**机器人检测实验报告**

**小组成员：**

**王柏程 2022211243**

代码（数据预处理、相关性分析），实验报告撰写

**陈虹瑾 2022211264**

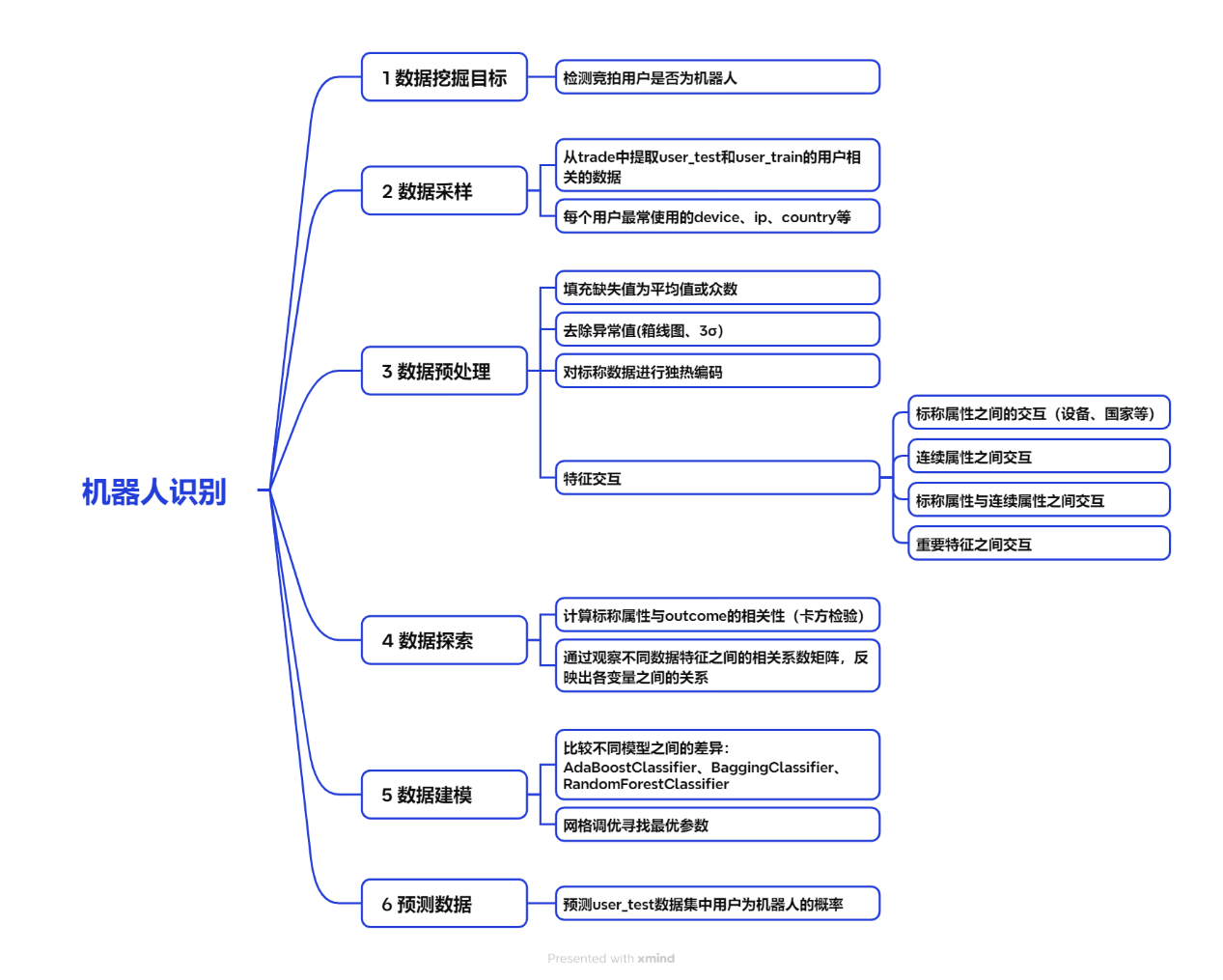
代码（特征交互与选择、模型选择及参数优化），实验报告撰写

1. **实验内容和目标**

**实验目标：**

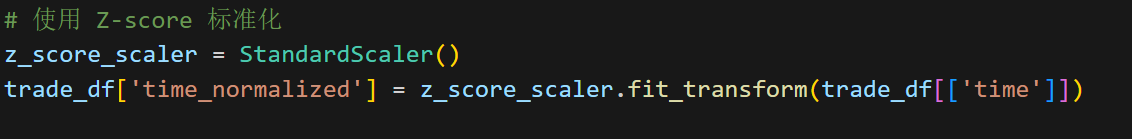
拍卖活动中大量机器人参与者严重影响了人类用户的拍卖体验，导致用户参与热情下降。为了解决这一问题，实验内容包括：数据收集与预处理，通过收集和清洗用户行为数据来提取特征；选择适合的机器学习模型进行训练，通过评估模型性能并进行调整来提高准确性，预测用户为机器人的概率，减少其对拍卖公平性的影响。

