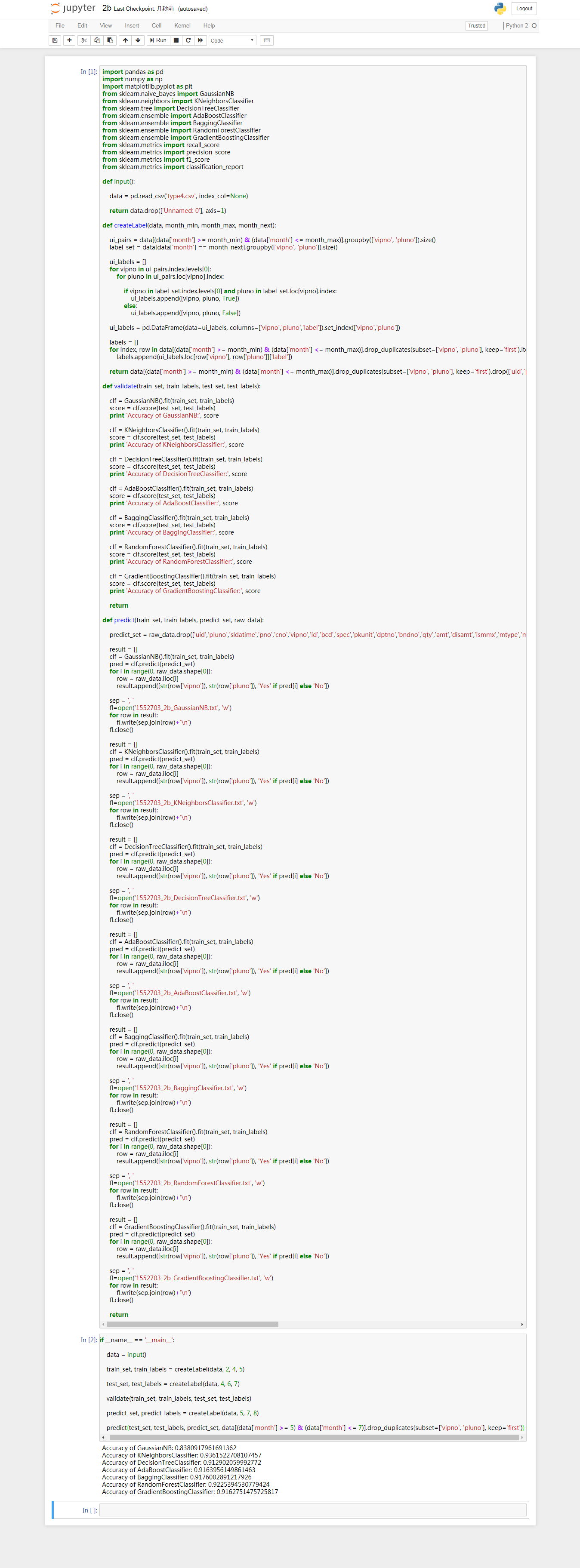
## 代码运行

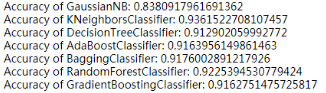
[点击查看原图](b.png)

运行所需type4.csv文件可在a的特征结果中找到



### 讨论分析

1. 本利用题目1a的分类器构建用户交易特征与用户购买商品IDs之间的多分类模型，测试和讨论该模型的准确性，并且预测8月份的购买情况，以[用户], [商品], [标签]的格式进行表示
2. 首先分析7个模型的特点：
   1. GaussianNB：借助贝叶斯概率公式直接计算“条件 - 结果”概率
   2. KNeighborsClassifier：使用KNN算法直接计算样本距离，获得一组数字上合理的近邻分类，和前者贝叶斯一起是暴力计算的方法，虽然缺少一定的逻辑支持但是具有不错的效果和性能。
   3. DecisionTreeClassifier：针对多个条件构造基于Y/N的树形分类，通过多重条件判断来确定最后的分类，相较于前两者具有更好的解释性
   4. AdaBoostClassifier：较为直接的集成方法，通过用同一组数据训练多个不同的分类器并进行集合来获得更好的预测结果
   5. BaggingClassifier：在AdaBoost的基础上增加了预测序列和投票机制，构造一个预测函数的系列，并按照投票的结果进行组合，以期获得最好的结果。比Adaboost相比是有选择的集成
   6. RandomForestClassifier：分裂多棵决策子树来复杂化决策过程，以在Y/N决策的基础上获得更为复杂的组合来预测更多的情况
   7. GradientBoostingClassifier：通过多次迭代并从中选优的过程，来从多个弱分类器获得一个较强的分类器。
3. 1a中出现的GradientBoostingClassifier极差状况在本题中并没有出现，分析是因为标签数较少，GradientBoostingClassifier能够正常进行迭代的结果
4. 实验过程中发现，如果使用全部的20k+条数据，进行预测，效果比仅使用一个数据训练一个标签的效果更差。实际预测准确率差距在10%左右。考虑到一个标签取一个数据的情况下，测试集的数据数量依然不足以导致足够大的偶然误差，认为这一结论并不是出自偶然因素，分析如下：
   1. 一个标签取一个数据的情况下，不同标签的数据差异更为明显，在之后的预测过程中就更容易分辨不同的数据，获得更好的结果。
   2. 用户在前三个月中的第一次购买行为和第四个月的第一次购买可能存在更为紧密的联系。
      1. 首先用户在三个月内第一次购买某项物品，必然是该用户在三个月中的某个月中的第一次购买该物品
      2. 因此之前的联系可以理解为，用户每个月第一次购买某样物品的记录和他之后的某个月是否会有购买该物品的第一次记录有着更为密切的联系
   3. 这样理解之后认为一个标签使用一个数据是一个合理的选项，故最后采用这种方法。
5. 各分类器的准确率如下：
6. 认为准确有效，其中准确率超过0.9的分类器认为可靠



### 性能比较

此处比较各个分类器训练并预测8月结果的耗时。