**dix平台（高校版）简易教程**

作者：自由在高处

时间：2017.7.27

写前语：由于我们队伍的队员在之前并不熟悉spark，这两天也是瞎弄，侥幸跑通了，若教程中有错误的地方，还请各位看官谅解。我们使用的是**pyspark+RandomForest**。

1. **特征提取**
2. 代码框架(fea\_extra\_\*.py)



1. 说明

训练集和测试集的特征提取都用这个框架就行了，训练集只是比测试集多一列【标签字段】，我是放在的最后。**需要特别注意的是，特征之间都需要用空格隔开，这是之后算法模型的输入要求的。**如果之后模型报错了，回来重新提特征吧，因为很可能是你的数据格式不规范导致的。

1. 平台运行
2. 将“输入”目录中的“初赛训练集(3k)”拖入画布
3. 将“组件”目录中的“PySpark”拖入画布，形成下方所示的图
4. 选中PySpark，配置参数



注意：上图中打马赛克的地方是训练集特征保存的文件路径，自己随便设置就行了，但是${output\_prefix}/ 这个前缀得保留，意思是在这个目录下，你随便建文件夹。



如上图所示，运行完成后，如果显示“成功“，训练集的特征就保存在你指定的目录下面了。



如上图所示，测试集和训练集一回事儿，按之前的代码框架改下脚本，改下输出路径，原始数据路径，就可以了，当然，测试集你得改下资源参数，这儿我也是乱设置的，仅供参考。



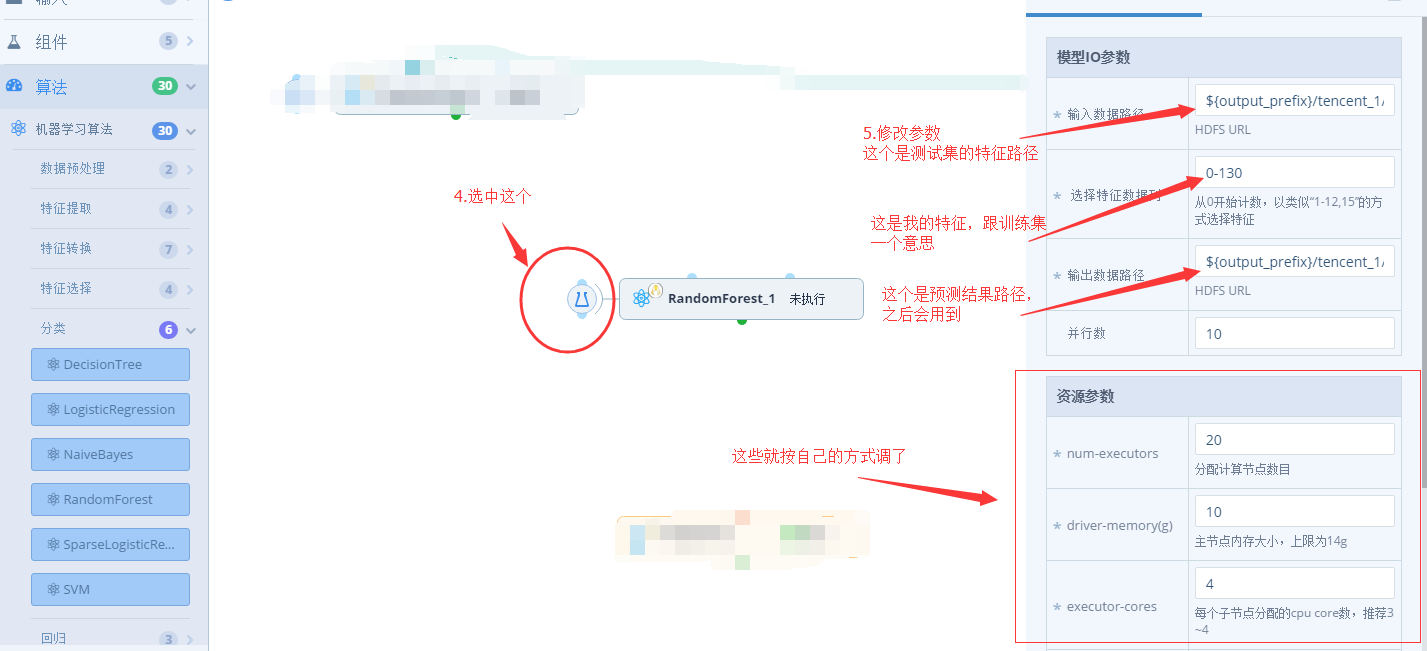
输出路径是可以自定义的，但是原始数据路径是严格的，查看方式见上图，将数据集拉入画布，选中，右方就是数据路径。