**实验内容框架：**



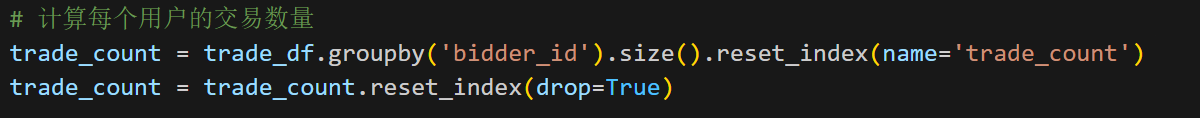
1. **实验步骤和过程**
   1. **分析、提取数据**

**1、使用Z-score标准化time属性：**

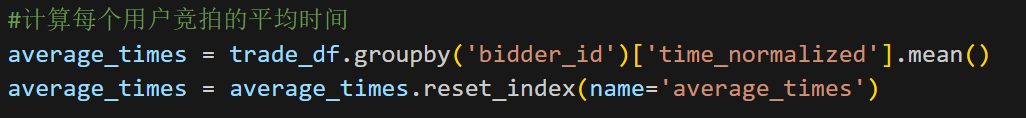


**2、定义一个函数trade\_data\_processing（）从trade.csv文件中提取用户交易次数、平均时间以及所有常用数据：**

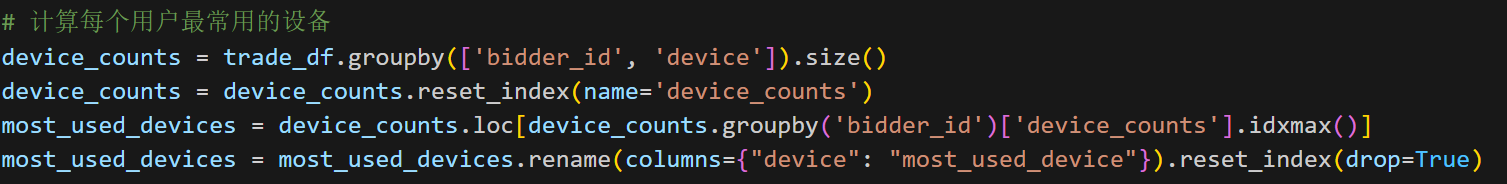
1. 用户交易次数trade\_count



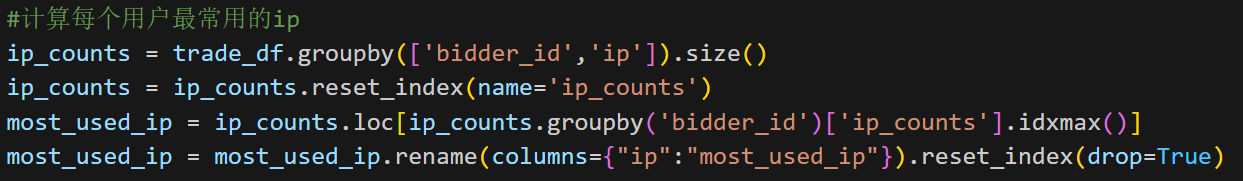
1. 用户平均时间average\_time



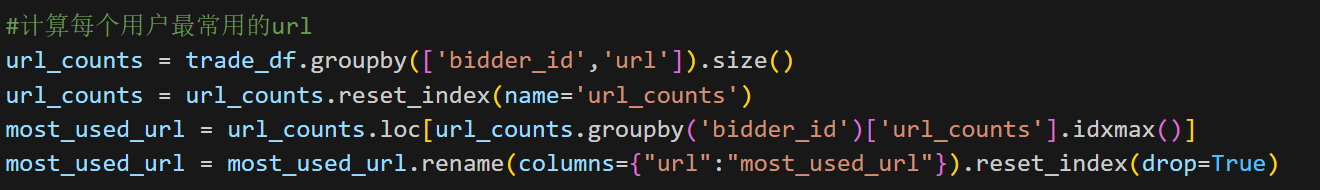
1. 最常用设备most\_used\_devices



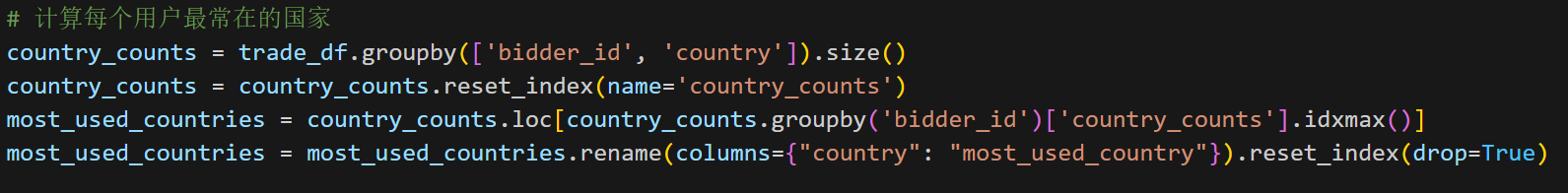
1. 最常用ip most\_used\_ip



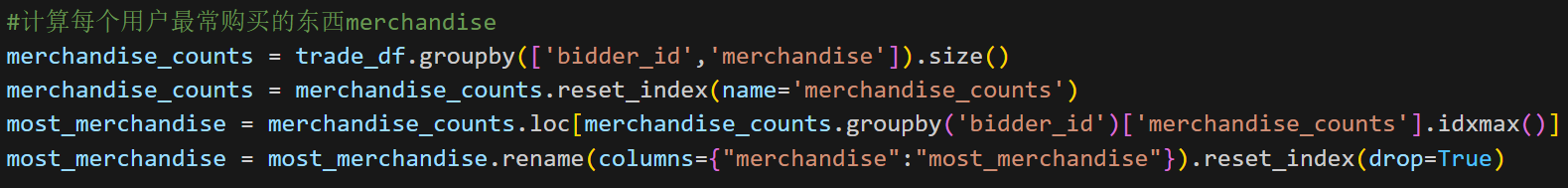
1. 最常用url most\_used\_url



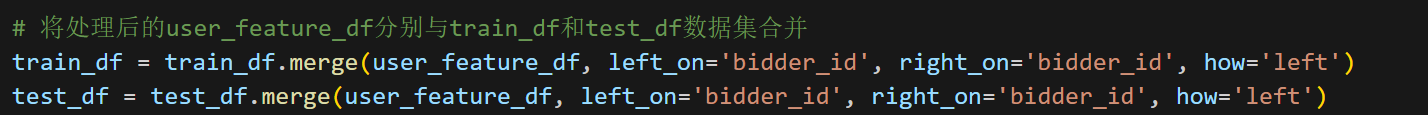
1. 最常用国家 most\_used\_countries



1. 最常购买的东西 most\_merchandise



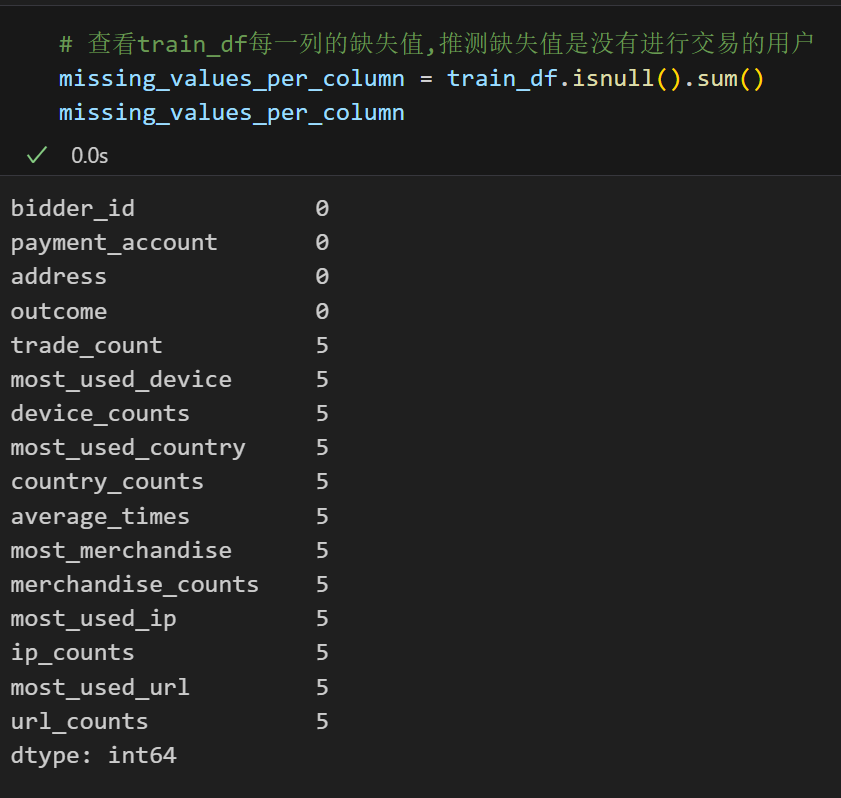
提取数据后将所有结果合并得到user\_feature\_df，然后将user\_feature\_df分别与train\_df和test\_df按照bidder\_id列合并得到初始训练集train\_df和测试集test\_df。



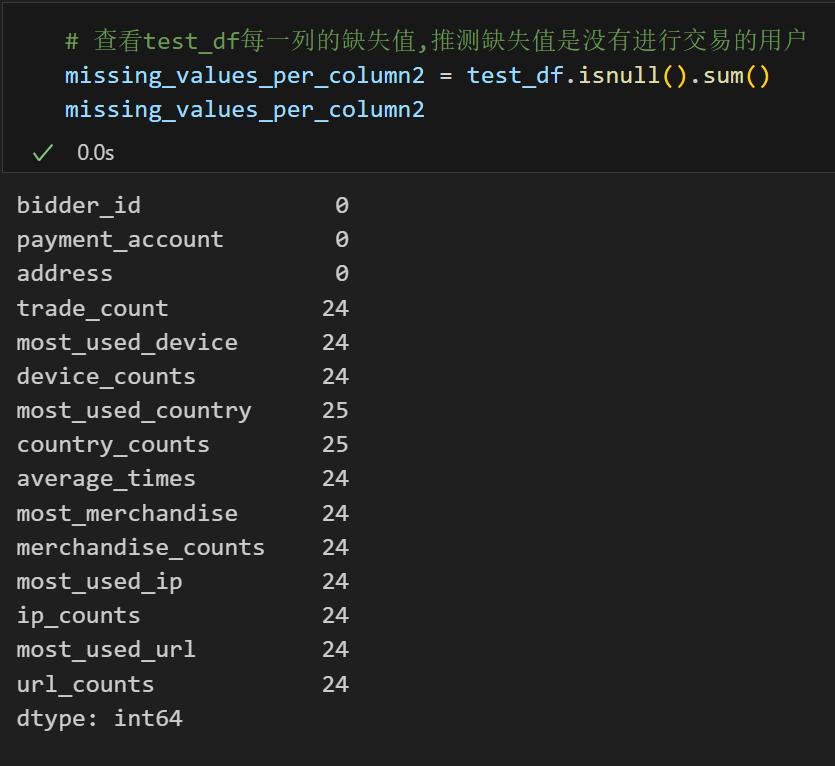
* 1. **数据预处理**
     1. **缺失值处理**

1. **查看是否有缺失值：**

训练集train\_df存在缺失值:

