**

**HUNAN UNIVERSITY**

《DM&BID》

Report 6

|  |  |
| --- | --- |
| **报告名称：** | 决策树模型 |
| **学生姓名：** | 杨超然 |
| **学生学号：** | 202106060220 |
| **专业班级：** | 电商2102班 |
| **学 院：** | 工商管理学院 |
| **指导老师：** | 江资斌 |
| **日 期：** | 2023.3.6 |

目录

[一、决策树模型代码实现 3](#_Toc2022910766)

[（一）分类决策树模型 3](#_Toc100144858)

[（二）回归决策树模型 3](#_Toc1654932805)

[二、 决策树模型案例-员工离职模型搭建 4](#_Toc247457691)

[（一）读取数据与预处理 4](#_Toc1493072045)

[（二） 模型搭建 4](#_Toc715445120)

[（三） 模型预测评估 5](#_Toc725192287)

[3.1直接预测是否离职 5](#_Toc1337070884)

[3.2预测不离职&离职 5](#_Toc881465180)

[3.3模型预测效果评估 6](#_Toc1443083254)

[3.4特征重要性评估 7](#_Toc219940760)

[三、 参数调优: K折交叉验证与GridSearch网格搜索 8](#_Toc724996833)

[（一） K折交叉验证 8](#_Toc199559153)

[（二） GridSearch网格搜索 8](#_Toc1768711504)

[2.1单参数调优 8](#_Toc1265605954)

[2.2参数调优效果检验 8](#_Toc213745343)

[四、 知识小记 10](#_Toc1825322017)

[（一）相关概念 10](#_Toc1383242324)

[（二） 决策树的建树依据 11](#_Toc1643260693)

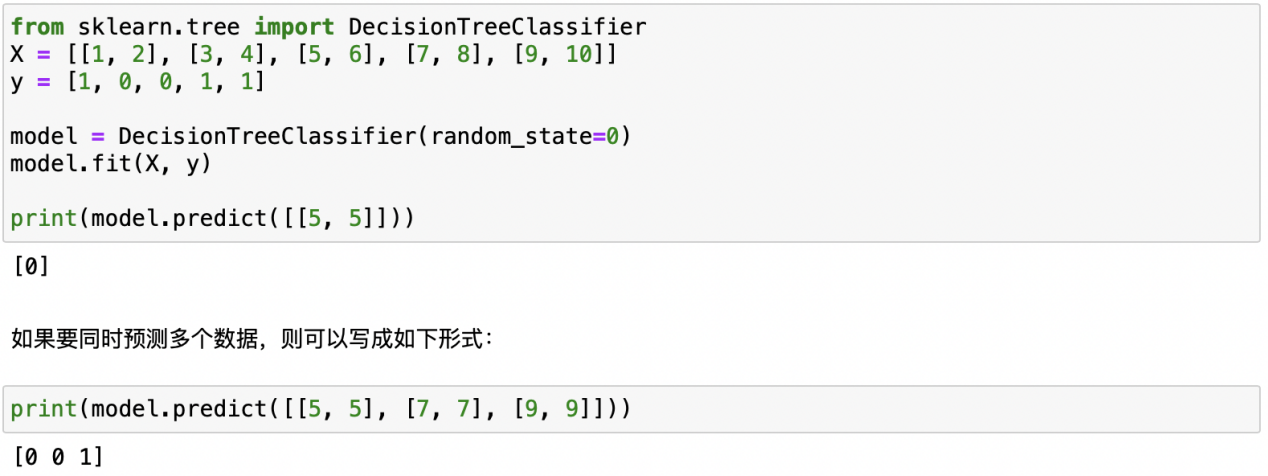
[（三） K折交叉验证 13](#_Toc1642766831)

[（四） GridSearch网格搜索 13](#_Toc1932362785)

[（五） 逻辑回归模型与树模型优劣对比 14](#_Toc826133914)

