****

Spark期末项目

实验报告

|  |  |
| --- | --- |
| 题 目： | UCI adult数据集二分类 |
| 学生姓名： | 刘宣辰 |
| 指导教师： | 高建良 |
| 学 院： | 计算机学院 |
| 专业班级： | 大数据1702 |

目录

[项目背景 3](#_Toc44448534)

[项目设计框架 3](#_Toc44448535)

[数据来源及格式 3](#_Toc44448536)

[数据处理过程步骤 6](#_Toc44448537)

[核心模块代码介绍及结果输出 7](#_Toc44448538)

[结果分析 15](#_Toc44448539)

# 项目背景

该数据来自于UCI公开数据集，研究的问题是根据户口普查资料，预测收入是否能超过 50000 美元/年。

# 项目设计框架

1.数据预处理

2.模型建立

3.模型训练

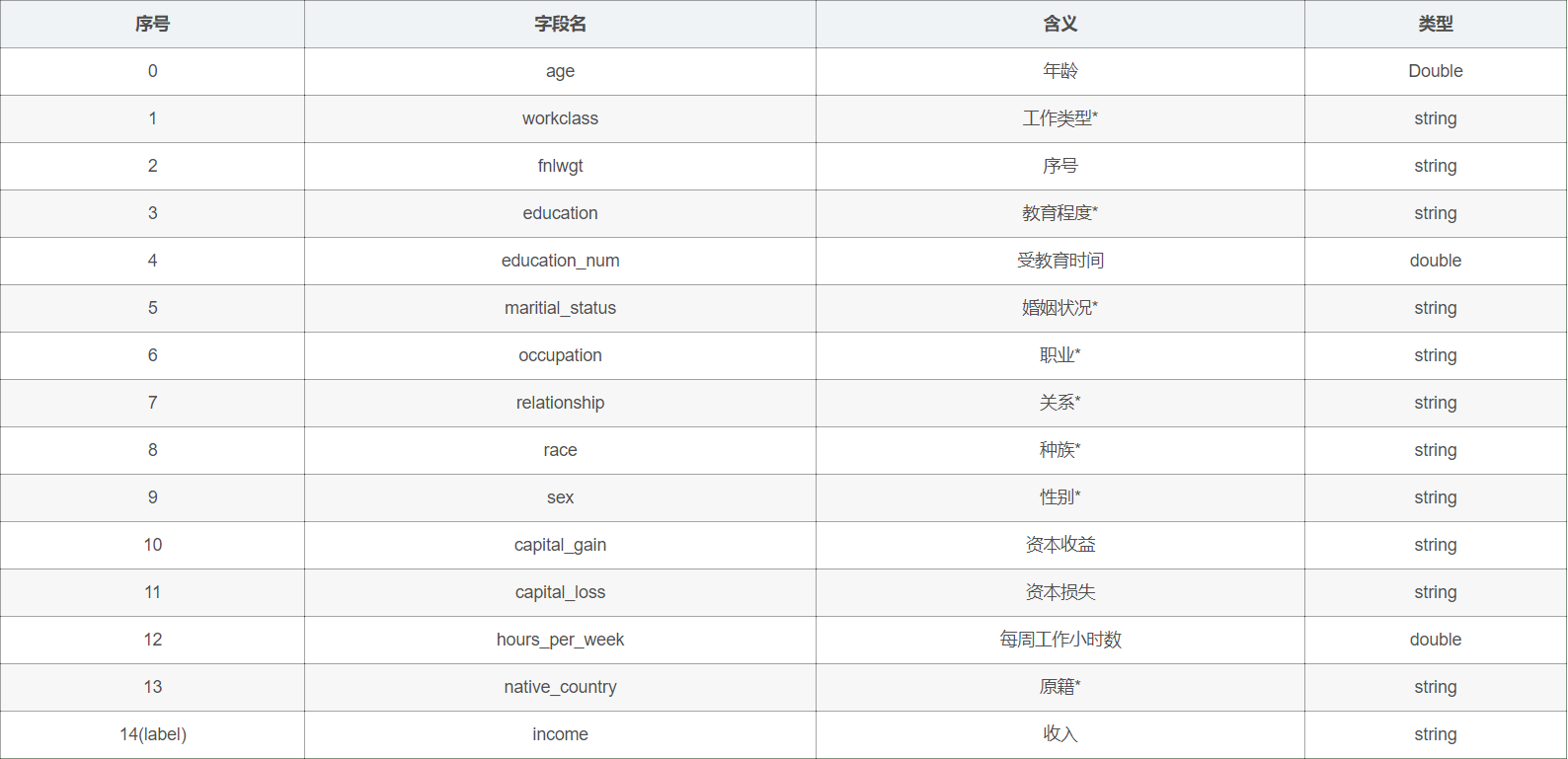
4.模型评估

# 数据来源及格式

该数据从美国1994年人口普查数据库抽取而来，具体使用以下条件提取了一组合理的干净记录：（（AAGE> 16）&&（AGI> 100）&&（AFNLWGT> 1）&&（HRSWK> 0））。可以用来预测居民收入是否超过50K/year。

该数据集标签为年收入是否超过50k，属性变量包含年龄，工种，学历，职业，人种等重要信息，值得一提的是，14个属性变量中有7个类别型变量.。数据集各属性是:其中序号0~13是属性，14是类别。

数据集各列具体描述如下：



年龄（age）：连续值

工作种类（Workclass）：个人（Private）, 无限责任公司（Self-emp-not-inc）, 有限责任公司（Self-emp-inc）, 联邦政府（Federal-gov）, 地方政府（ Local-gov）, 州政府（State-gov）, 无薪人员（Without-pay）, 无工作经验人员（Never-worked）离散值

Fnlwgt：连续值

教育情况（Education）：Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool ）离散值

受教育年限（Education-num）：连续值

婚姻状况（Marital-status）：已婚（Married-civ-spouse），离婚（Divorced），未婚（Never-married），离异（Separated），丧偶（Widowed），已婚配偶缺席（Married-spouse-absent）、 再婚（Married-AF-spouse），离散值

职业情况（Occupation）：技术支持（Tech-support），维修工艺（Craft-repair），服务行业（Other-service）、 销售（Sales）、 执行管理（Exec-managerial）、 专业教授（Prof-specialty），清洁工（Handlers-cleaners），机床操控人员（Machine-op-inspct）、 行政文员（Adm-clerical）、 养殖渔业（Farming-fishing）、 运输行业（Transport-moving），私人房屋服务（Priv-house-serv），保卫工作（Protective-serv）， 武装部队（Armed-Forces）职业情况，离散值

