1．实验目的

（1）掌握sklearn中决策树方法的调用。

（2）掌握决策树方法基本调参技术。

（3）掌握基于决策树的组合方法，包括Adaboost和 Random Forest

（4）能够根据准确率比较、评价不同的分类模型。

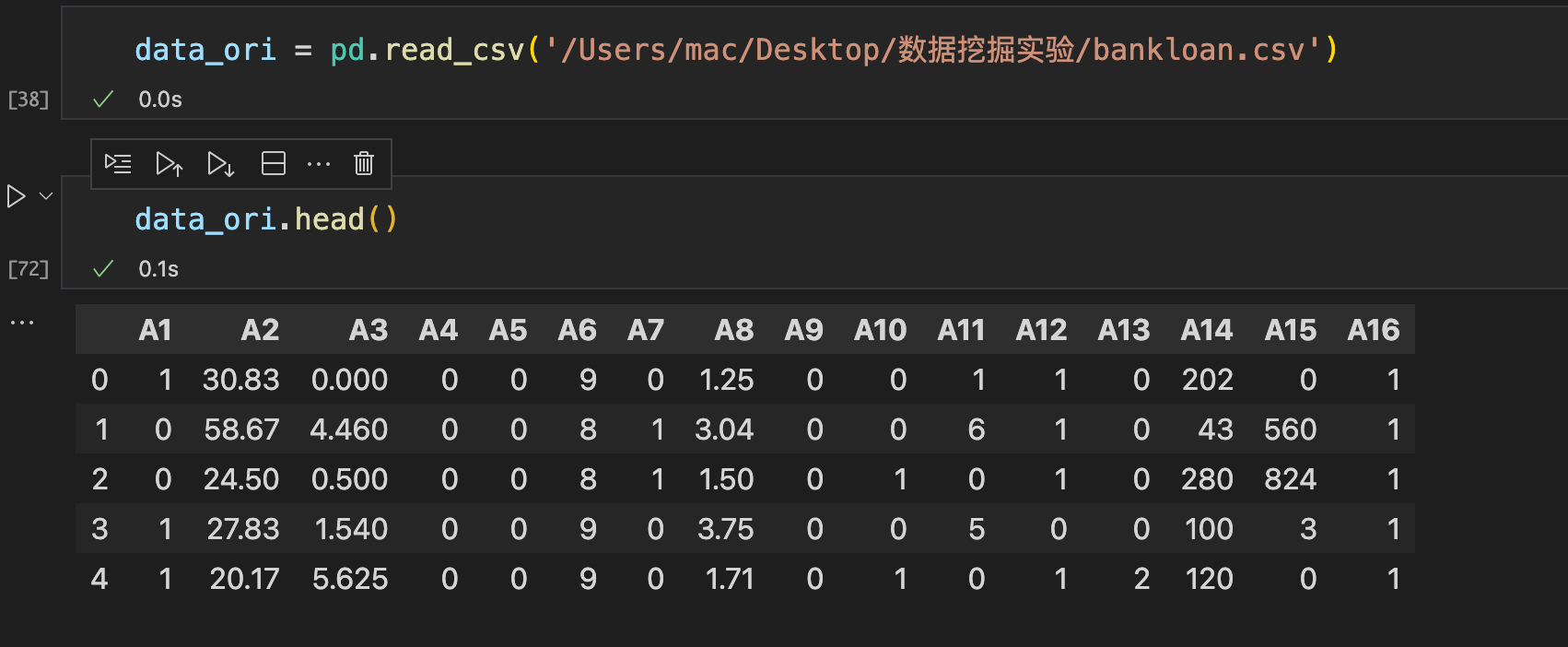
2．实验数据

数据文件bankloan.csv。银行贷款数据，一共包含690个样本，16个属性。基于隐私保护的考虑数据已经脱敏，属性名称用A1，A2，…，A16来表示，其中A16是类标号属性，有两个取值：“+”和“﹣”，“+”表示信贷申请被通过，“﹣”表示信贷申请被否决。需要注意的是，一些属性的值为“NA”，表示值缺失。