# 一、决策树模型代码实现

## （一）分类决策树模型



## （二）回归决策树模型



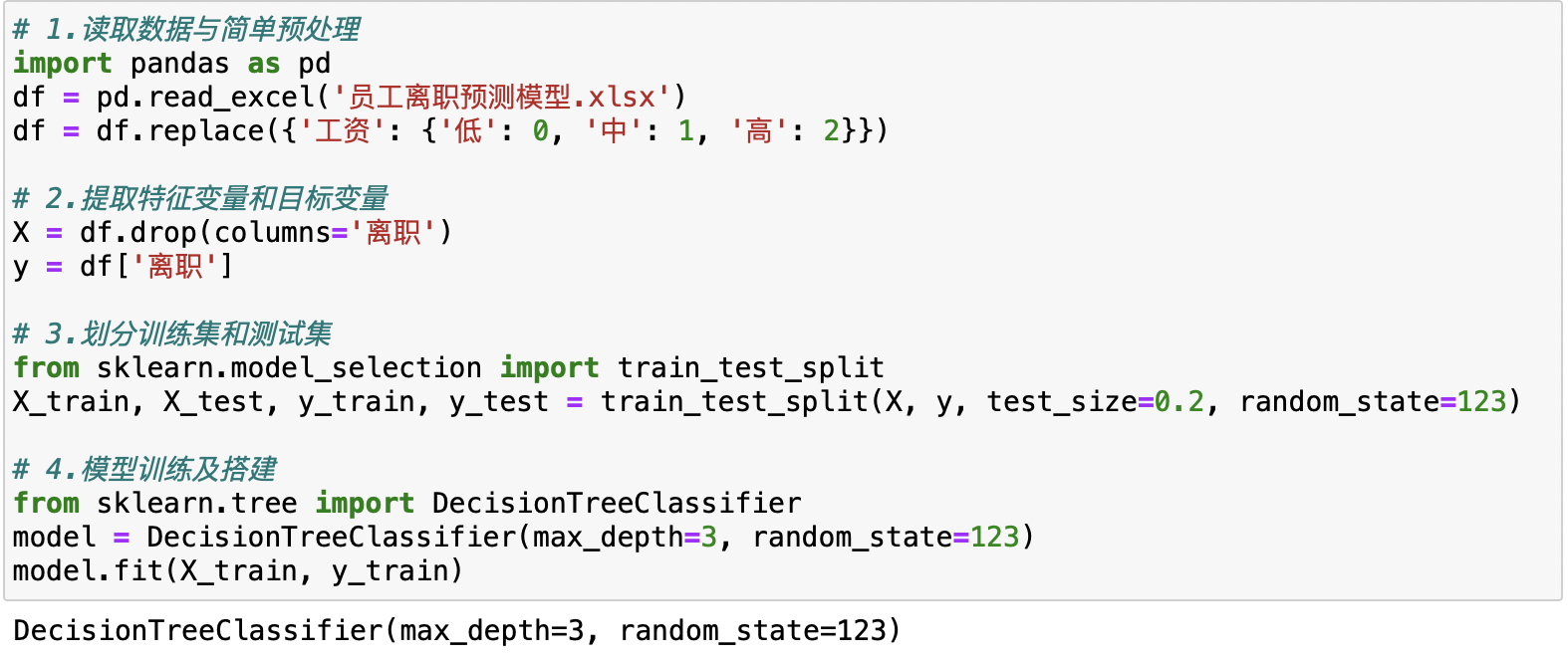
# 决策树模型案例-员工离职模型搭建

## （一）读取数据与预处理





## 模型搭建

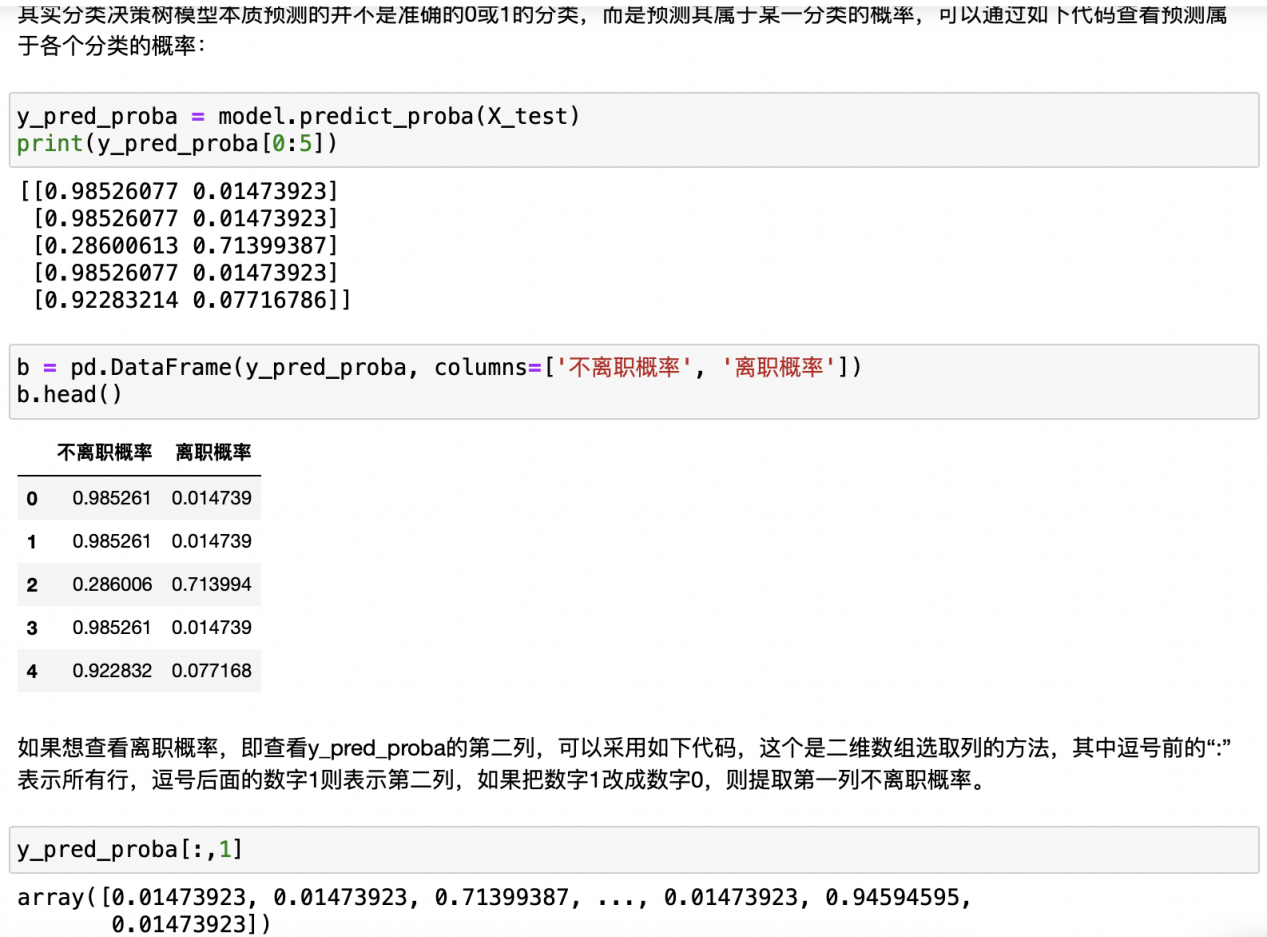