亲属情况（Relationship）妻子（Wife），子女（Own-child），丈夫（Husband），外来人员（Not-in-family）、 其他亲戚（Other-relative）、 未婚（Unmarried），离散值

种族肤色（Race）：白人（White），亚洲太平洋岛民（Asian-Pac-Islander），阿米尔-印度-爱斯基摩人（Amer-Indian-Eskimo）、 其他（Other），黑人（Black）离散值

性别（Sex ）：男性（Female）,女性（ Male），离散值

资本盈利（Capital-gain ）：连续值

资本损失（Capital-loss）：连续值

每周工作时间（Hours-per-week ）：连续值

国籍（Native-country ）：美国（United-States）、 柬埔寨（Cambodia）、 英国（England），波多黎各（Puerto-Rico），加拿大（Canada），德国（Germany），美国周边地区（关岛-美属维尔京群岛等）（Outlying-US(Guam-USVI-etc)），印度（India）、 日本（Japan）、 希腊（Greece）、 美国南部（South）、 中国（China）、 古巴（Cuba）、 伊朗（Iran）、 洪都拉斯（Honduras），菲律宾（Philippines）、 意大利（Italy）、 波兰（Poland）、 牙买加（Jamaica）、 越南（Vietnam）、 墨西哥（Mexico）、 葡萄牙（Portugal）、 爱尔兰（Ireland）、 法国（France）、多米尼加共和国（Dominican-Republic）、 老挝（Laos）、 厄瓜多尔（Ecuador）、 台湾（Taiwan）、 海地（Haiti）、 哥伦比亚（Columbia）、 匈牙利（Hungary）、 危地马拉（Guatemala）、 尼加拉瓜（Nicaragua）、苏格兰（Scotland）、 泰国（Thailand）、 南斯拉夫（Yugoslavia），萨尔瓦多（El-Salvador）、 特立尼达和多巴哥（Trinadad&Tobago）、 秘鲁（Peru），香港（Hong），荷兰（Holland-Netherlands）离散值

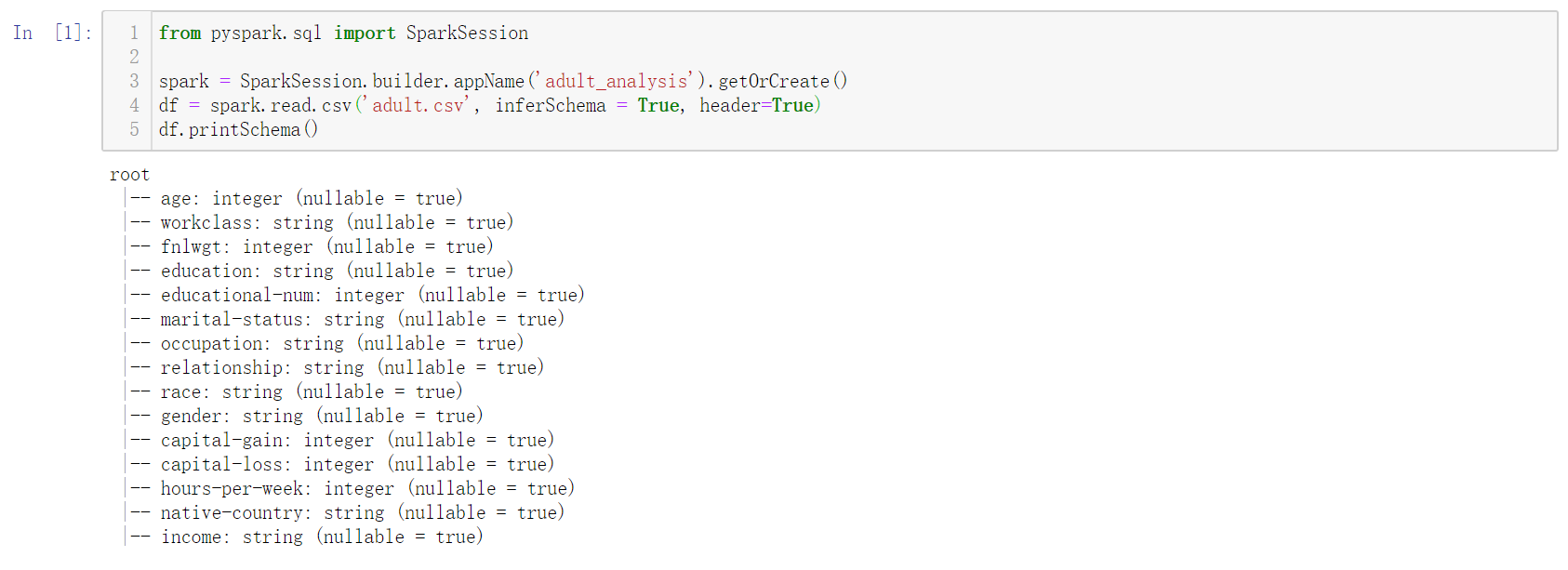
收入 （income）：>50K, <=50K ，离散值

# 数据处理过程步骤

该数据集比较规整，唯一存在的缺陷是存在大约7%的缺失值，但是由于spark的ml包中的分类算法几乎都对缺失值不敏感，这里为了保留所有有效信息，再加上考虑到未来收集到的需要进行预测的数据仍然极有可能存在缺失值，所以不对缺失值处理，而是单独将缺失值作为一类。