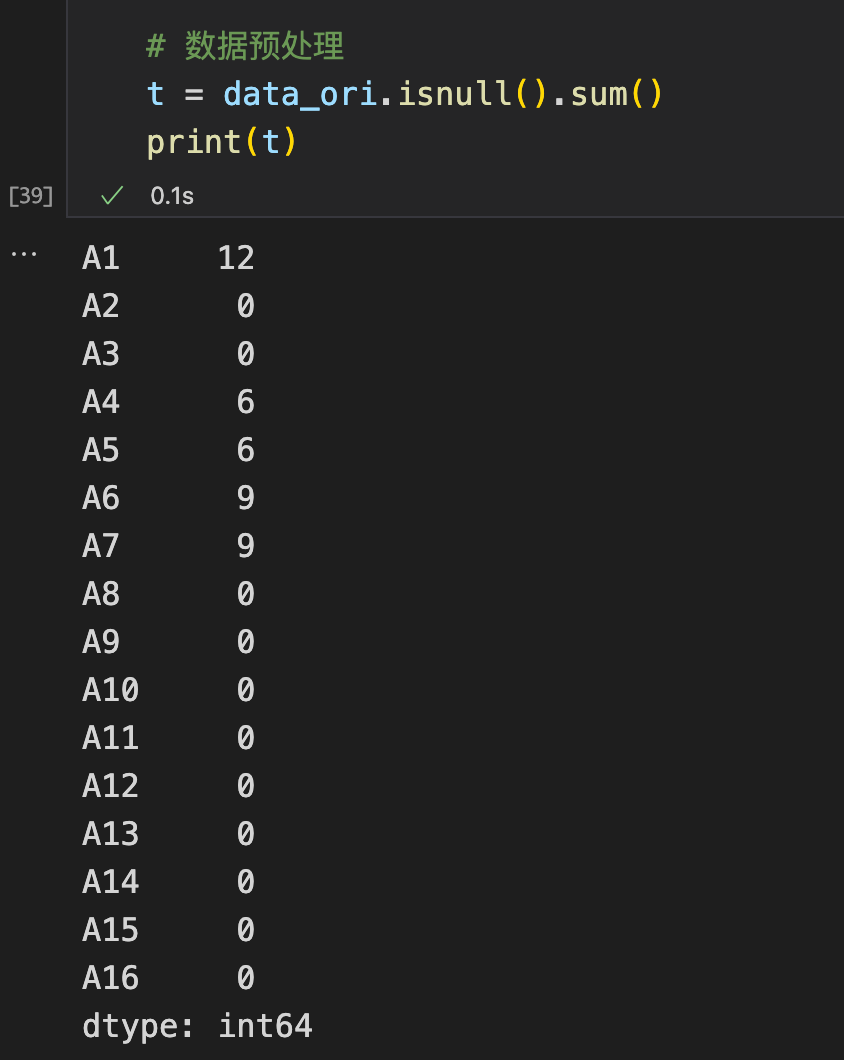
3．实验要求和步骤

（1）读取数据，完成数据预处理

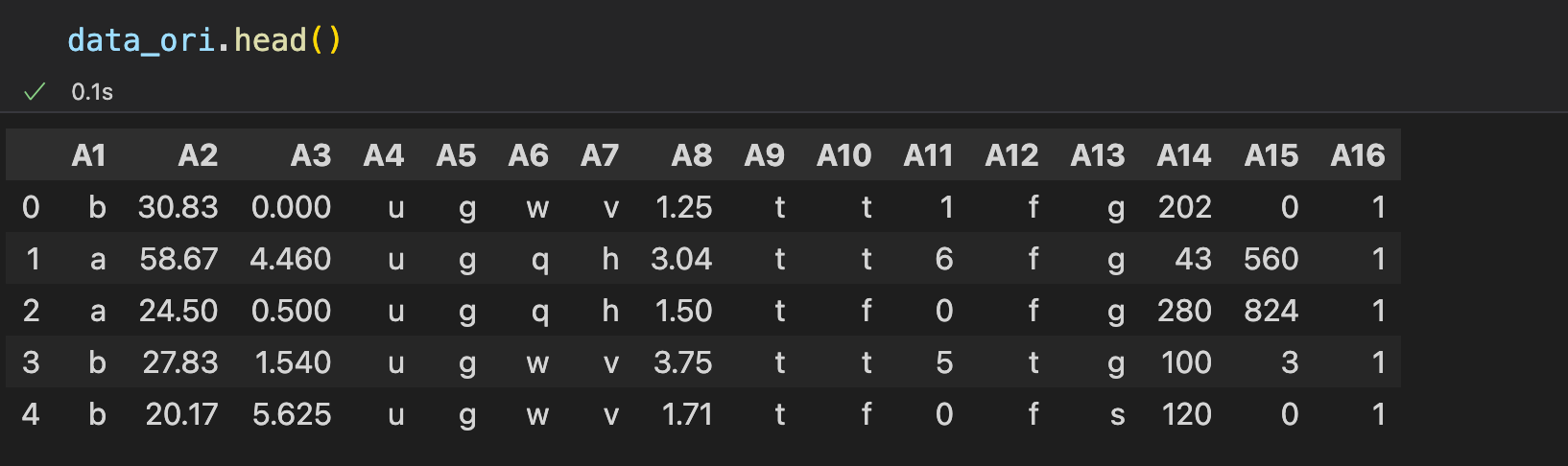
利用Panada库完成对于数据的读取：



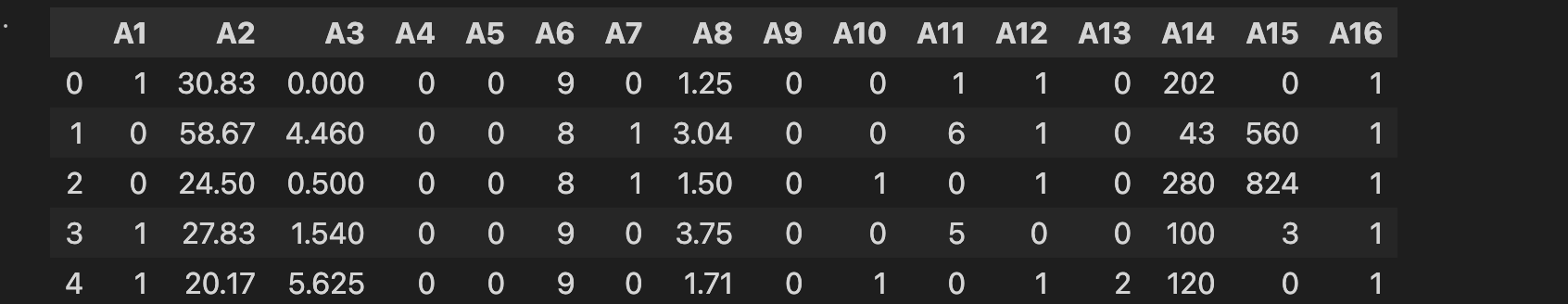
统计各个属性当中缺失值的个数并用众数对于离散数据中的缺失值进行补全:



完成缺失值处理之后得到的数据

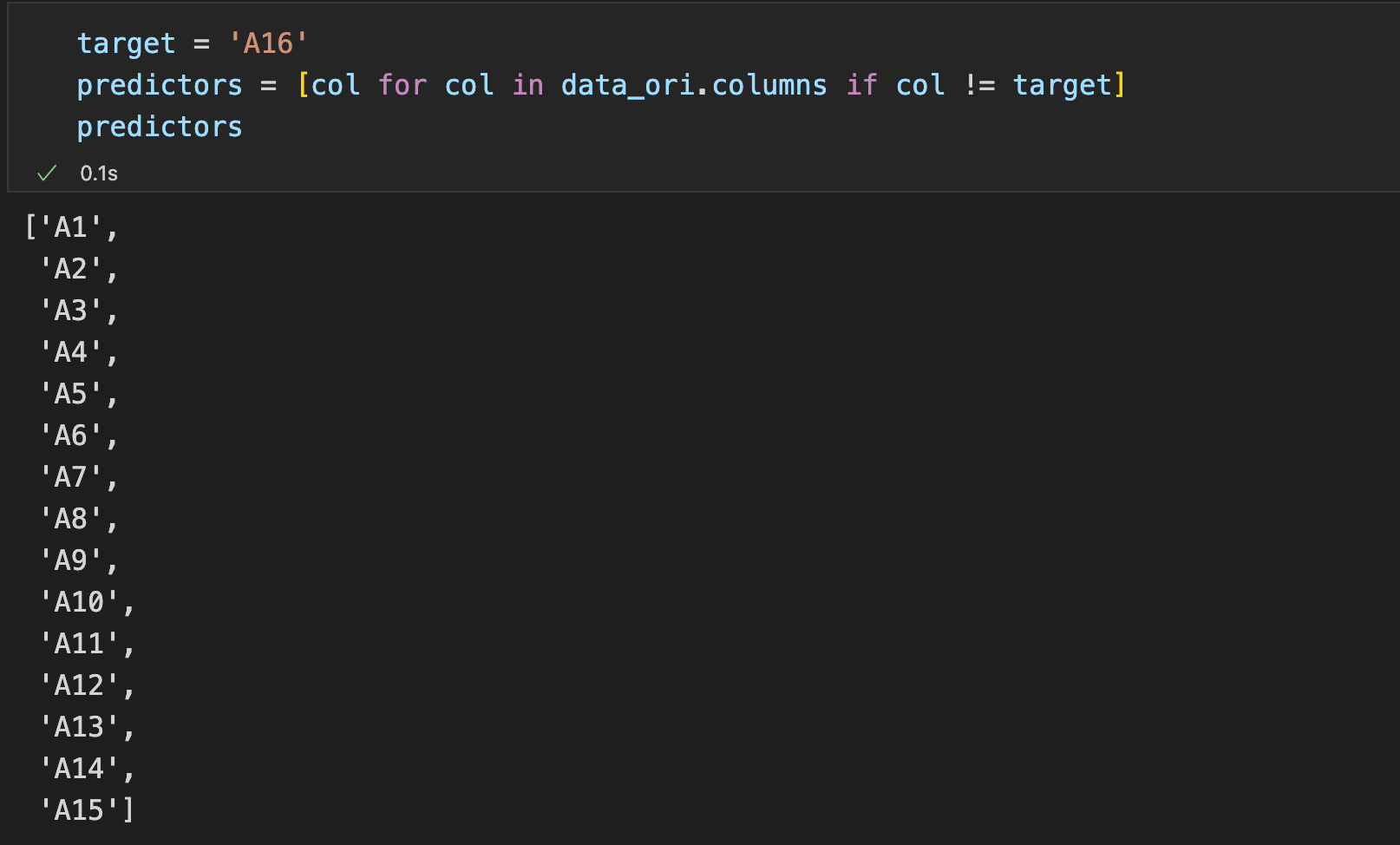


对于非数据化的数值进行数据化之后得到的数据

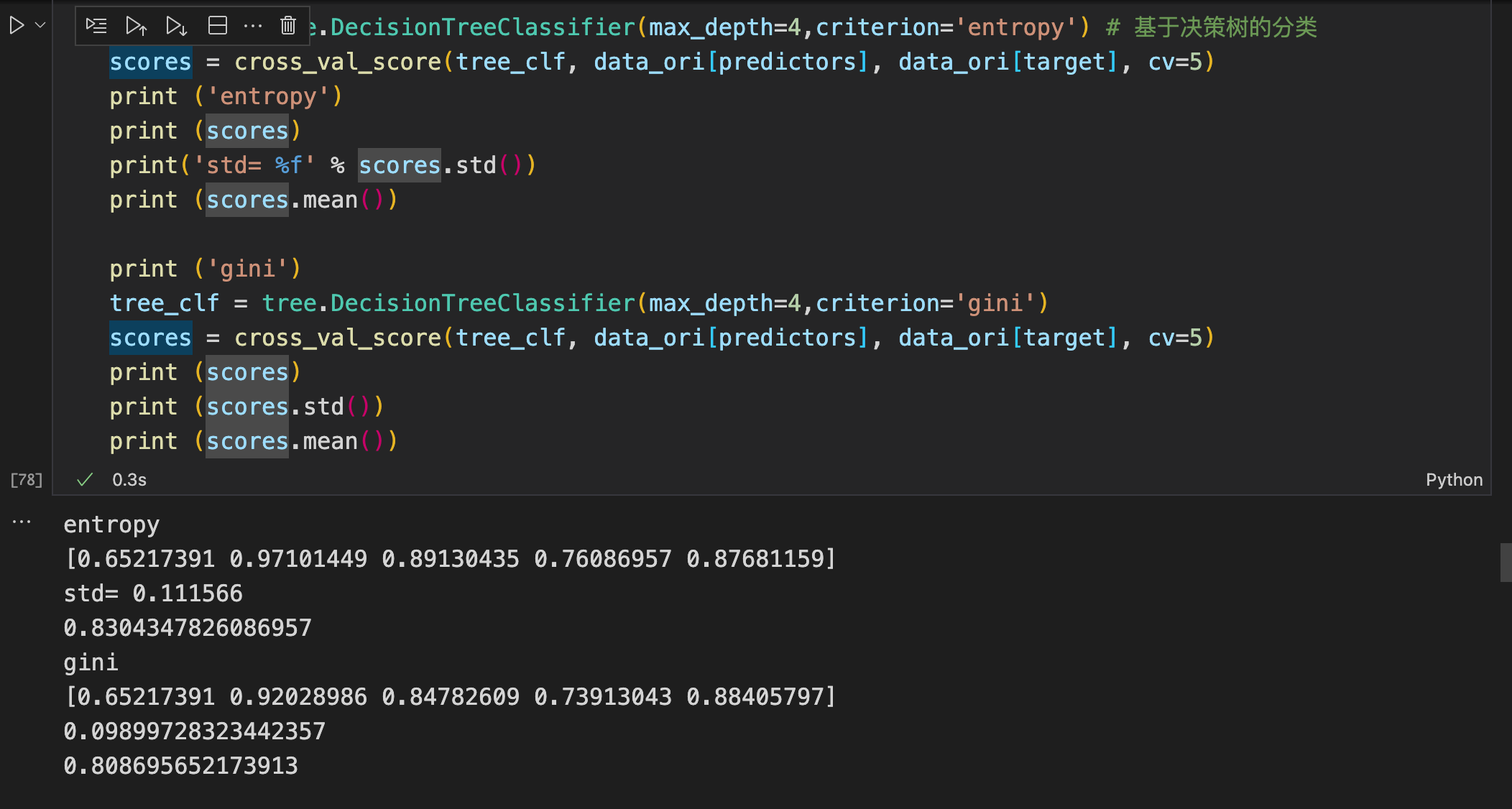


1. 选择合适的特征集合

暂时将其中除了target之外的所有数据都设置为predictors



1. 基于交叉验证方法，训练一个决策树模型，参数criterion分别选择gini和entropy，优化参数，比较分类准确率,可以发现采用entropy的准确率略高于gini