## 模型预测评估

3.1直接预测是否离职

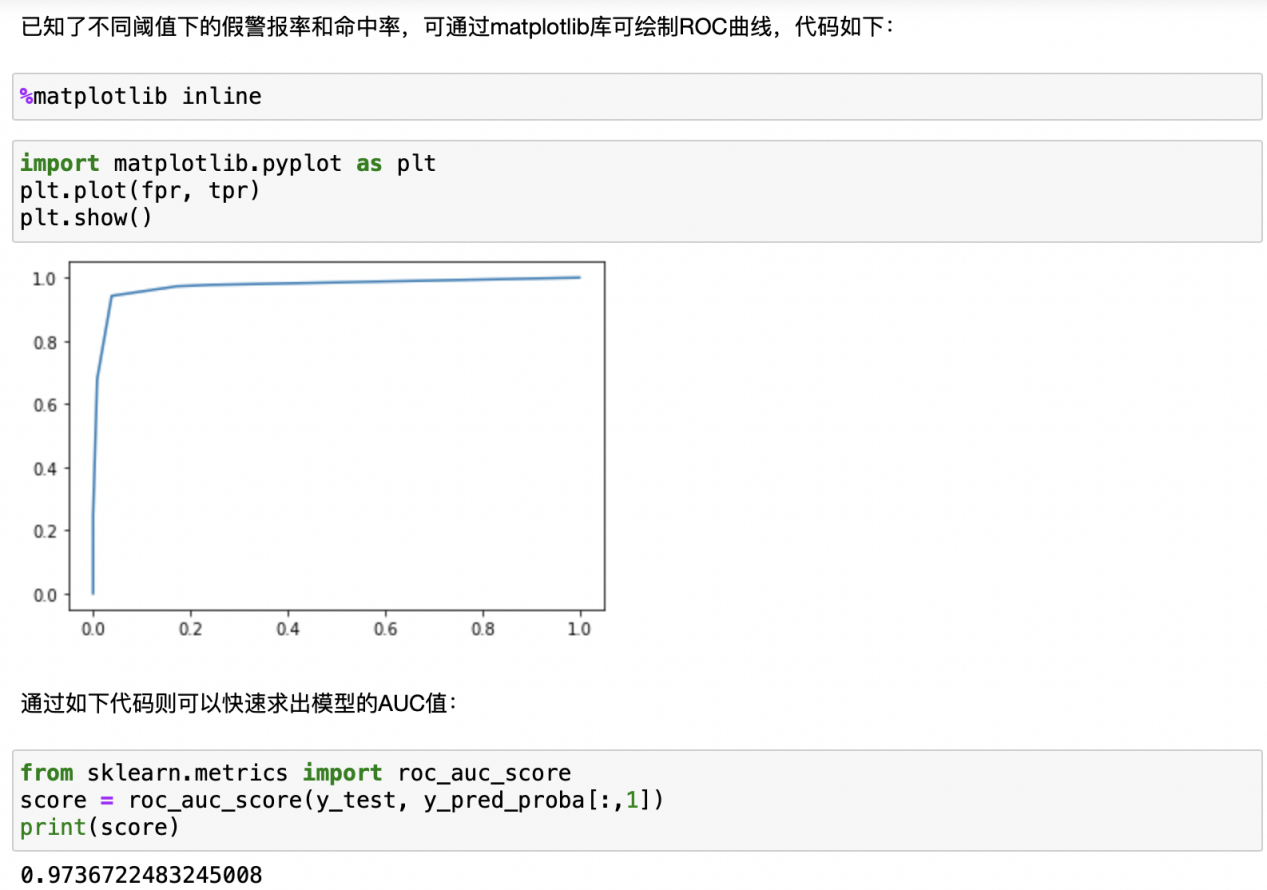


3.2预测不离职&离职

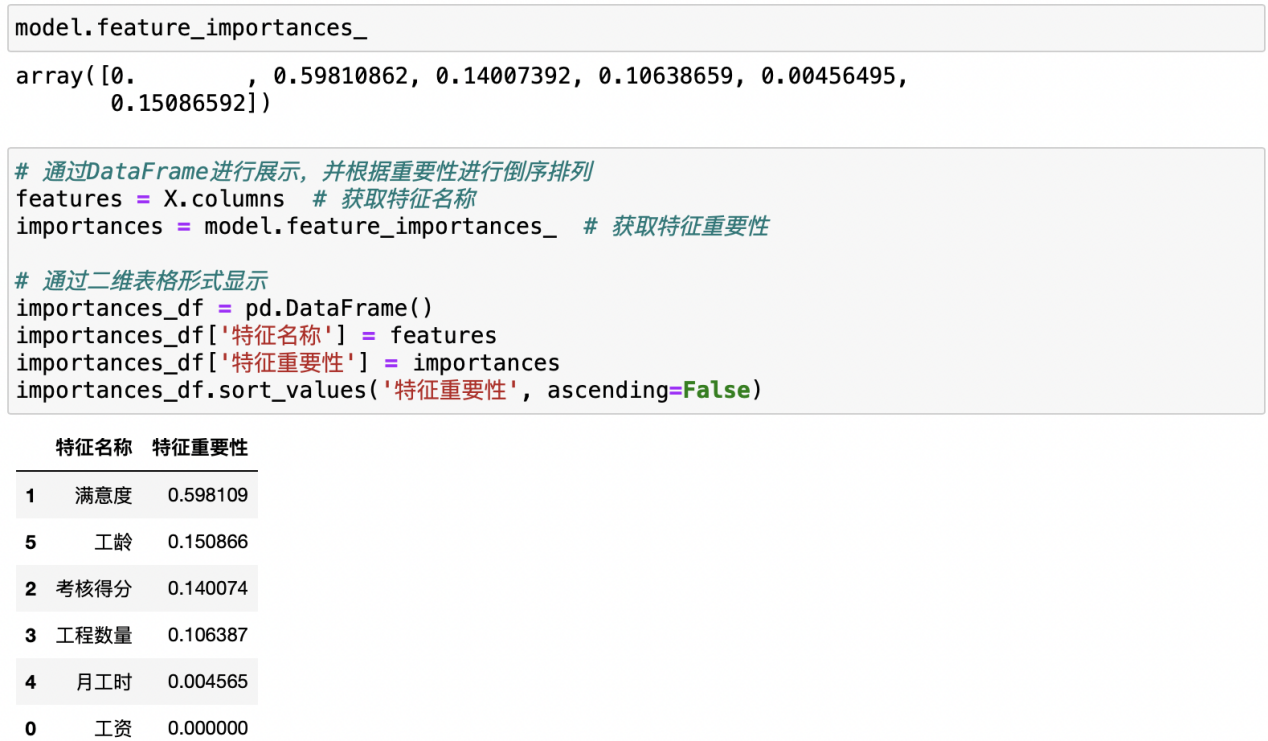


3.3模型预测效果评估



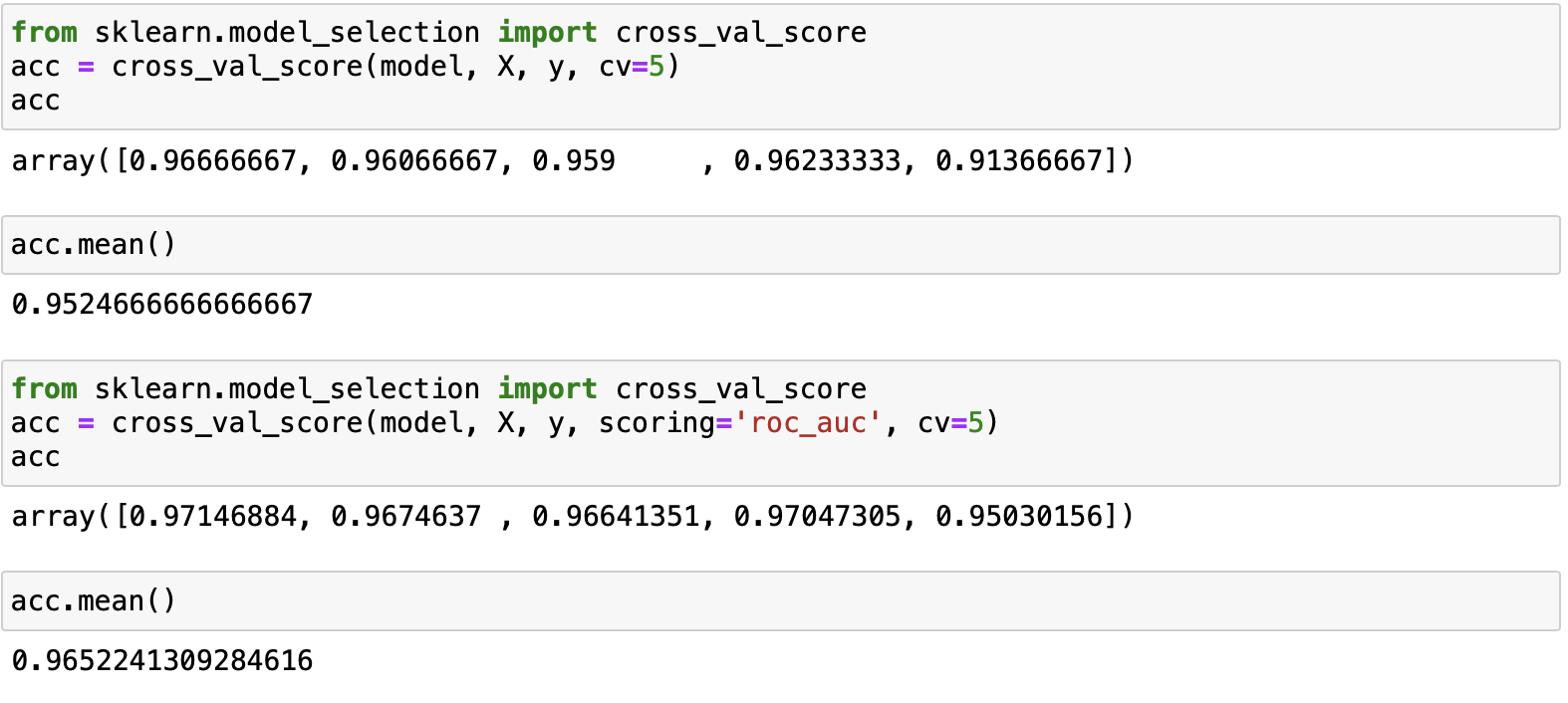