# 核心模块代码介绍及结果输出

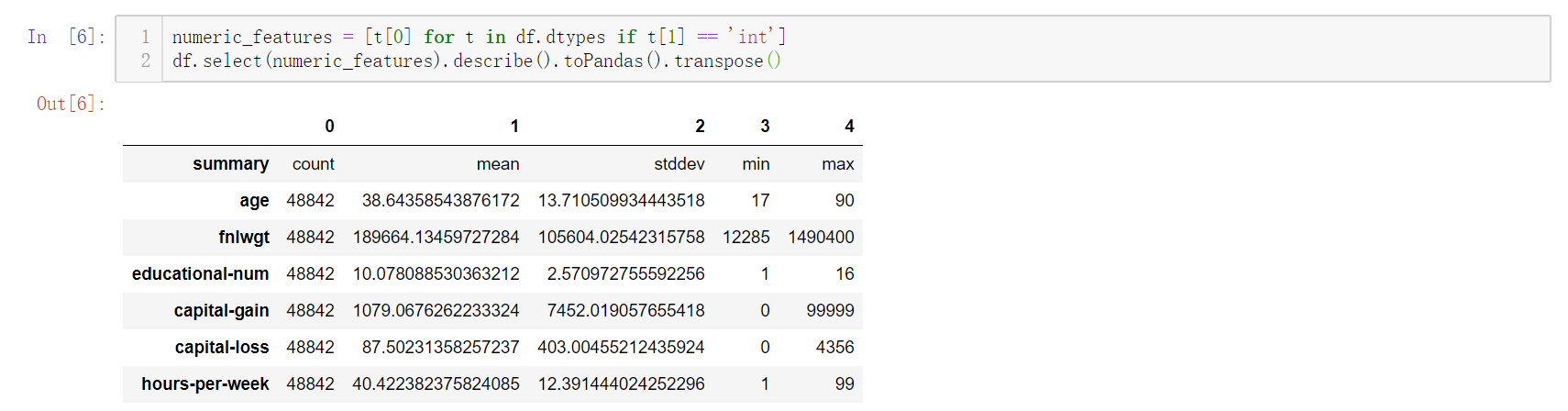
导入数据及初步查看



查看正负类的样本个数，发现存在类别不平衡现象

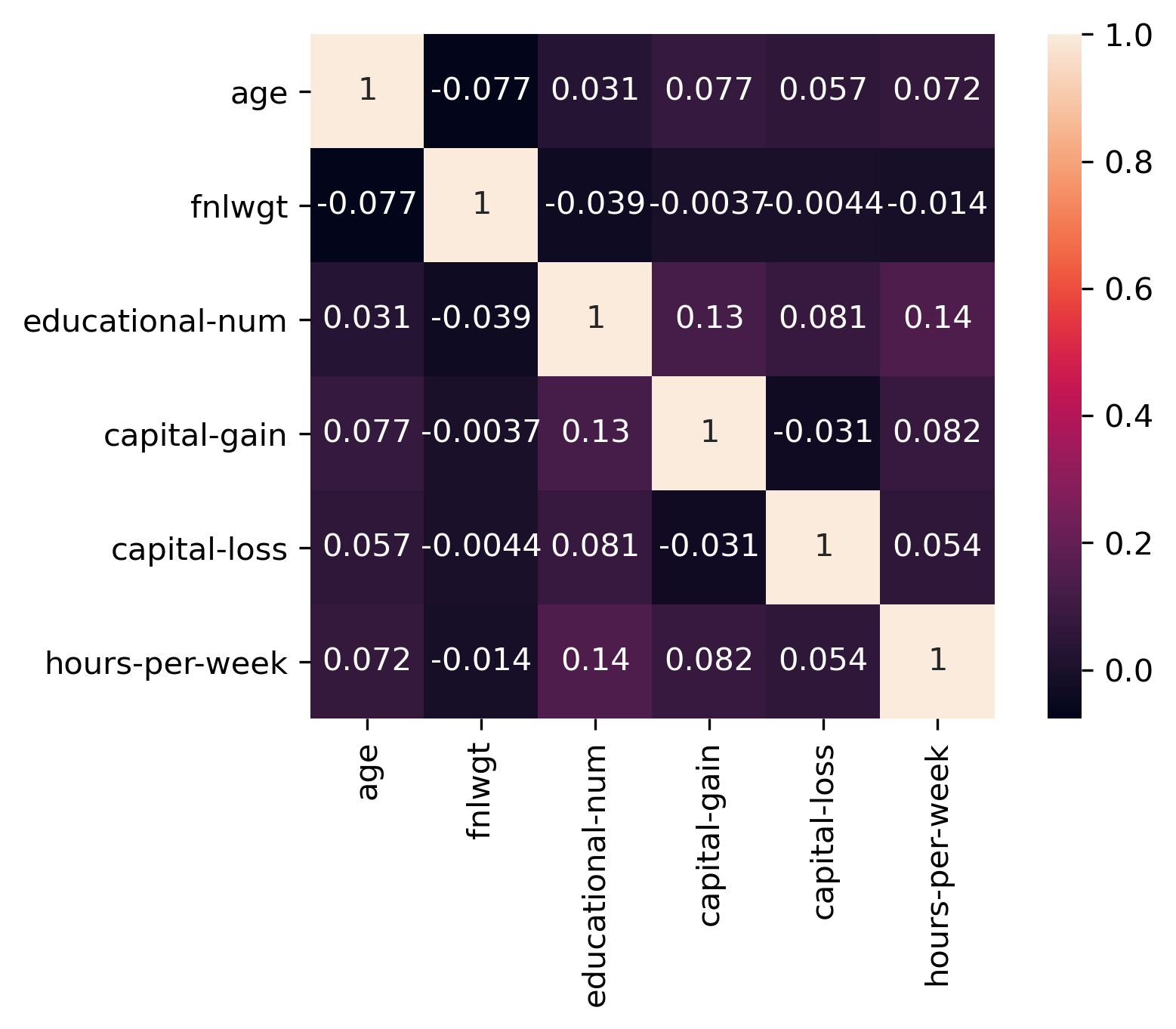


查看数值变量



计算出所有数值变量的协方差矩阵，然后用热力图可视化。





观察热力图发现，所有变量之间的一阶相关性不强，所以所有的数值变量都保留。

开始对数据进行处理，建立spark的pipeline对数据进行一系列转换。

第一种pipeline是没经过归一化的