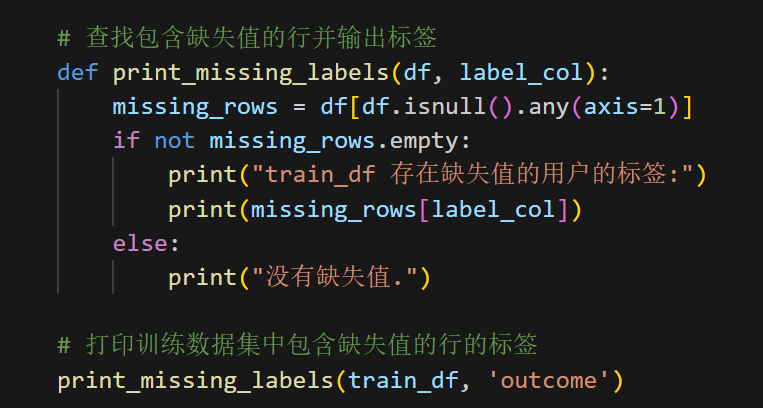


训练集test\_df存在缺失值:

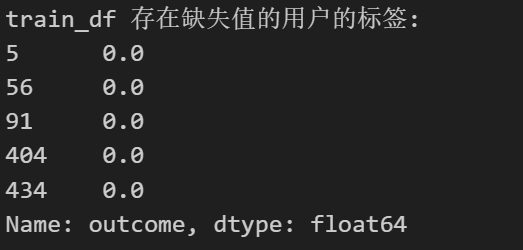


从缺失值所在列和数量观察，train\_df和test\_df中存在缺失值的用户**均为未参加过竞拍出价的用户**。

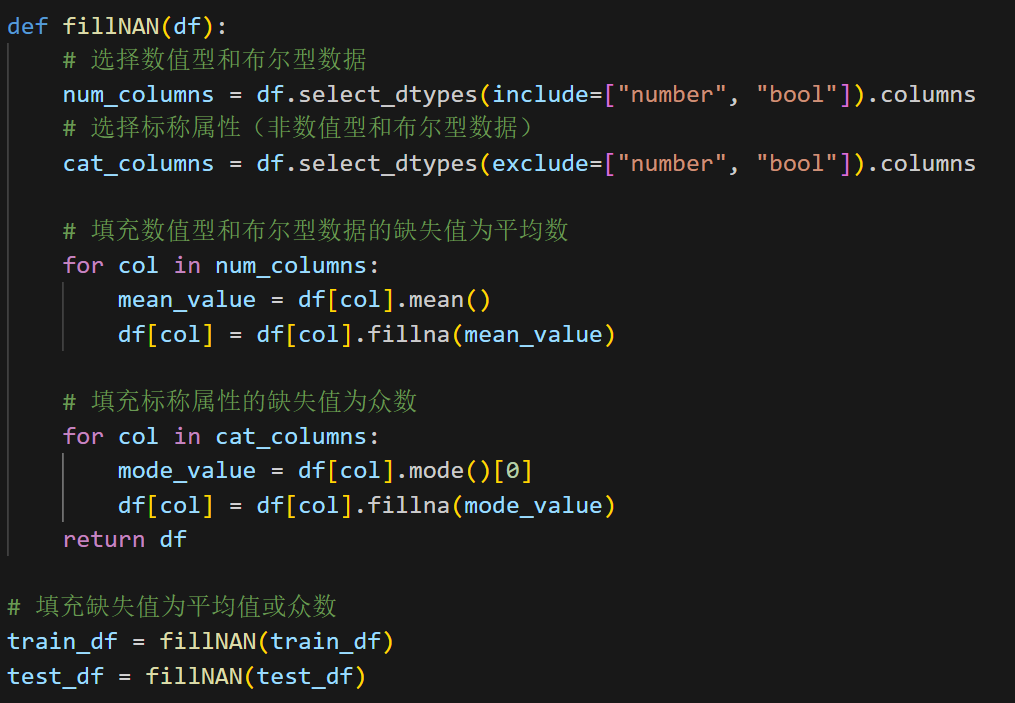
查看train\_df中存在缺失值（未参加竞拍）的用户的标签值：



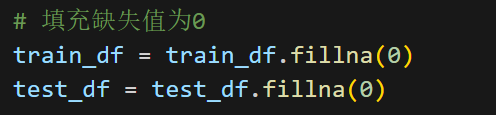
**发现train\_df中存在缺失值（未参加竞拍）的用户均不是机器人：**



1. **填充缺失值：**
   1. **将数值属性的缺失值填充为平均值，标称属性的缺失值填充为众数：**



* 1. **将所有缺失值填充为0：**



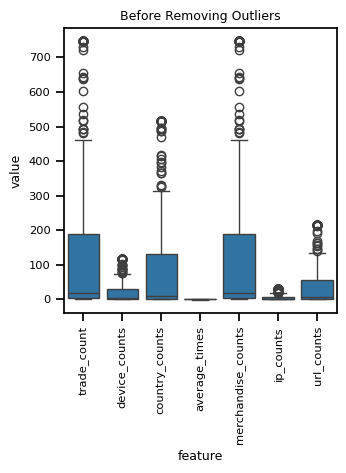
由于缺失值用户为未交易用户，trade\_counts填充为0是正确的，average\_times已标准化，平均值也为0，所以填充为0也就是填充为平均值；