3.4特征重要性评估



# 参数调优: K折交叉验证与GridSearch网格搜索

## K折交叉验证

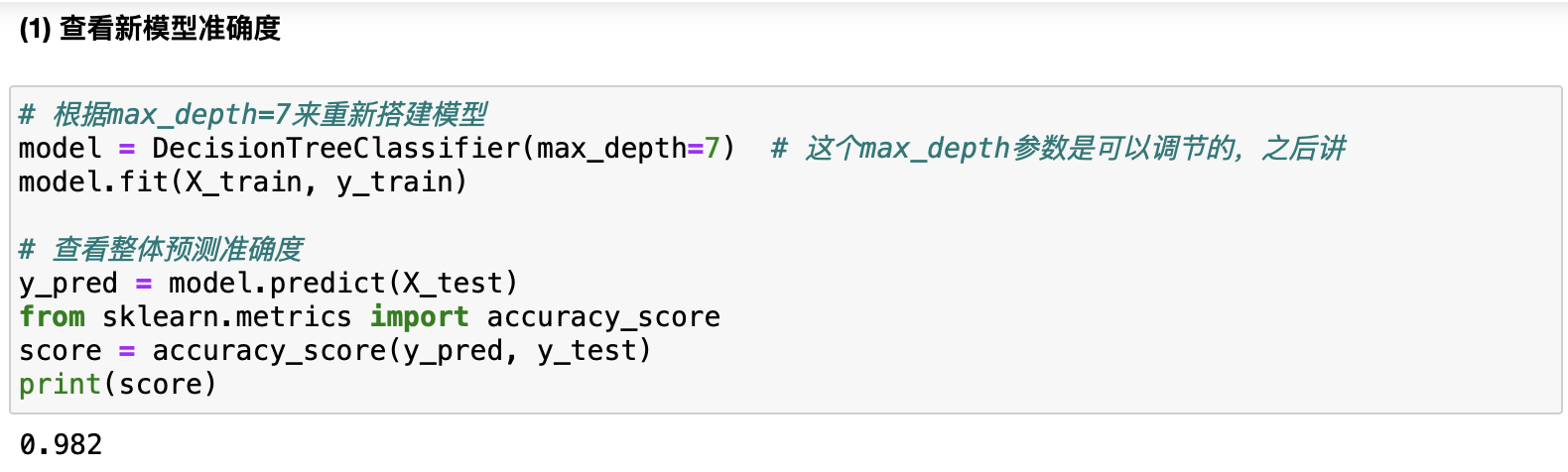


## GridSearch网格搜索

2.1单参数调优



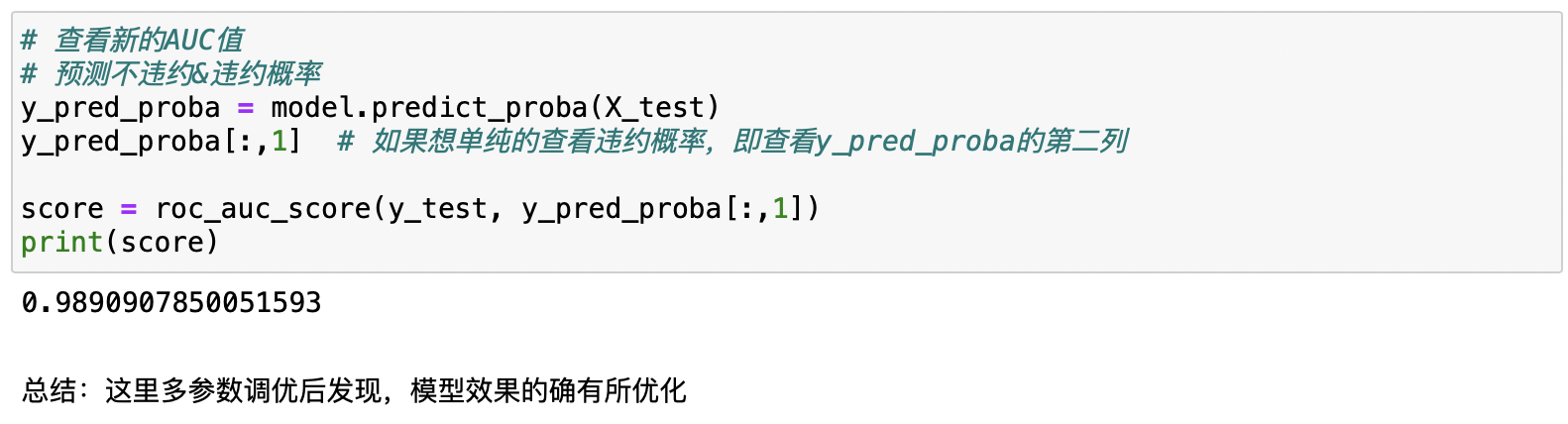
2.2参数调优效果检验