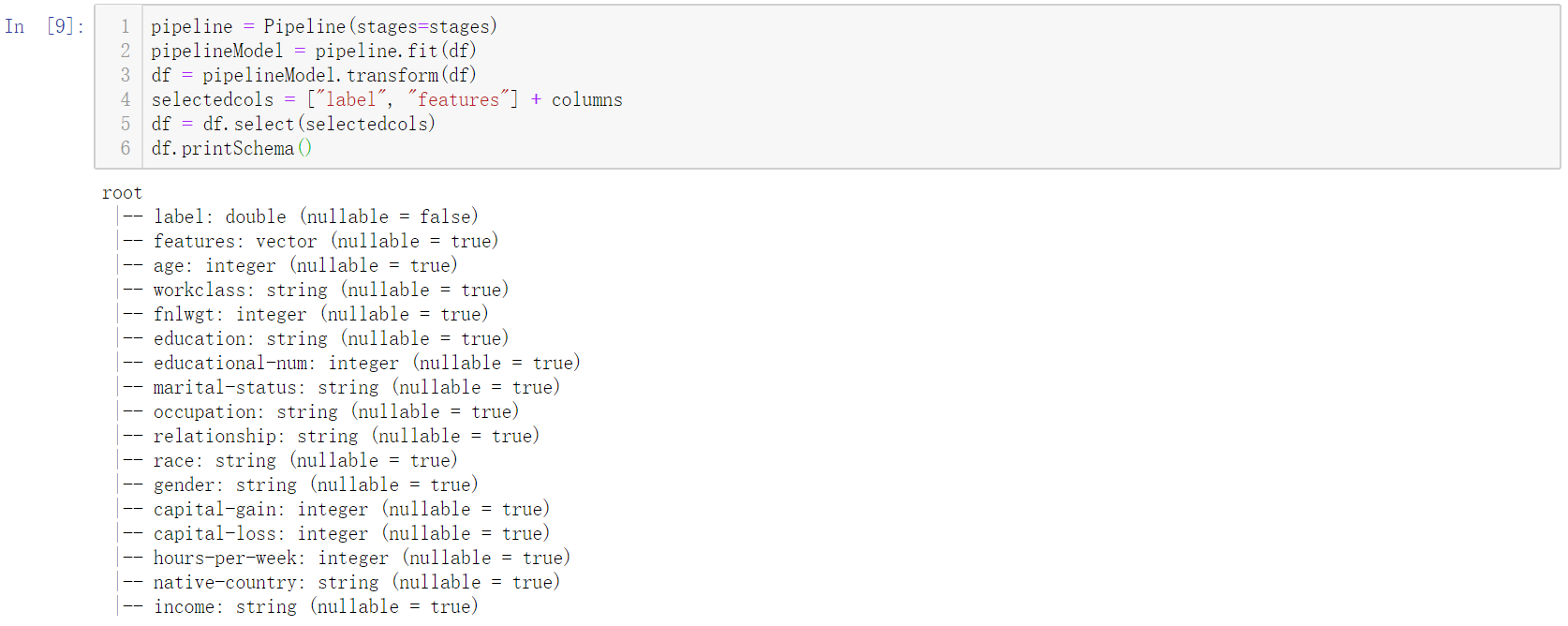


第二种pipeline是经过归一化的



下面只详细分析未归一化的情况

Pipeline已创建好，对数据集进行fit和transform。



数据处理好之后，开始模型训练。

首先是划分训练集和测试集。



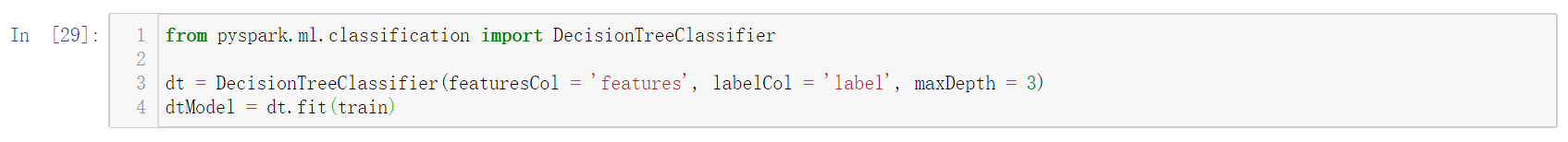
逻辑回归算法



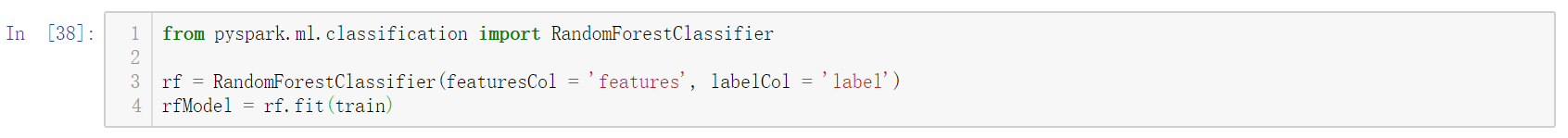
支持向量机算法



决策树算法



随机森林算法



梯度提升树算法

