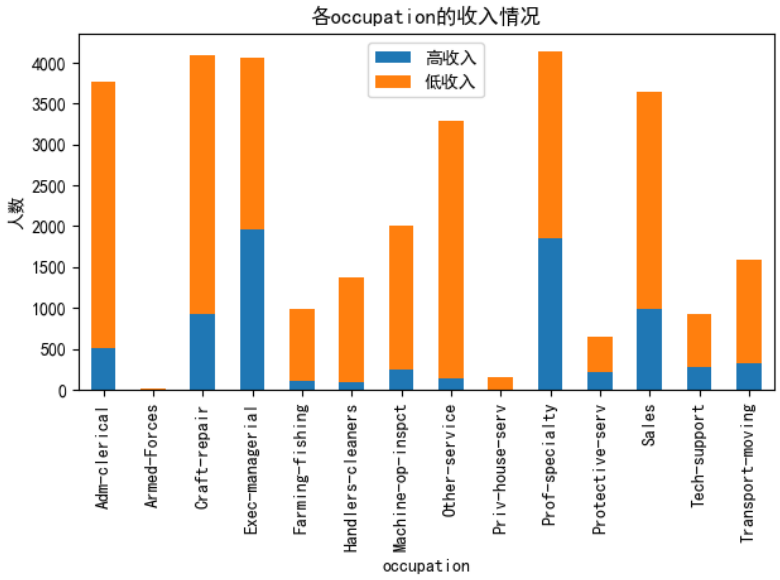
随着教育级别的增加，高收入的比例逐渐增加，说明是否为高收入和教育级别关系很大。

⑧ 各婚姻状况的收入情况：



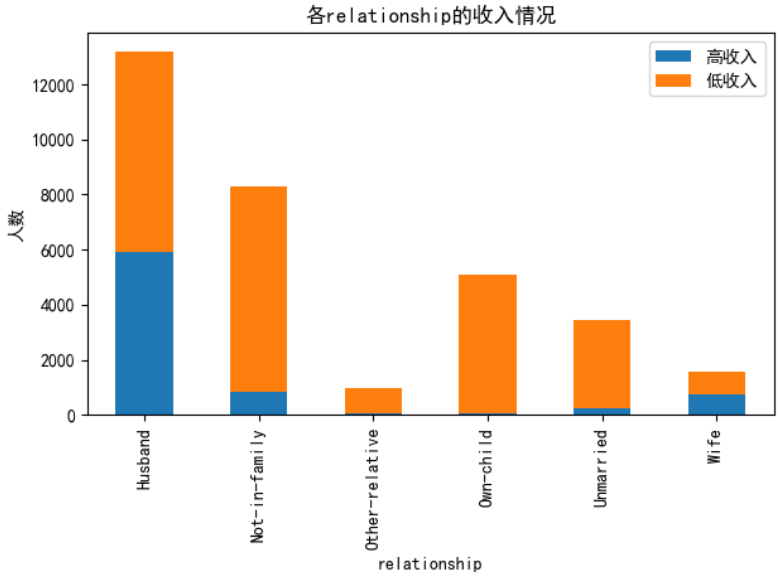
当婚姻状况为 `Married-civ-spouse` 时，为高收入的比例最高，说明是否为高收入和婚姻状况关系很大。

⑨ 各 occupation 状况的收入情况：



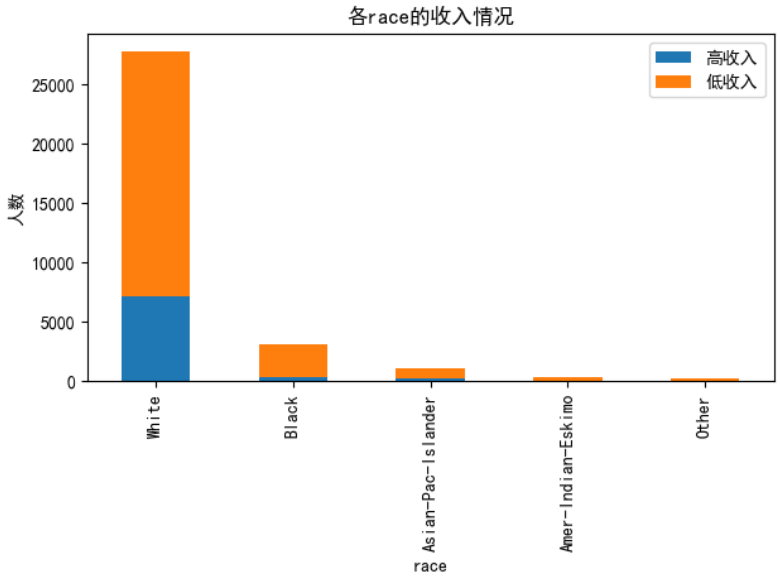
各 occupation 之间差距较大，说明 occupation 对收入的影响显著

⑩ 各 relationship 状况的收入情况：



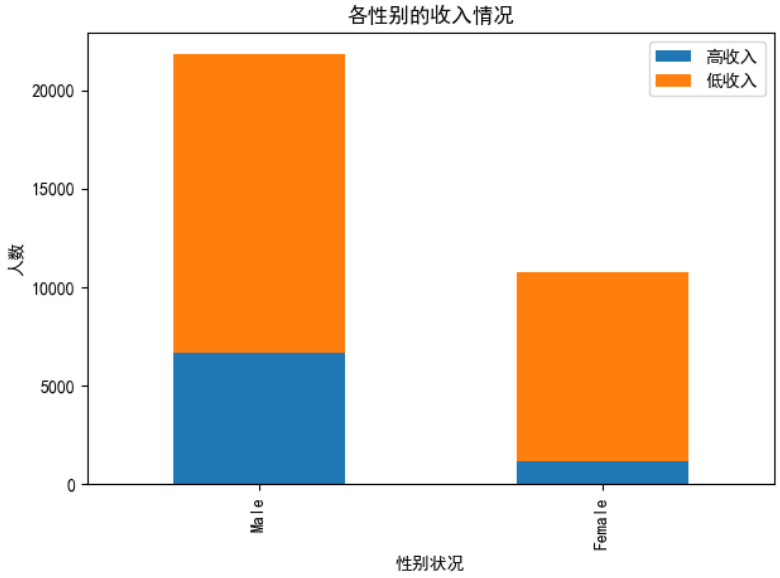
各 relationship 之间差距较大，说明 relationship 对收入的影响显著

⑪ 各 race 状况的收入情况：



Race 似乎对收入的影响不大

⑫ 各性别的收入情况：



性别对收入影响似乎不是太大。

七、训练结果：
将以上对收入影响不大的属性丢弃，能达到 78%的正确率。
如果保留所有的属性则可以达到 84%的正确率
说明保留更多的信息有助于提高预测的准确率。

八、具体处理细节：
处理缺失值的方法：将缺失值填为 0
处理类目型数据的方法：国籍将美国国籍标为 1，其余标为 0；性别男性标为 1，女性标为 0。其余属性采用 one-hot 编码

九、分类器构建、分类器训练过程与结果：见程序，程序里写得很详细。