2.3多参数调优





# 知识小记

## （一）相关概念

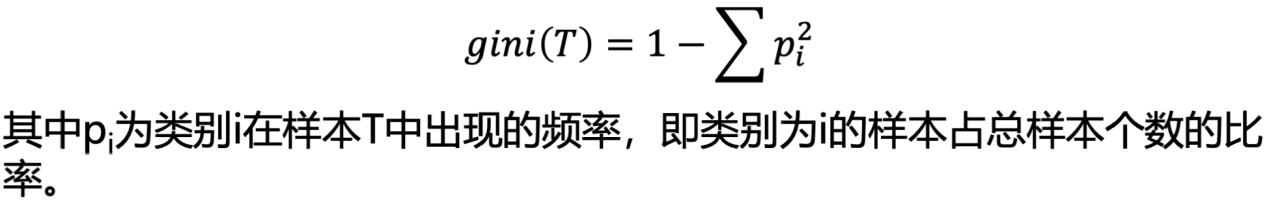
***1.1父节点&子节点***：子节点由父节点分裂而来，然后作为父节点继续分裂，直到不能分裂为止

***1.2根节点***：没有父节点的节点，即初始节点

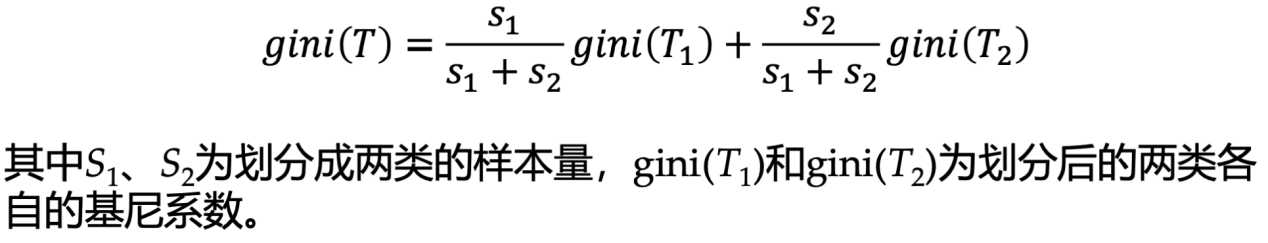