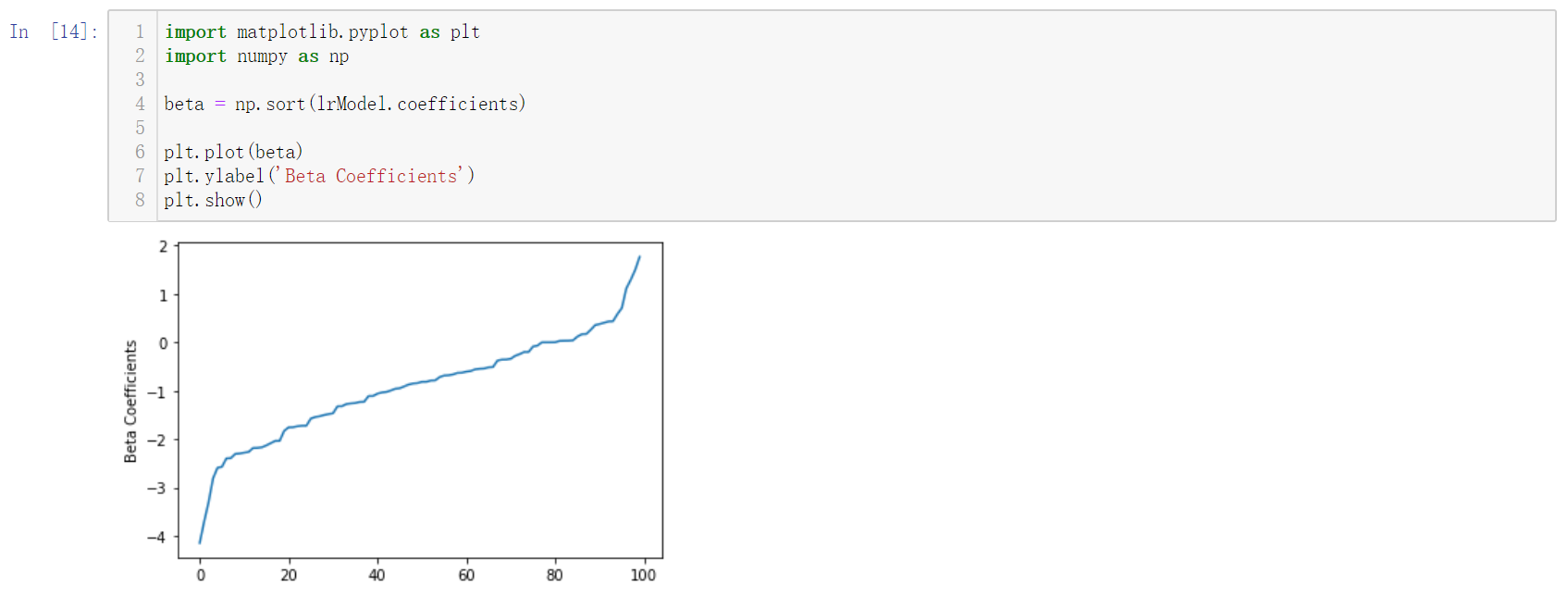


总共使用了五种算法进行训练。

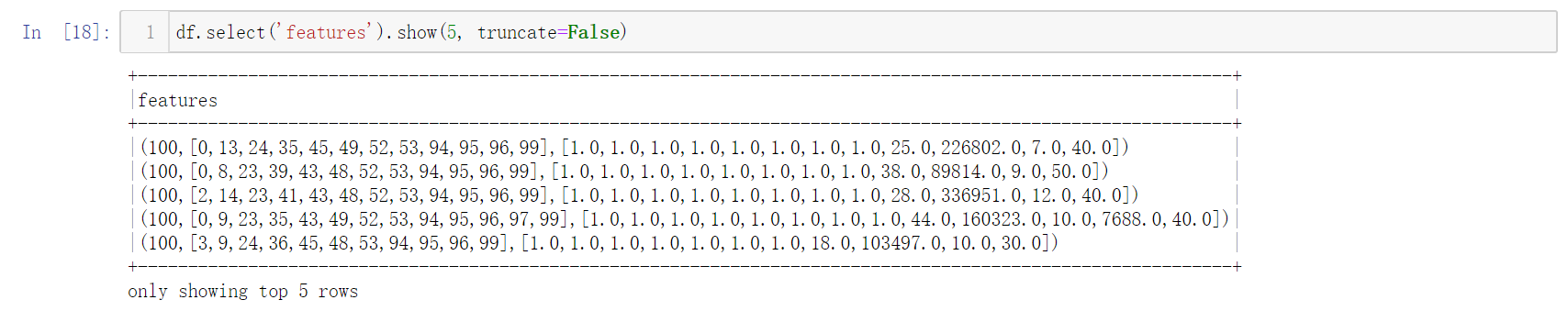
之后仅以逻辑回归算法为例展示后续流程。

首先观察逻辑回归算法的参数。

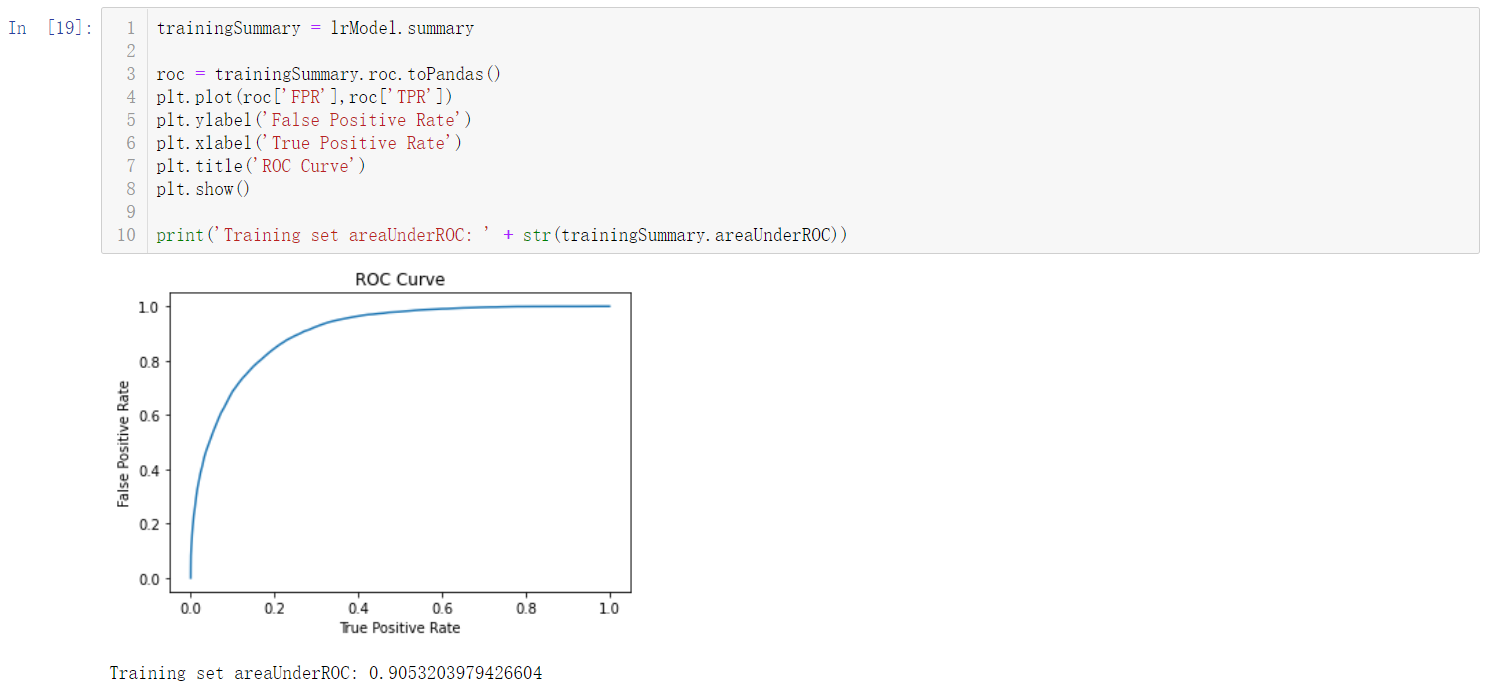
特征向量是100维的向量，所以有100个参数。



观察前5行的features，以Density Vector形式展示



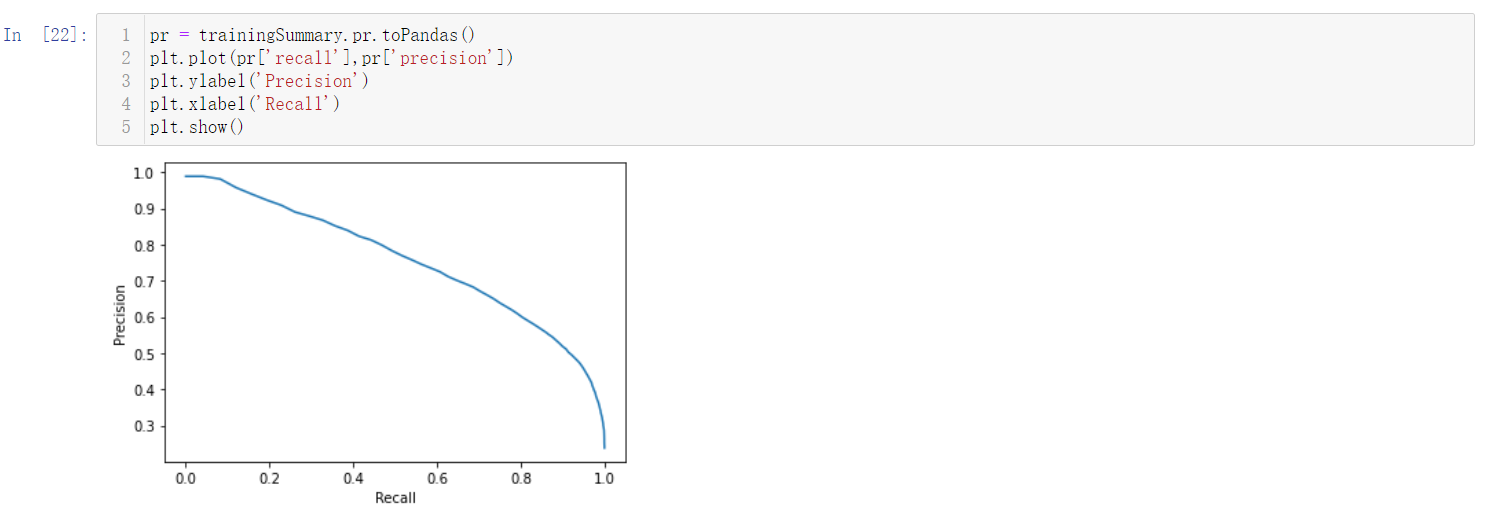
然后绘制训练集ROC曲线，初步分析。