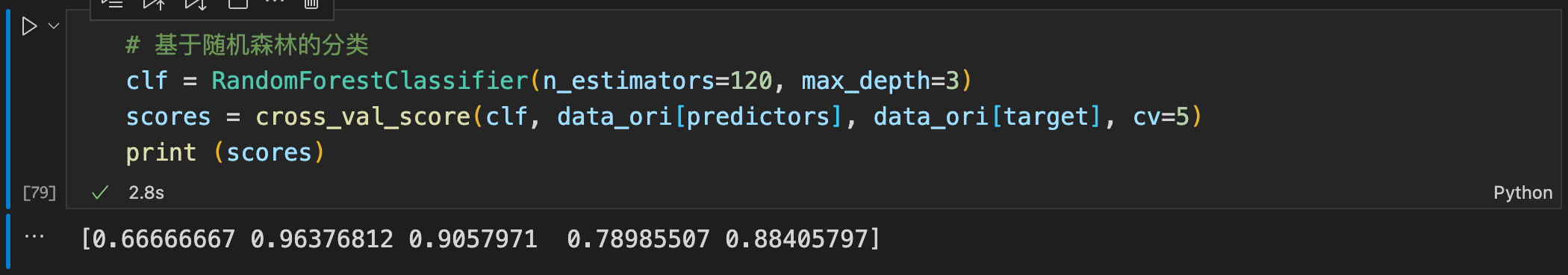


1. 基于交叉验证方法，分别训练Adaboost和 Random Forest模型，优化参数。建议，某个重要参数，例如n\_estimator，调参时，同时对比两个算法的性能，并绘制为图表，例如accuracy与基分类器数量的关系曲线图。