***1.3叶子节点***：没有子节点的节点，即最后的节点

## 决策树的建树依据

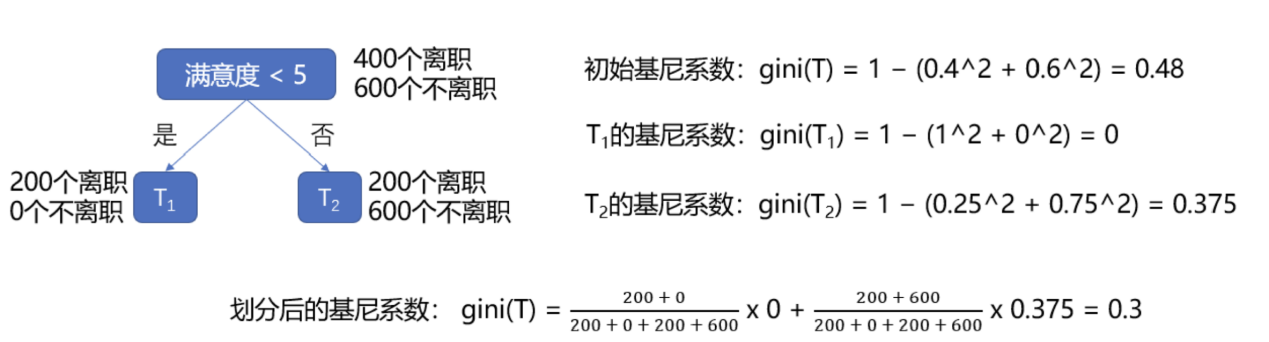
***2.1基尼系数***：用于**反映系统的混乱程度**，gini越大，系统的混乱程度越高，采用gini运算的决策树称为**CART决策树**：