输出训练集准确率和召回率。

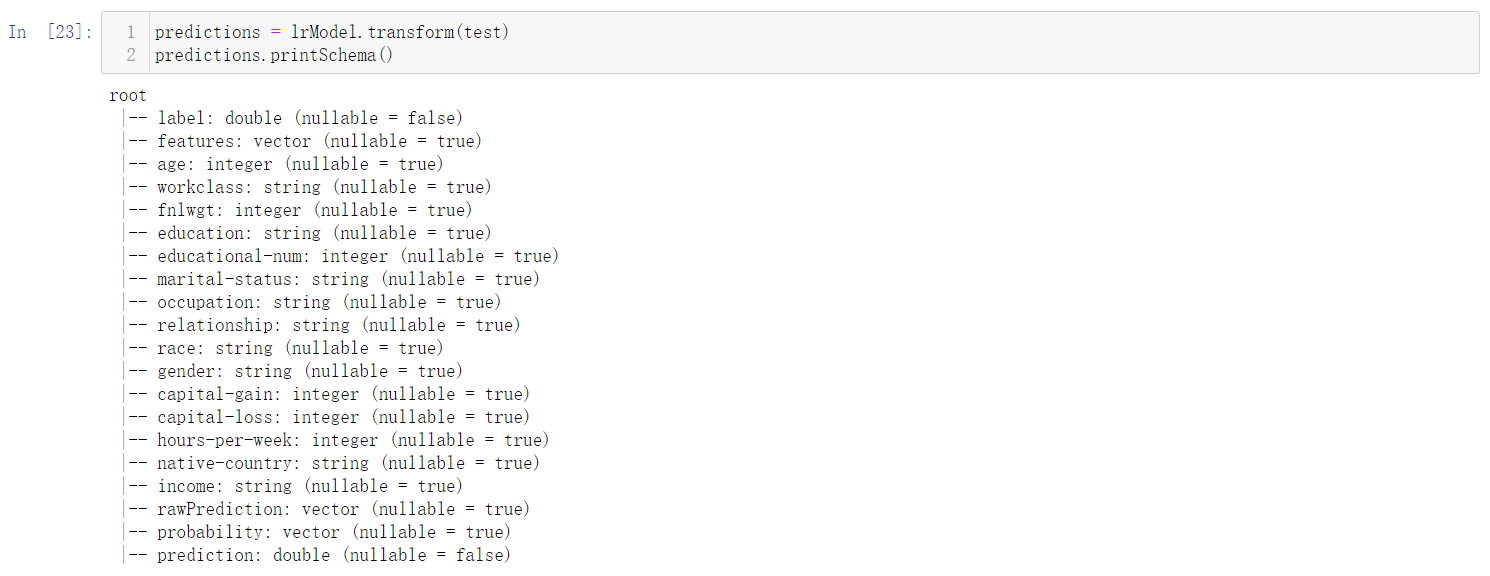


绘制P-R曲线

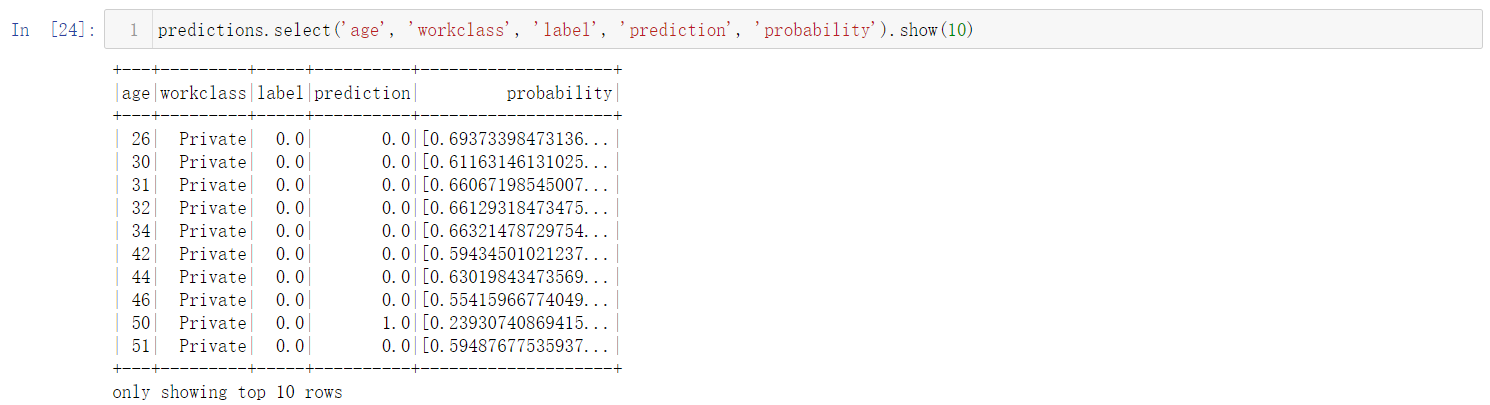


根据上述在训练集上的指标进行评估，选出最佳的模型，对其在测试集上评估。

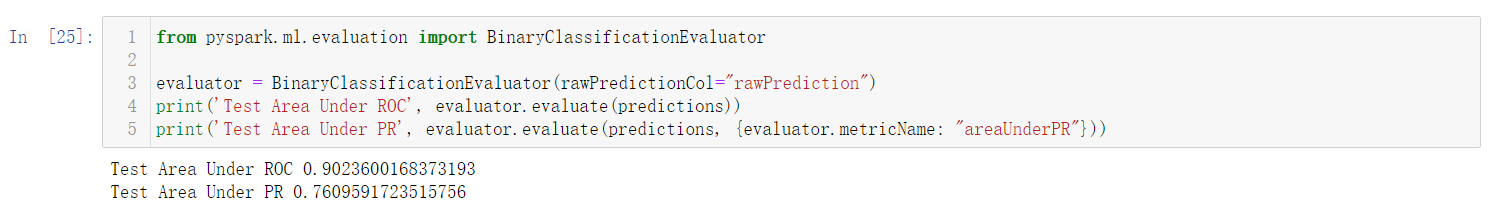
之后开始评估模型在测试集上的表现



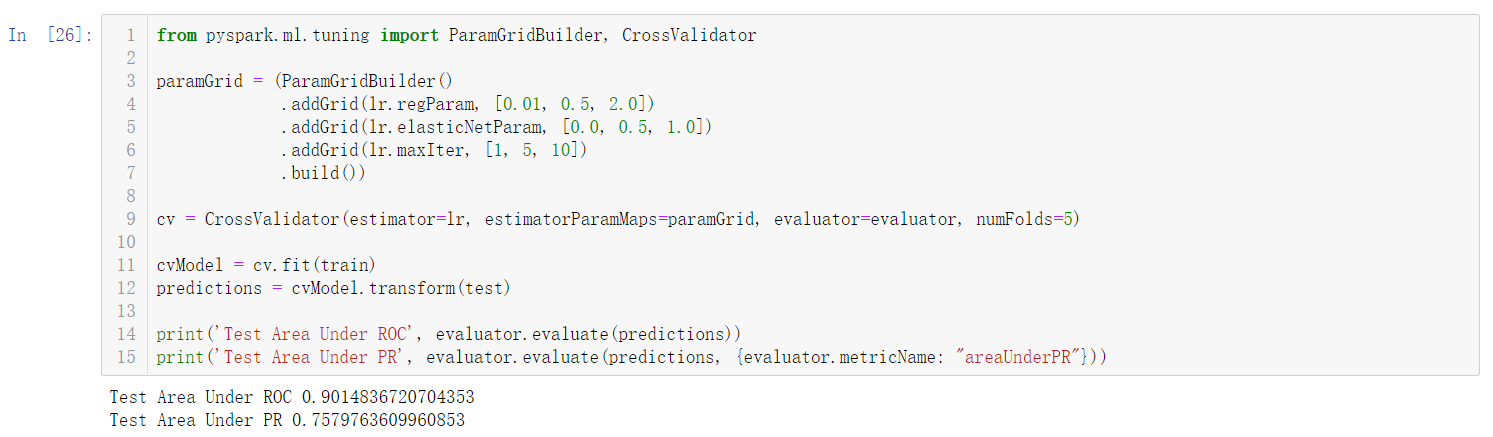
观察前十个测试集样本的真实类别和预测类别及其概率