标称属性填充为0可以当作一个新的类别，保留缺失信息，并在模型训练时提供更多的信息，便于判断test\_df中存在缺失值的未交易用户。

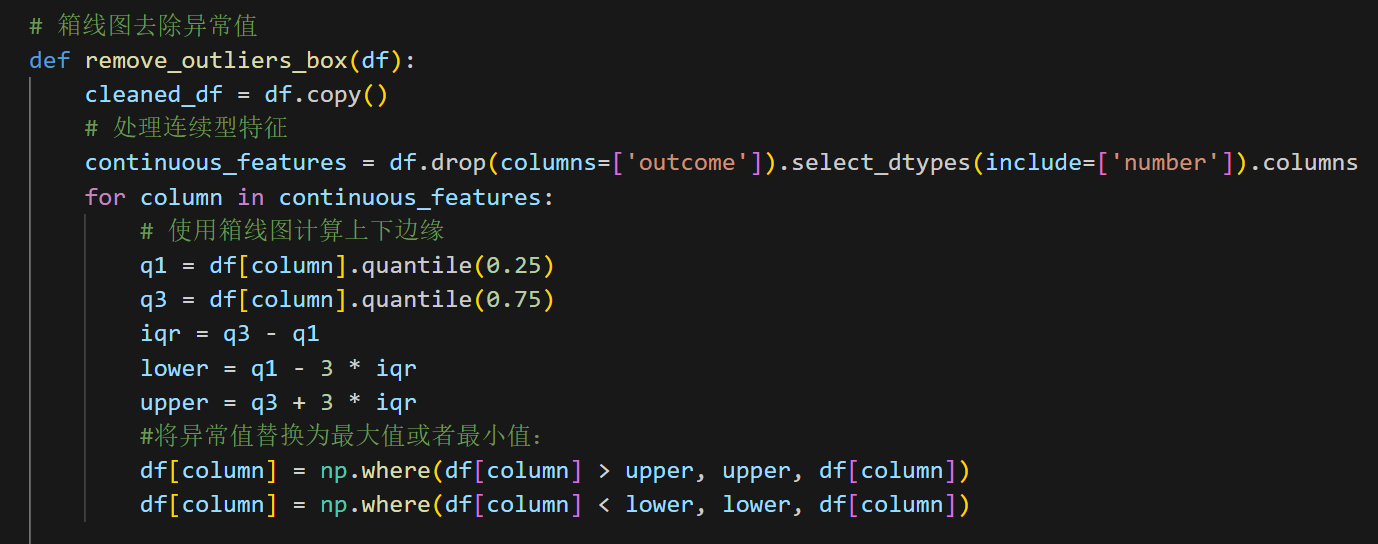
经过多次测试，比较两种填充方案，决定把缺失值填充为0。

* + 1. **去除异常值**

**首先画出箱线图查看数据分布情况：**

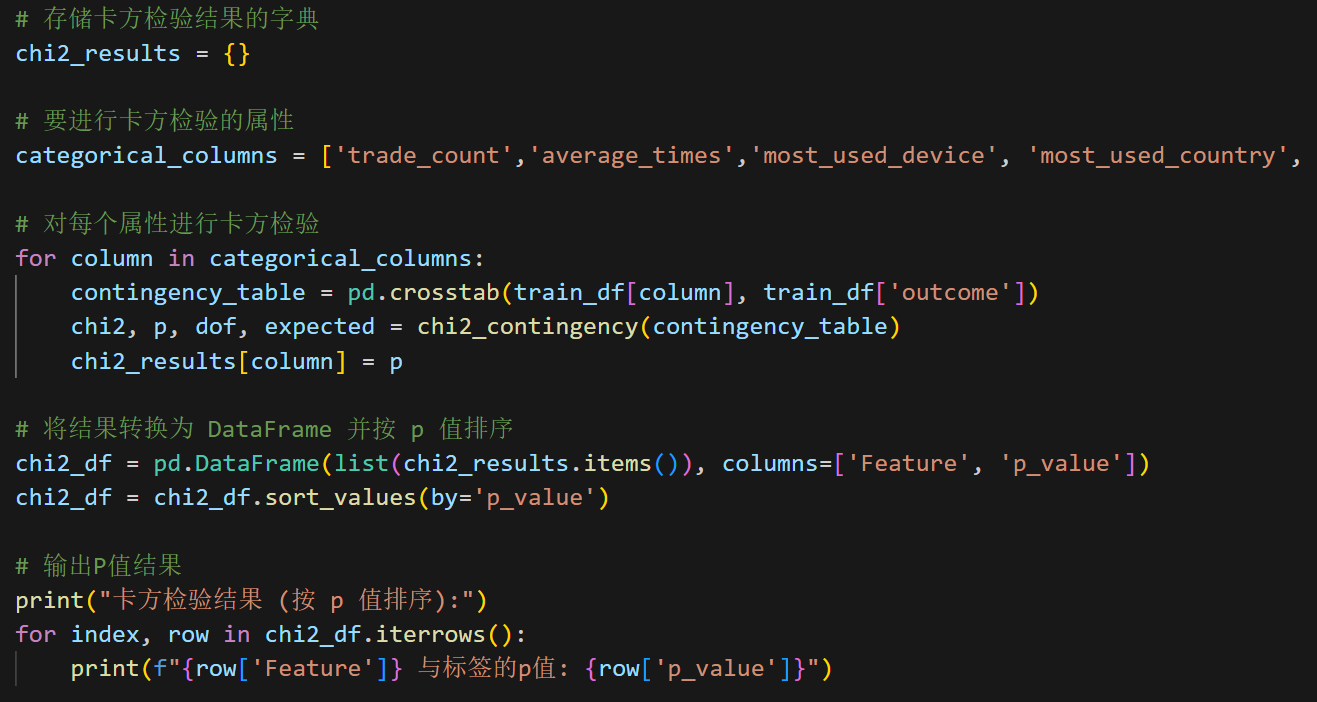
****

**使用箱线图去除异常值，并且将异常值替换为最大值或最小值：**

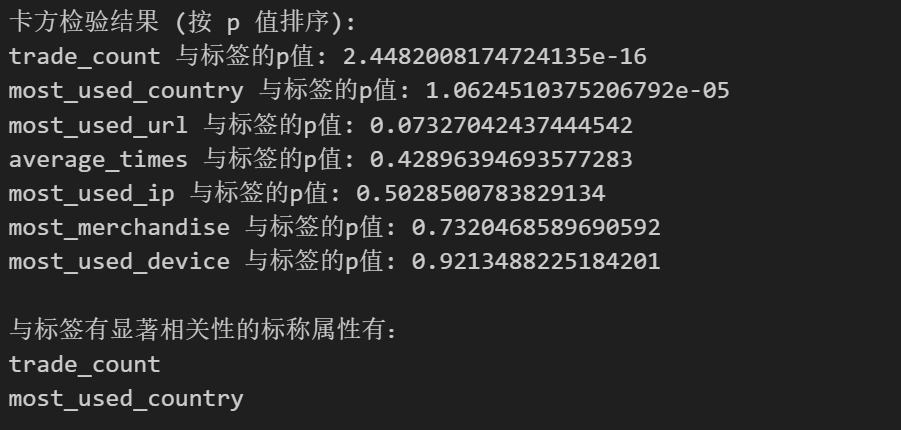
****

* + 1. **相关性分析**

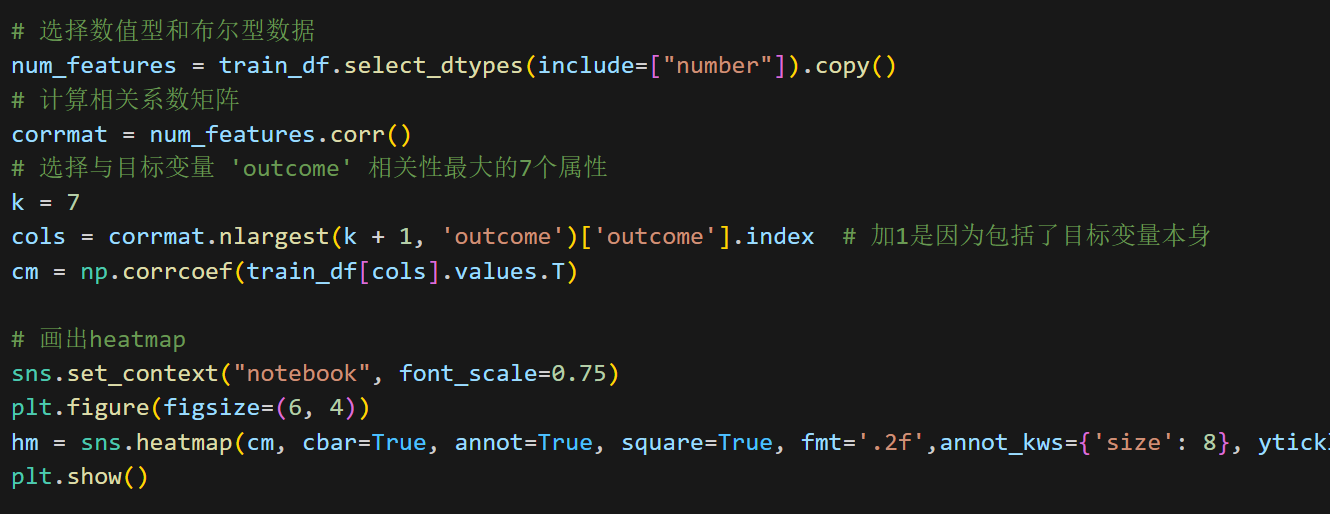
1. **通过卡方检验计算标称属性、trade\_count,和average\_times与outcome之间的相关性：**