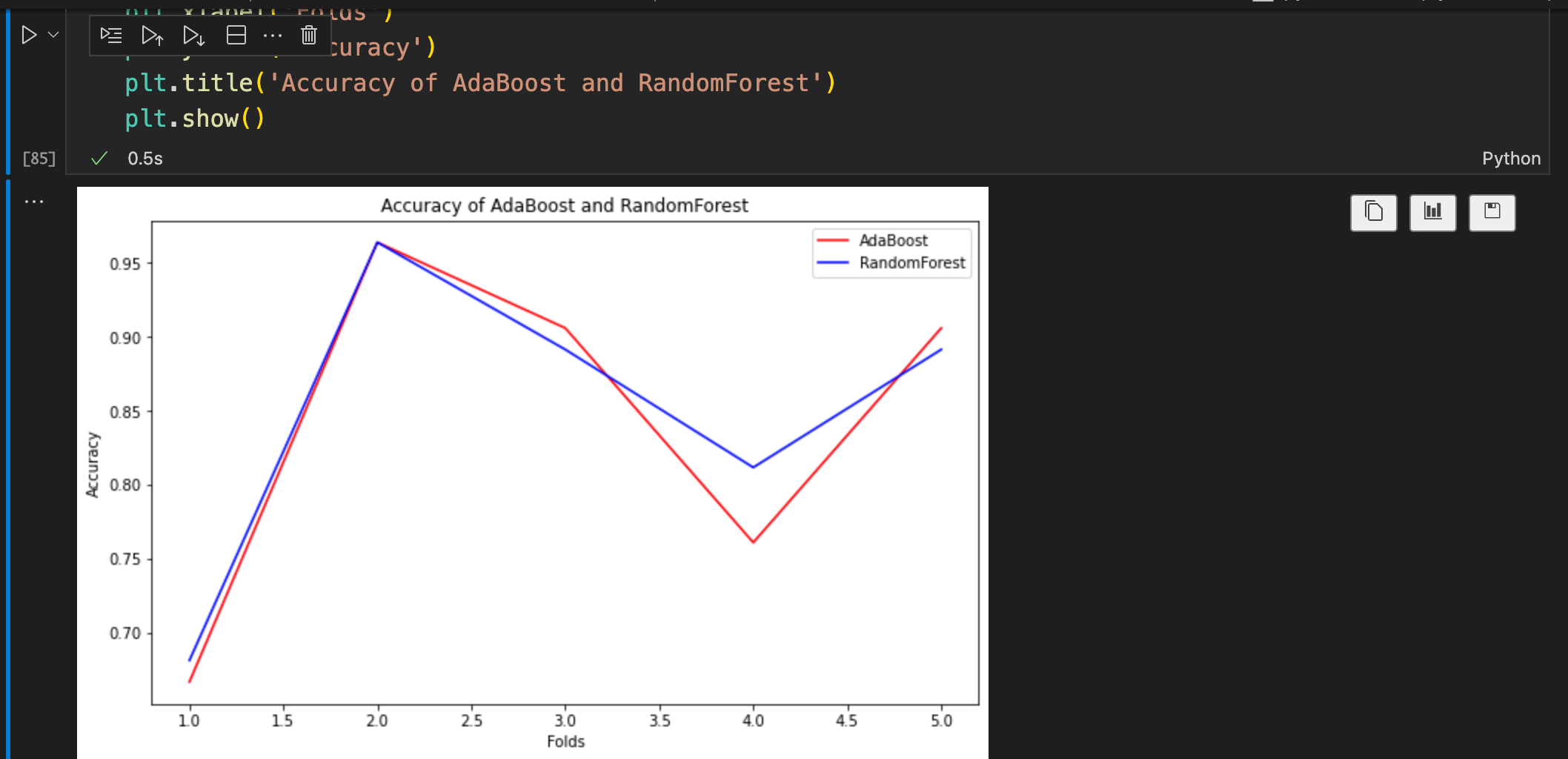
基于adaboost 进行分类得到的结果



基于随机森林分类得到的结果

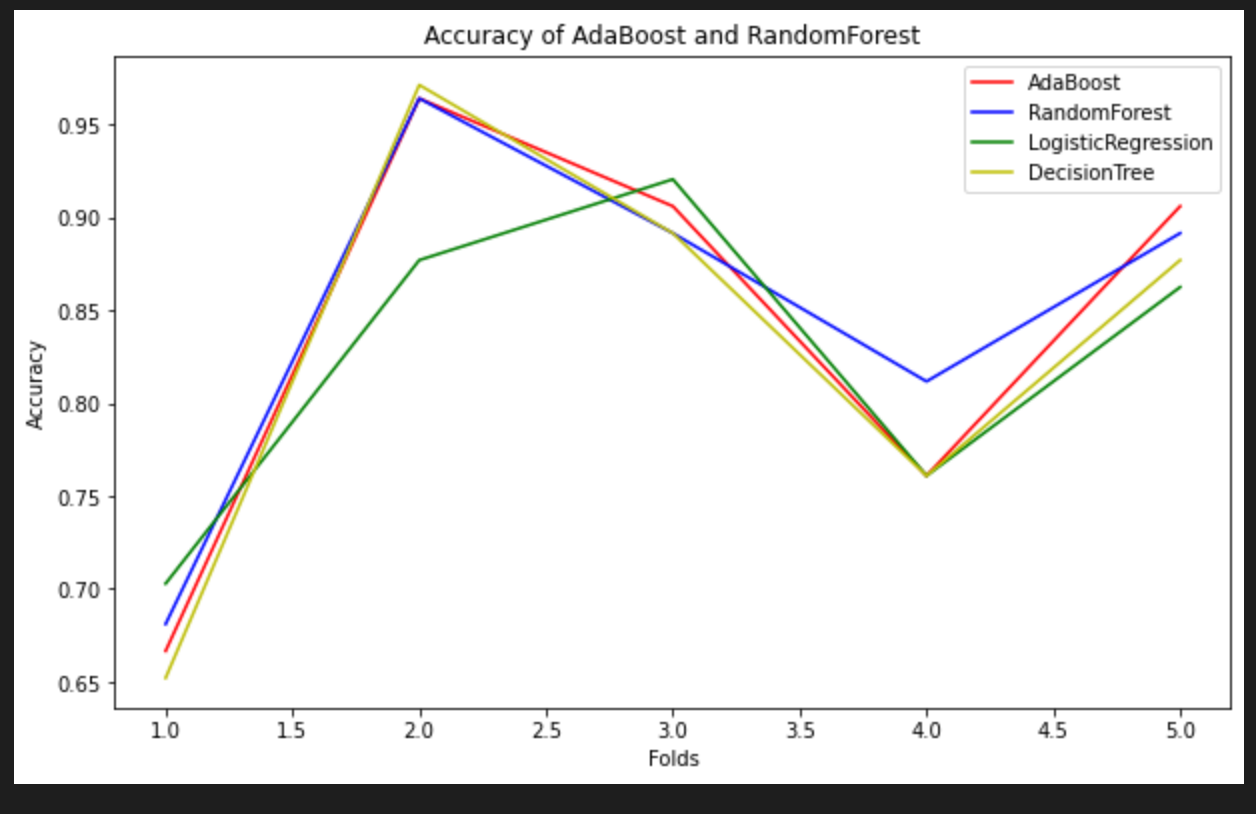


绘制两种模型的对比图之后可以得到



1. 综合比较多个模型的性能，尝试优化特征工程，改进各模型的效果