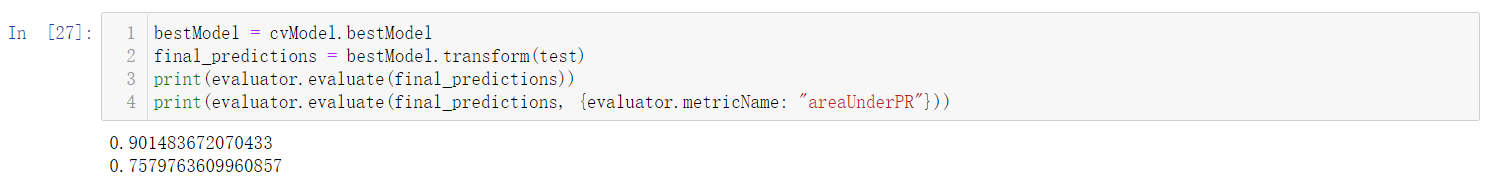
计算评估指标



再用网格调参和交叉验证进行参数调整。输出交叉验证的平均ROC曲线面积和PR曲线面积。



再输出最佳参数的ROC曲线面积和PR曲线面积。

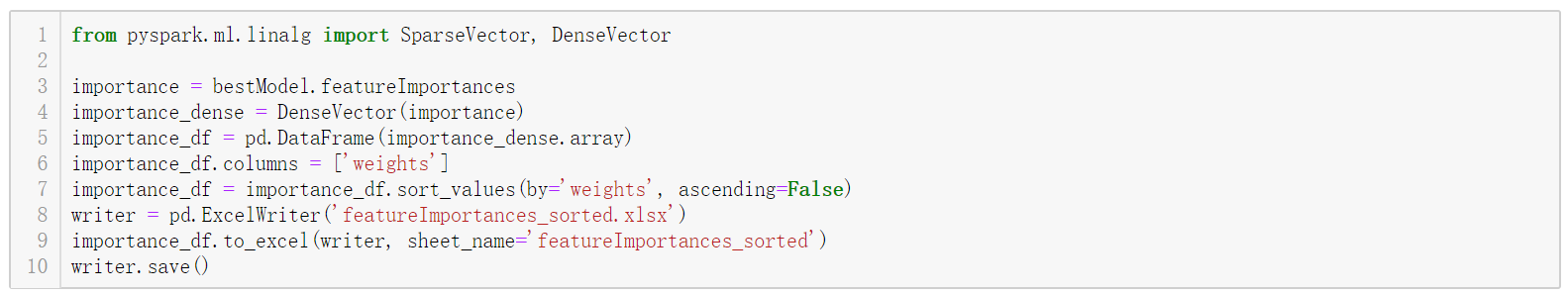


评估完毕。

接着还要根据模型的权重判断那些因素和收入十分相关。

由于之前已经利用pipeline将数据转化成features和label两列，features是一个100列的特征矩阵，所以我们要根据之前pipeline处理的顺序找到features每个列对应的意义。