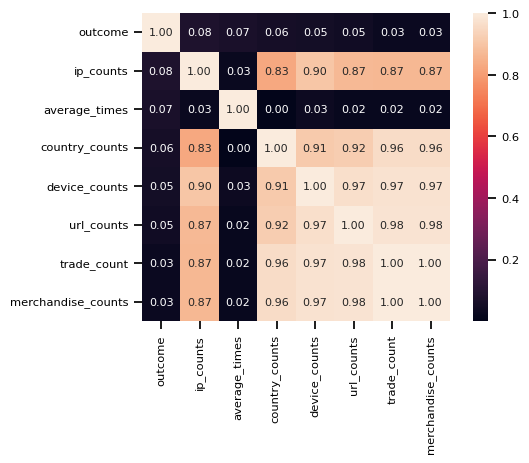


****

1. **计算所有数值属性与outcome之间的相关性并输出相关系数矩阵：**

****

**输出结果：**

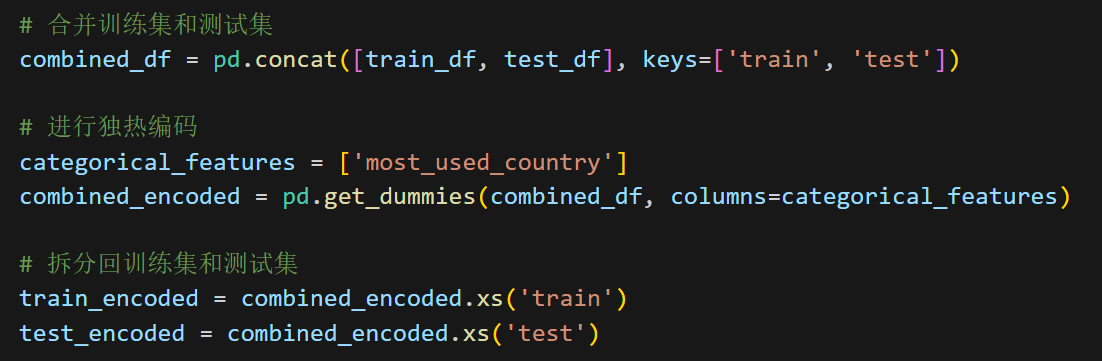
****

* + 1. **独热编码**

**1、将与outcome相关性较强的属性进行独热编码：**

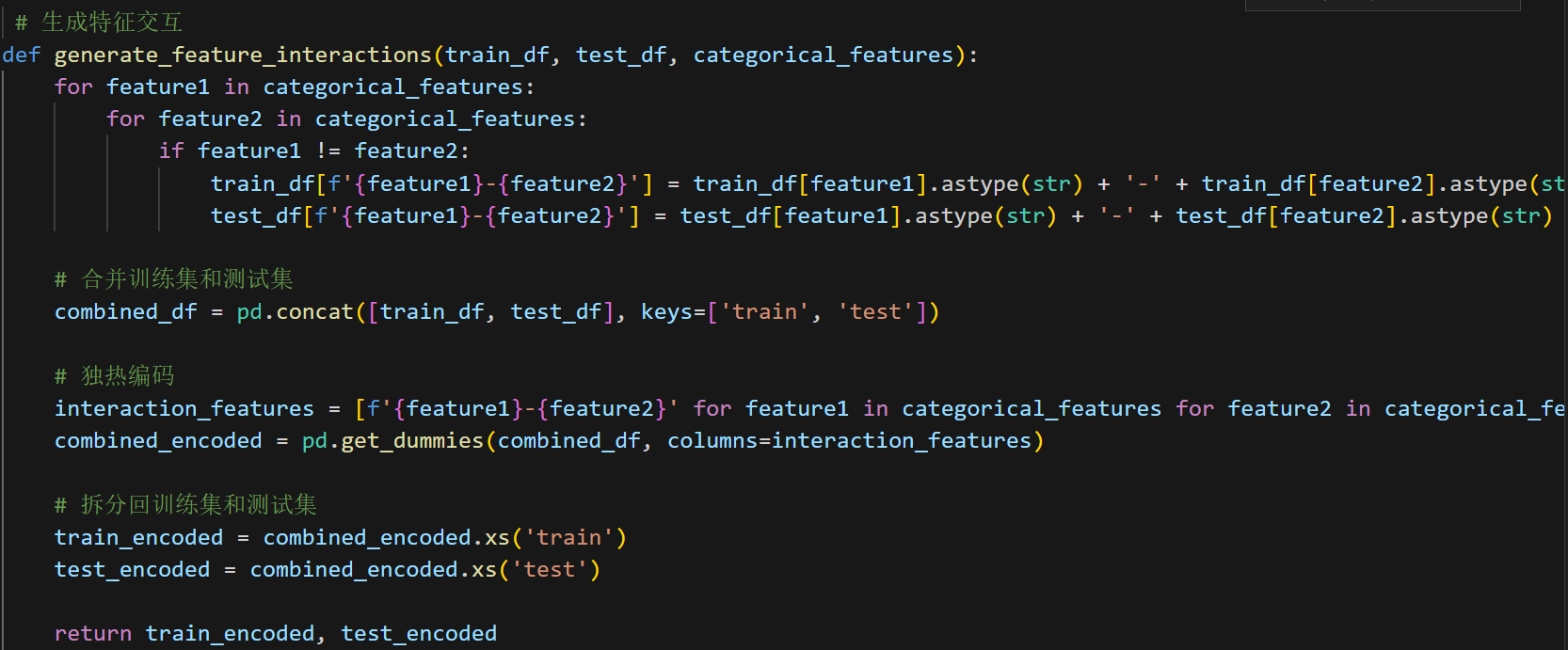
当分别对train\_df和test\_df进行独热编码时，由于两个数据集包含的数据不同，某些类别可能只在其中一个数据集中出现，导致编码不一致，在训练和与测试输入的数据不一致出现错误。所以合并后进行编码，确保训练集和测试集中相同的类别具有相同的编码。

将两个数据集合并后对most\_used\_country进行独热编码，再将训练集和测试集分开：



**2、特征交互：**

函数 generate\_feature\_interactions（），用于生成不同特征之间的交互特征，并对这些交互特征进行独热编码，返回编码后的训练集和测试集。



①标称属性之间交互：

most\_used\_device 和 most\_used\_country

most\_used\_device 和 most\_used\_ip

1. 连续型属性之间交互：

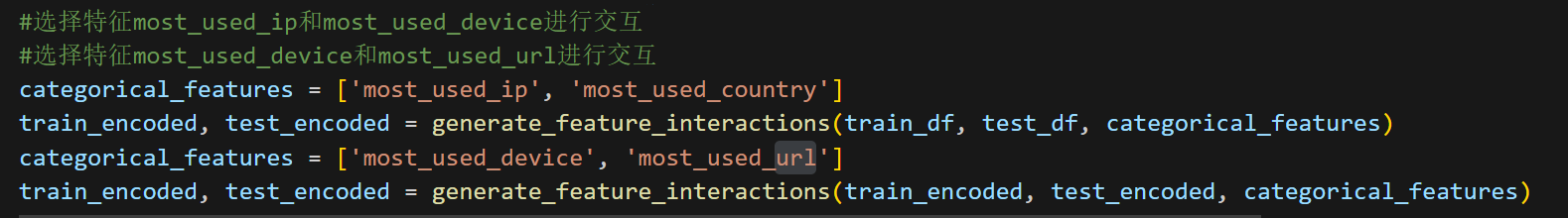
trade\_count 和 average\_times

1. 标称属性与连续型属性交互：

most\_used\_device 和 average\_times

1. 选择重要特征(与outcome相关性较高)进行交互：

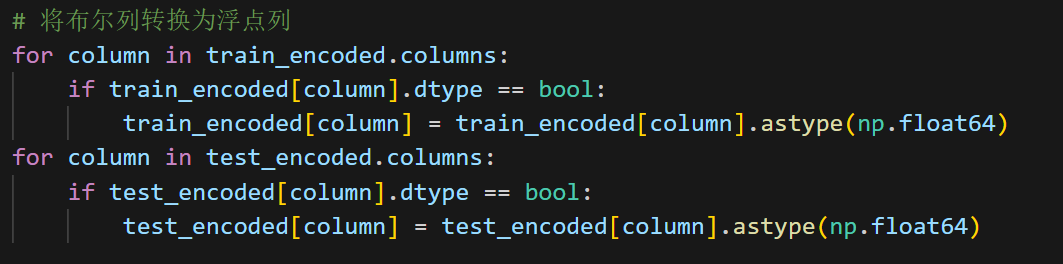
most\_used\_country和most\_used\_url



经过过次尝试，选择most\_used\_device 和 most\_used\_url交互，most\_used\_ip和most\_used\_country交互时模型准确率较高。