将上面用到的四种模型绘制于同一个对比图当中



（6）书写图文并茂的数据分析报告。性能比较，应该列一个表格。

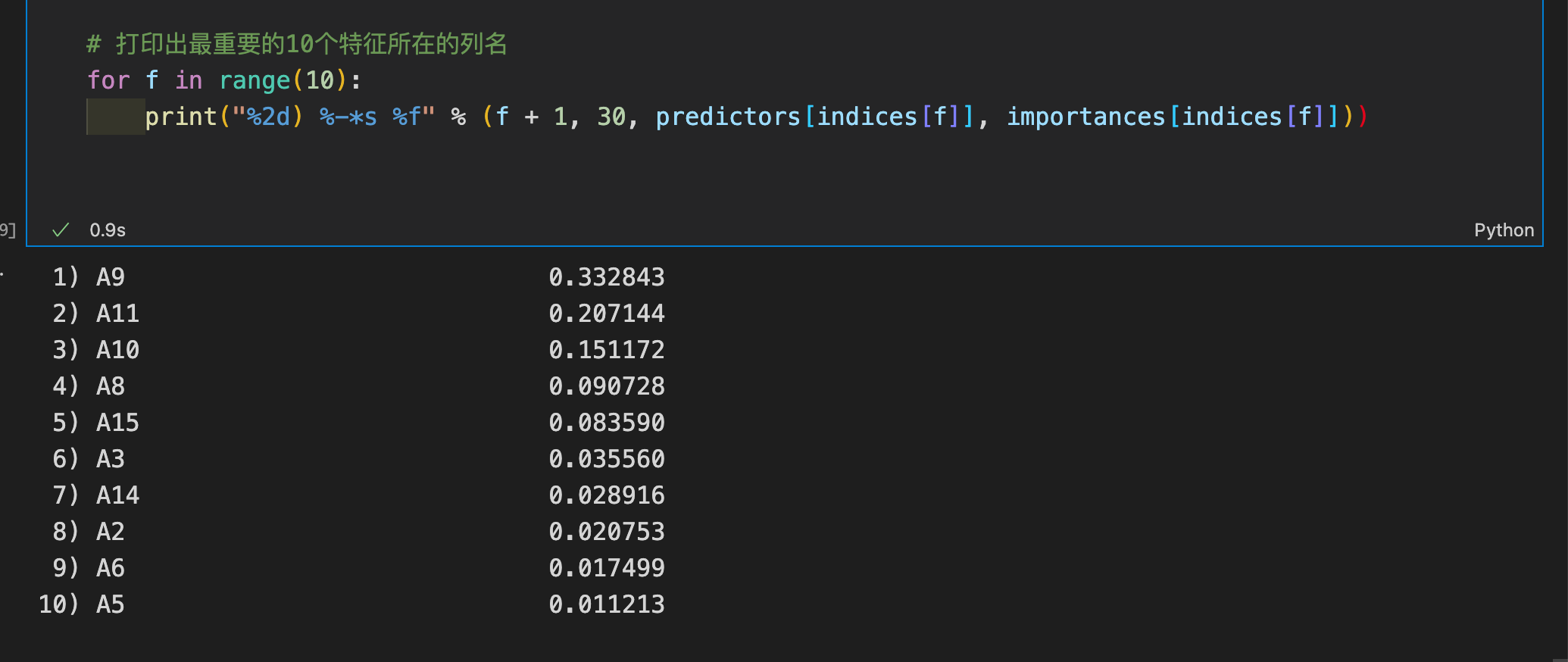
列表之后综合比较随机森林法拥有最高的准确率并且方差比较小，逻辑回归法准确率最低



4．分析与讨论

（1）实验数据中，如何评价各特征维度对分类结果的重要性？列出Top10重要 特征，给出必要 的说明。

利用sklearn库当中的feature\_importances\_函数给出对于各个特征的重要性进行评估之后可以得出最重要的10个特征如下



（2）本实验任务，降维方法是否可改善分类性能？

回答问题，记录个人知识总结、调试经验等

5．附录

附实现代码