先保存最好的模型中每个特征的权重，让其按照权重大小降序排列，写入到excel表格中保存。



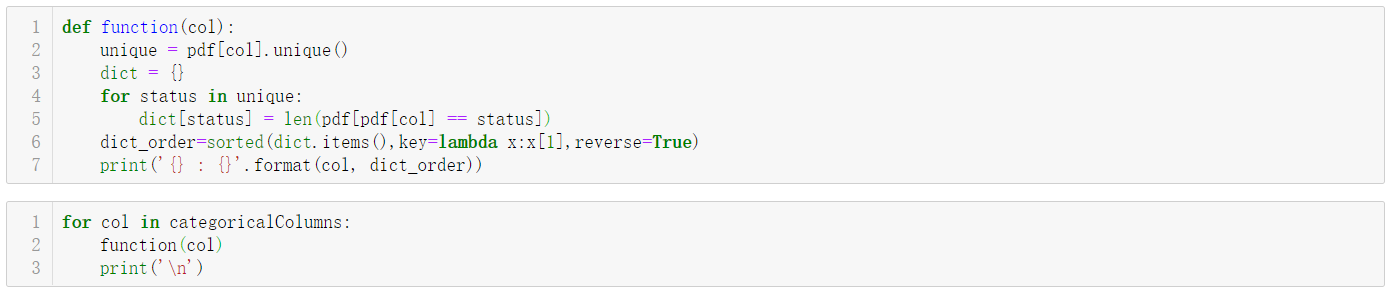
观察权重发现，权重大于0.01的并不多，所以我们着重分析这些较重要的特征。

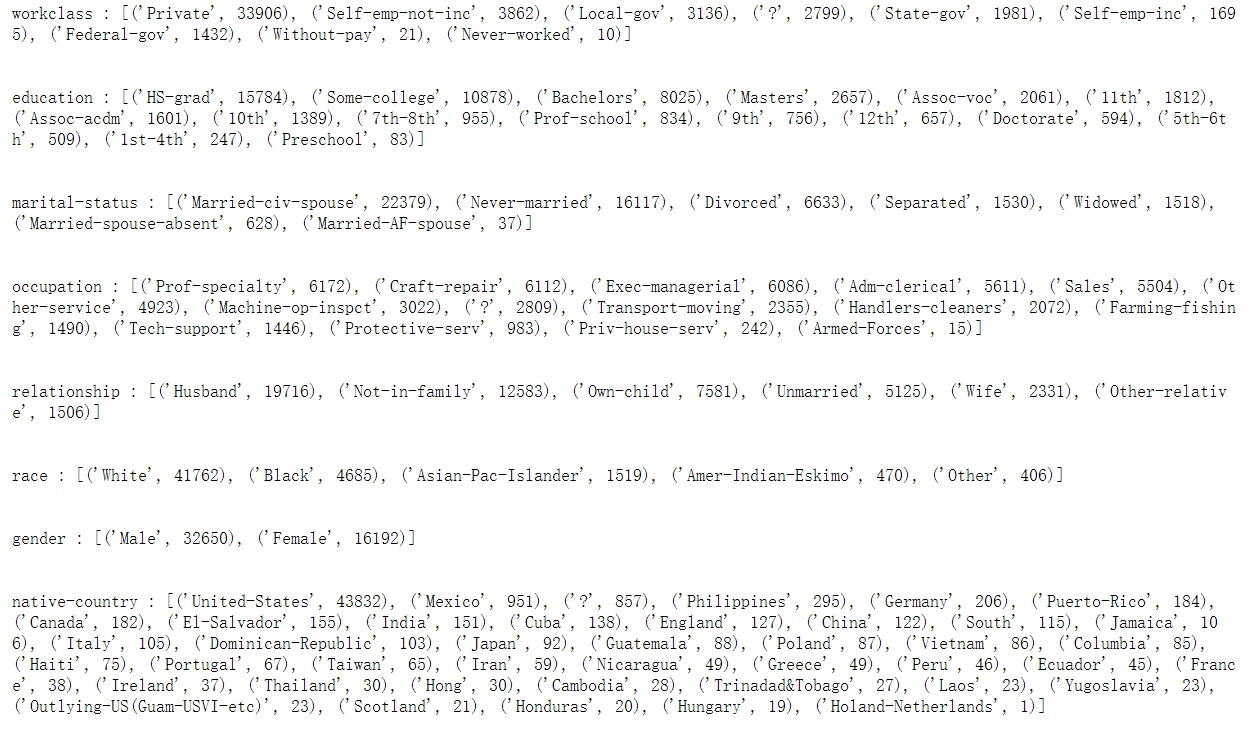


由于之前将类别变量进行了one-hot处理，所以找到features中每列对应的类别。



由于spark的stringIndexer方法是按照每个类别的数目大小排列进行编号的，所以建立一个字典统计每个类别的数目，然后降序排列输出，其对应了one-hot从全零到最后一位为1。





# 结果分析

本次实验对UCI的adults数据集用了五种分类模型进行训练，并进行了综合多种指标的性能评估，最后发现梯度提升树的效果最好，决策树的效果最差。我后来又在pipeline里加上了对数值变量的归一化处理，相比未归一化的情况分类器的性能有所提升。之后可以进一步通过观察模型的具体参数来分析那些是重要特征，哪些是不重要的特征。权重大于0.01即较为重要的特征如下表格。

|  |  |  |
| --- | --- | --- |
|  | **weights** | **meanings** |
| **97** | 0.204319752 | capital-gain |
| **23** | 0.191742522 | marital-status:Married-civ-spouse |
| **94** | 0.097764629 | age |
| **99** | 0.082315988 | hours-per-week |
| **96** | 0.067811 | educational-num |
| **98** | 0.045293405 | capital-loss |
| **31** | 0.041227707 | occupation:Protective-serv |
| **95** | 0.030503287 | fnlwgt |
| **29** | 0.025068265 | occupation:Prof-specialty |
| **1** | 0.018136328 | workclass:Self-emp-not-inc |
| **34** | 0.017428714 | occupation:Other-service |
| **39** | 0.014273829 | occupation:Tech-support |
| **47** | 0.012860402 | relationship:Other-relative |
| **33** | 0.010310473 | occupation:Sales |