* + 1. **数据准备**

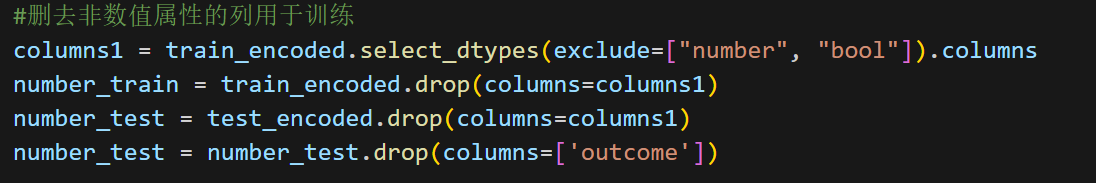
1、将 train\_encoded 和 test\_encoded DataFrame 中的所有布尔类型列转换为浮点数类型；



2、删除非数值属性的列，仅保留数值和布尔类型的列，得到number\_train 和 number\_test：

number\_train准备用于训练模型和模型测试；

number\_test 训练好后用该数据集进行预测是否为机器人。



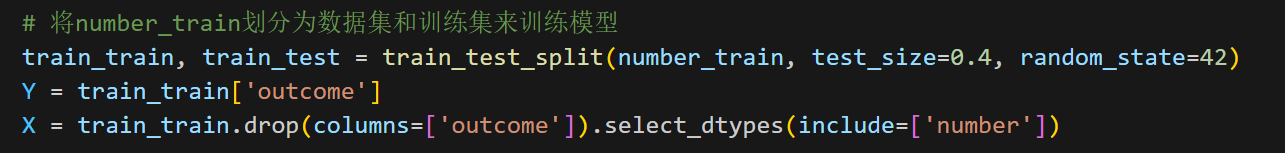
3、将number\_train 划分为训练集train\_train和测试集train\_test:

Train\_train用于训练模型

Train\_test用于计算AUC来测试模型

X为特征数据集

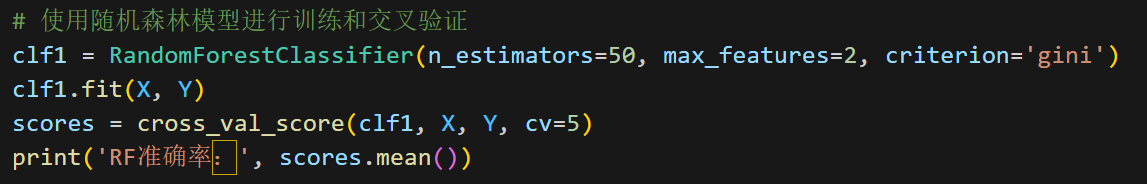
Y标签值

****

* 1. **模型训练、参数优化**
     1. **RandomForest**

**原理：**随机森林是由多个决策树组成的集成模型。每棵树在训练时都从训练集中随机抽取样本（有放回抽样）。每个节点的分裂时，随机选择特征子集，再从中选取最佳特征进行分裂。

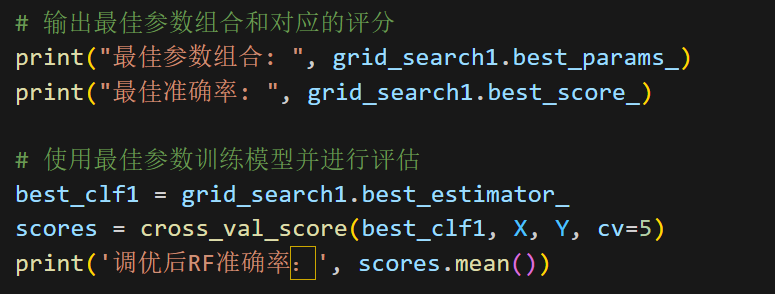
首先使用默认参数训练一个随机森林模型，并进行初步评估：

****

**网格调参：**

定义参数网格param\_grid1，通过网格搜索找到RF最佳参数组合：

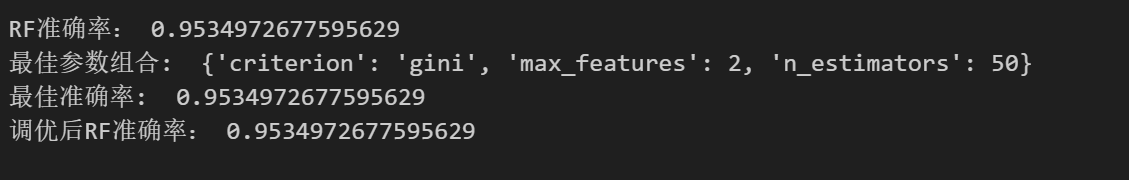
****

****

使用最佳参数重新训练模型并进行评估，输出调优后的模型准确率：



输出结果：

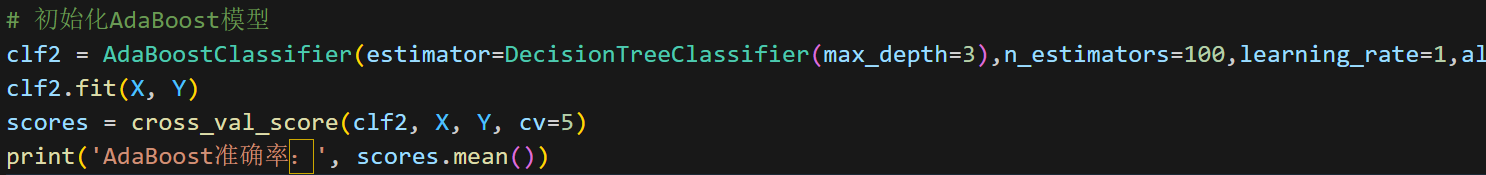


**RandomForest模型：**预测准确率较高且比较稳定，对训练数据的过拟合问题有较好的抗性。

* + 1. **AdaBoost**

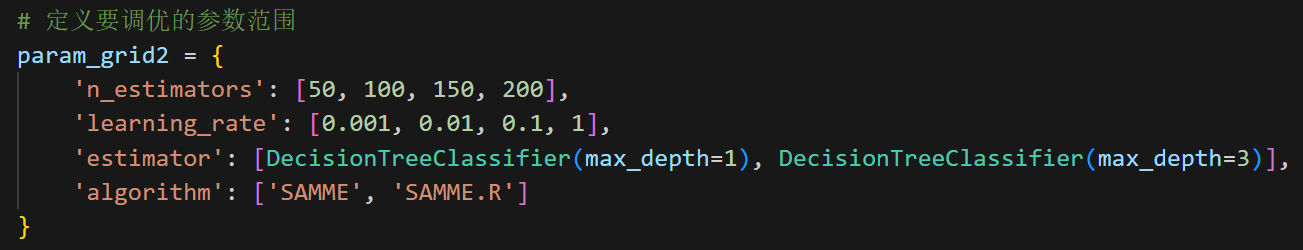
**原理：**AdaBoost通过加权组合多个弱分类器来提升模型的性能。每轮训练时，调整样本的权重，增加被前一轮分类器误分类样本的权重，最终模型是各个弱分类器的加权和。