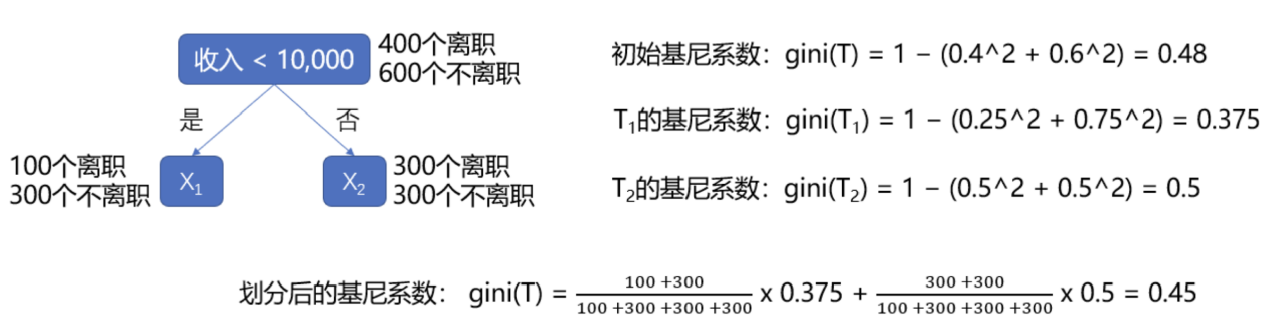


经分类后的基尼系数：当决策树依据某条件进行划分后，gini的计算方式如下：



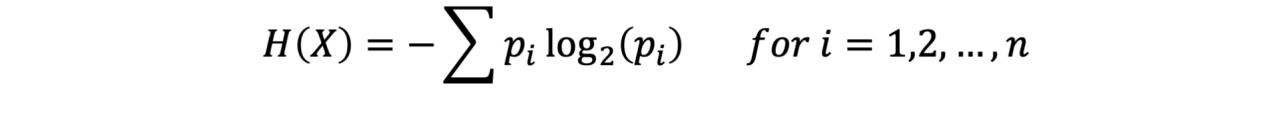
可通过下例了解分类gini的计算：



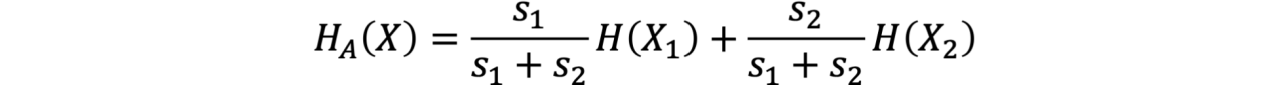


相比之下，以“满意度<5”划分的基尼系数较低，区分度较好，因此选择其作为初始节点。

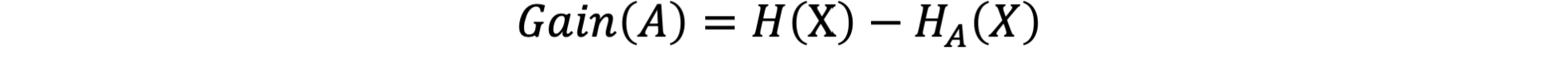
***2.2信息熵***：信息熵也是用于衡量系统的混乱程度，从而进行合理的节点划分，公式如下：