使用AdaBoost分类器对数据进行训练，并进行交叉验证初步评估模型的性能：

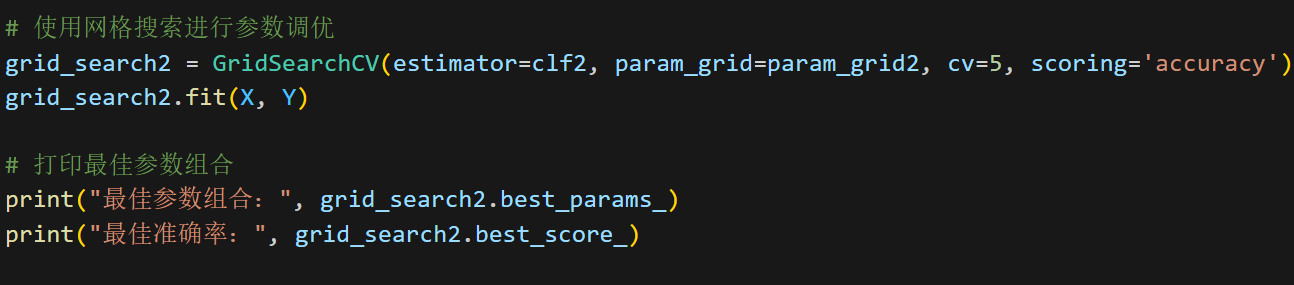
****

**网格调参：**

定义参数网格param\_grid2:



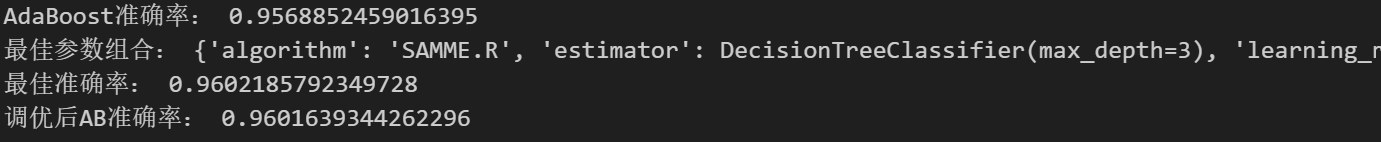
通过网格搜索找到AdaBoost最佳参数组合：



使用最佳参数重新训练模型并进行评估，输出调优后的模型准确率：



输出结果：

****

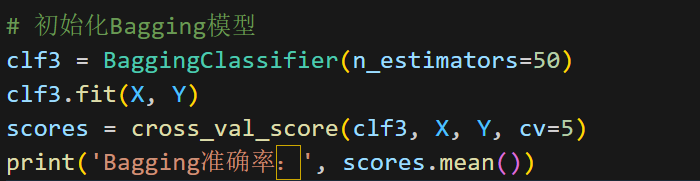
**AdaBoost模型：**准确率较高，但是可能会过拟合，提交实验结果时AUC值不高。

原因可能是该模型关注误分类样本，如果数据中有噪声，AdaBoost可能会过度拟合噪声数据。

**2.3.3 Bagging**

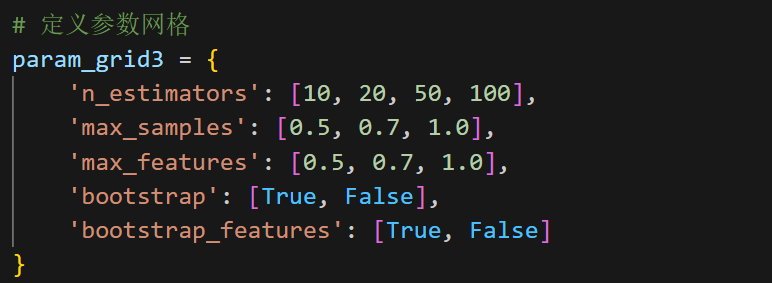
**原理：**基于自助采样法获取数据集，在原始数据集选择T次后得到T个新数据集，通过放回取样得到。基于每个采样集训练出一个基学习器，再将这些基学习器结合。

使用Bagging分类器对数据进行训练，并进行交叉验证初步评估模型的性能：

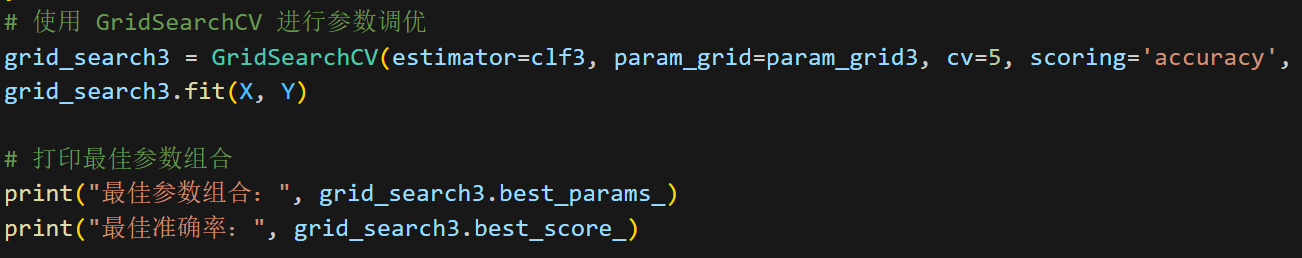


**网格调参：**

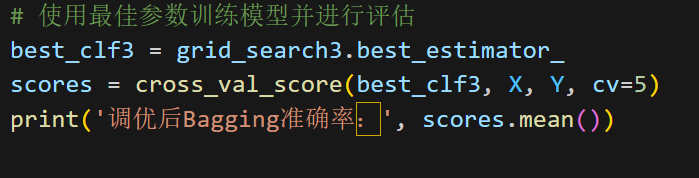
定义参数网格param\_grid3:



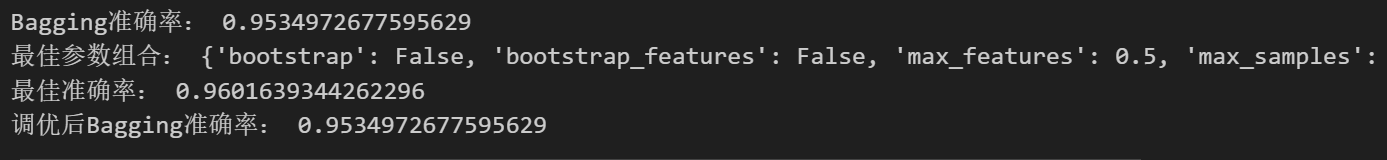
通过网格搜索找到Bagging最佳参数组合：



使用最佳参数重新训练模型并进行评估，输出调优后的模型准确率：



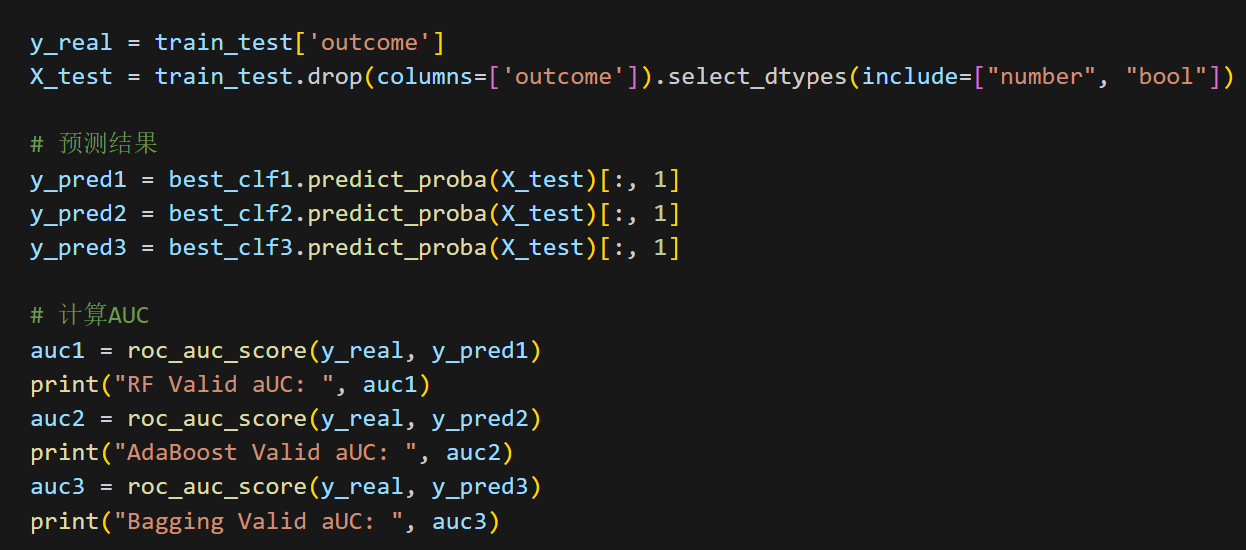
输出结果：



1. **实验结果**
   1. **计算每个模型的AUC值：**

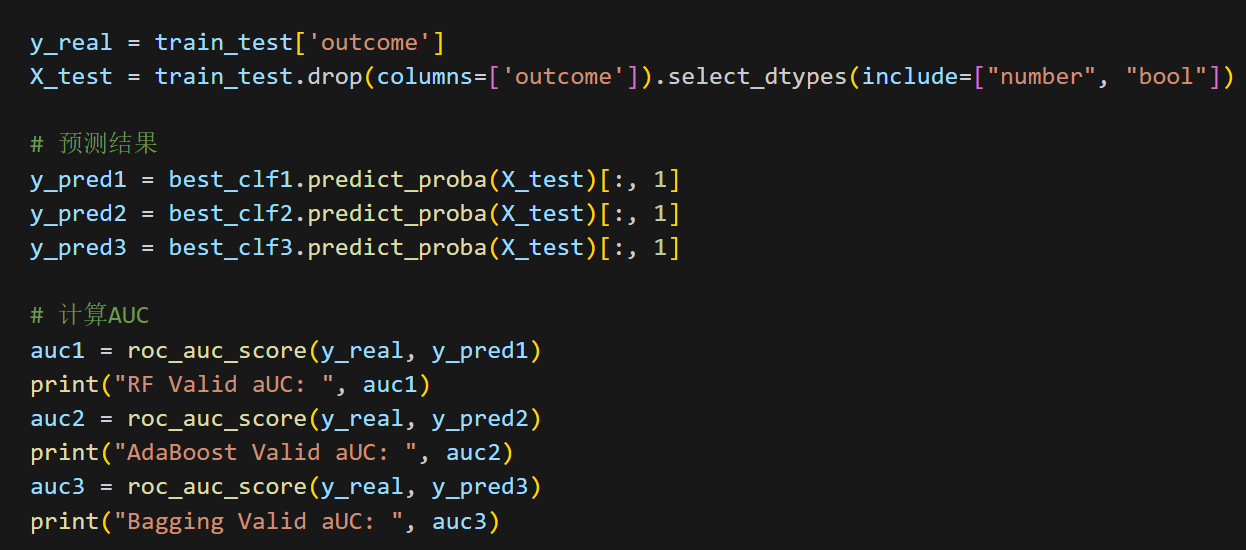
y\_real：从 train\_test中提取标签列 outcome 作为真实值。

X\_test：从 train\_test中删除 outcome 列并选择数值和布尔类型的特征作为测试数据。

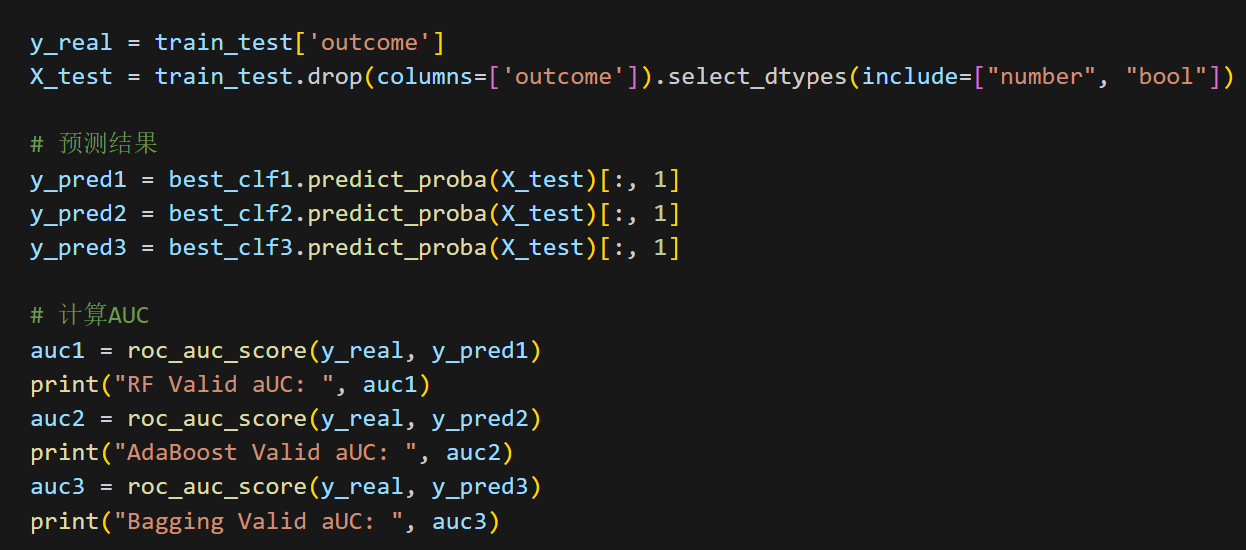


best\_clf1, best\_clf2, best\_clf3 分别是经过调优的随机森林模型、AdaBoost模型和Bagging模型。

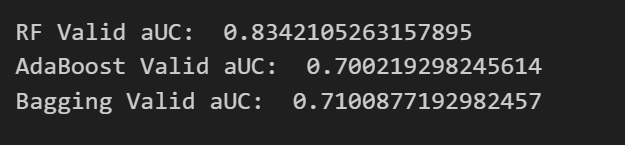
predict\_proba(X\_test) 方法返回每个样本属于每个类别的概率，[:, 1] 提取样本属于正类的概率。



roc\_auc\_score(y\_real, y\_pred) 计算真实标签 y\_real 和预测概率 y\_pred 之间的AUC值：

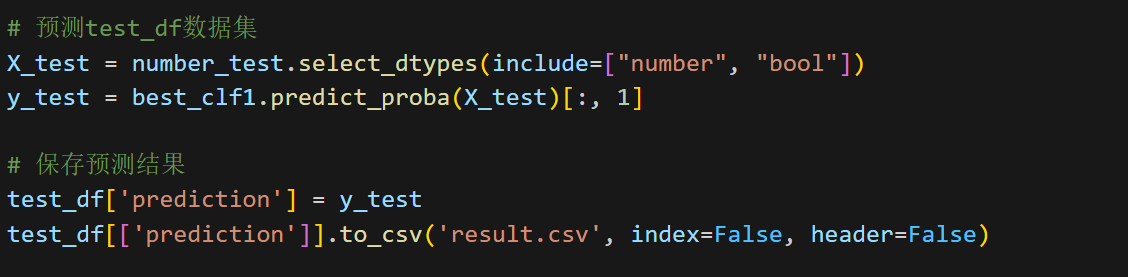


**输出结果：**



* 1. **比较选择适合模型并预测数据：**

多次调试选择，经过比较三个模型的AUC值，大部分情况下，RandomForest模型平均预测效果较好，用其来预测已经处理好的测试集number\_test,并保存结果：

****

1. **实验总结**

通过对特征进行提取、交互和编码，以及对数据进行处理和清洗，在训练集上训练了多个模型，并进行了参数调优。最终，调优后的随机森林模型效果最好，

使用最佳模型对测试集进行预测，得到预测结果。

**实验收获**：

本次实验涉及多方面的数据处理和分析工作，增强了我们对数据的理解和处理能力。通过不同模型的训练和调优过程，熟悉了多种模型的使用和参数调优方法，提升了模型构建的综合能力。

从问题定义、数据处理、模型构建到结果评估，整个实验过程增强了我们系统地解决问题的能力。