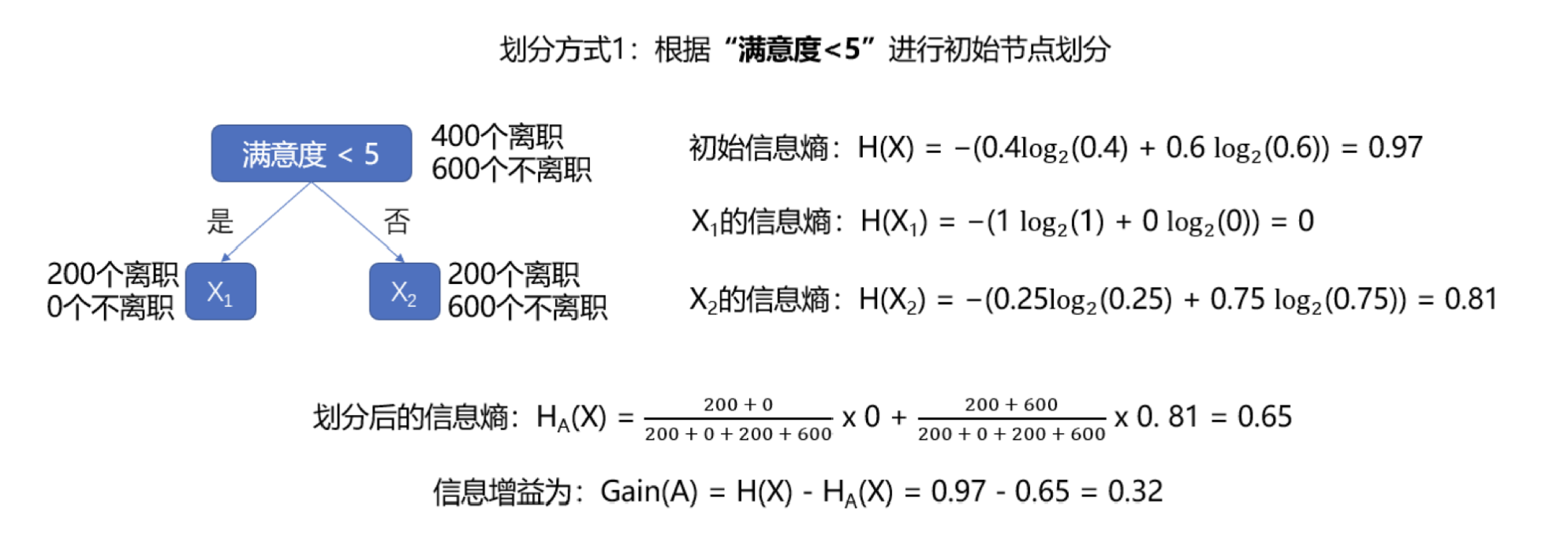


被分类后的信息熵被称为**条件熵**，计算形式类似于分类后的gini：



**信息熵的减少值**（原系统熵-划分后系统熵）称为**熵增益或信息增益Gain(A)**,**其值越大说明分类后的混乱程度越低，即分类越准确**：

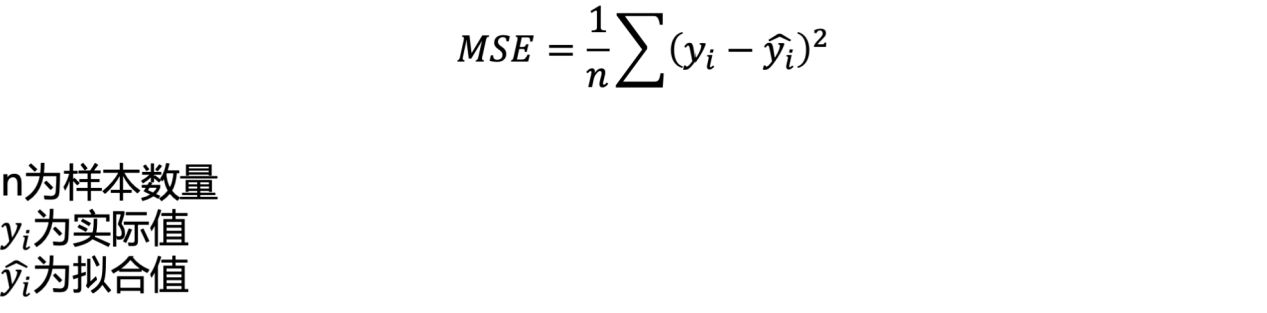






相比之下，根据方式1划分后的信息增益为0.32，大于根据方式2划分后的信息增益 0.046，因此我们选择根据方式1来进行决策树的划分。

2.3回归决策树模型的切分依据：**均方误差（MSE）**

****

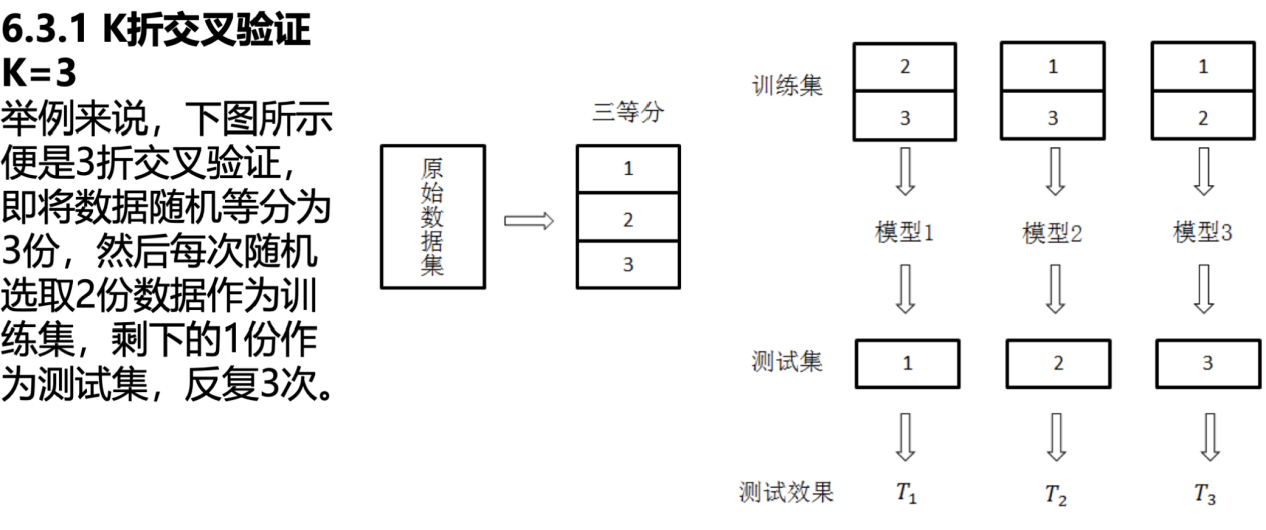
决策树**停止分裂的条件**：已经结束无法再分裂&达到极限分裂条件

## K折交叉验证

***3.1交叉验证***：对原始样本进行切分，分为多组不同的训练集与测试集，其中某次的训练集可能是下次的测试集，因此称之为交叉验证。

交叉验证有**简单交叉验证、K折交叉验证**与**留一交叉验证**，其中K折运用较为广泛

3.2K折交叉验证：K折交叉验证是指将数据集随机**等分为K 份**，每次选取**K-1份为训练集**训练模型，然后用**剩下的1份作为测试集**，得 到K个模型后将这**K个模型的平均测试效果作为最终的模型效果：**

****

通常来说，如果**训练数据集相对较小，则增大k值**，这样在每次迭代过程

中将会有更多的数据用于模型训练，同时算法时间延长；如果**训练集相对较大，则减小k值**，这样降低模型在不同的数据块上进行重复拟合的性能和评估的计算成本，在平均性能的基础上获得模型的准确评估。

## GridSearch网格搜索

网格搜索是一种穷举搜索的调参手段：遍历所有的候选参数，循环建立模型并对模型的有效性和准确性进行评估，选取表现最好的参数作为最终结果。

## 逻辑回归模型与树模型优劣对比

**逻辑回归模型**：变量不需要太多，不容易过拟合，泛化能力较强，可能1年才换一次模型，但是逻辑回归模型有时不够精确，不能有效剔除潜在违约人员。（**需变少、泛化强、不精确**）

**树模型**（决策树、随机森林、XGBoost等模型）：不太稳定(一个变量可以反复用)，容易造成过拟合，泛化能力较弱，一段时间后换一波人可能就不行了，但它的拟合度强，区分度高，可以快速去掉坏人。（**区分好、拟合强、不稳定、泛化弱**）