看到这儿，你的训练集和测试集的特征应该提完了，并保存好了。有两个东西我们之后会用到，**一个是训练集的特征保存路径，一个是测试集的特征保存路径。**

1. **模型的训练与预测**

请按照图中所示进行操作。





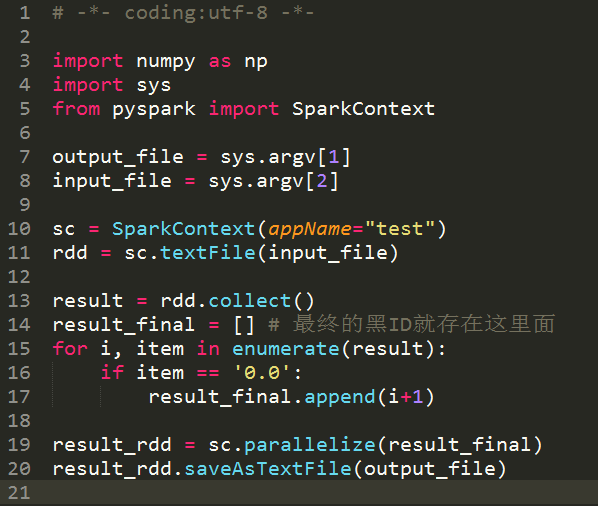


按照上述步骤操作，如果显示“成功“，就可以得到了模型的预测结果。该结果保存在你刚才设置的”预测结果路径“中（步骤5中设置的参数）。

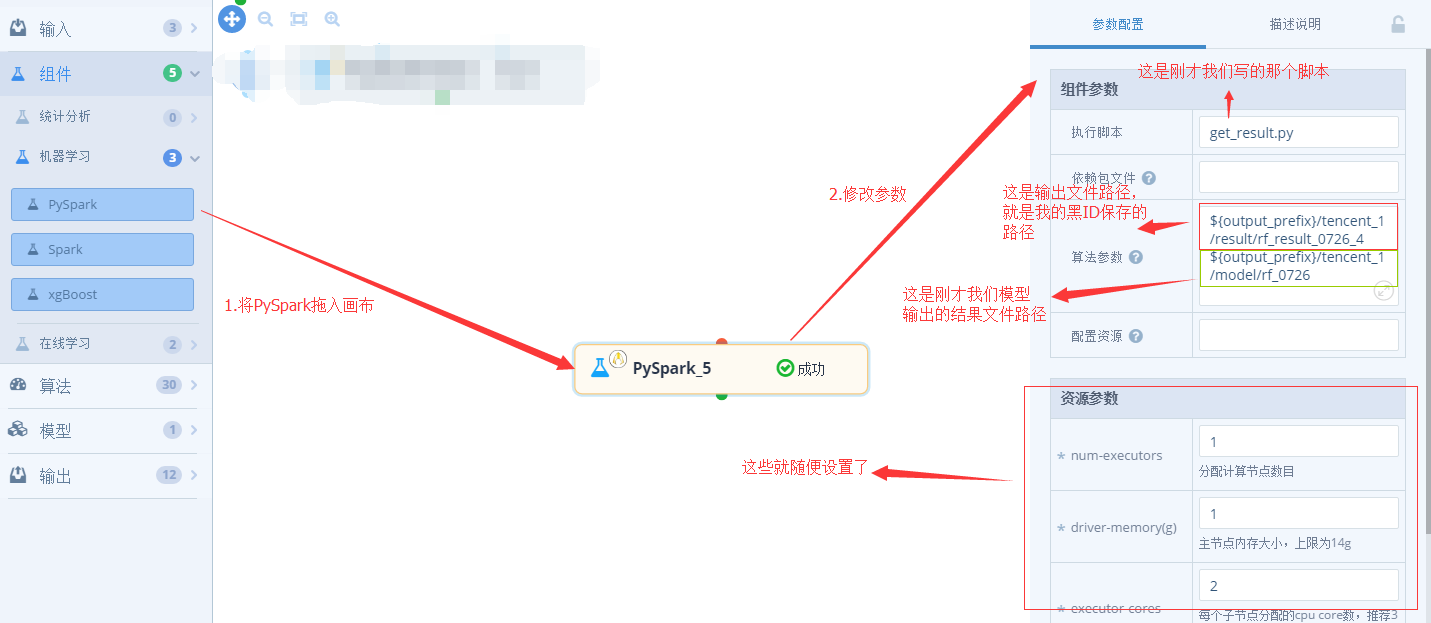
1. **根据模型预测结果提交成绩**

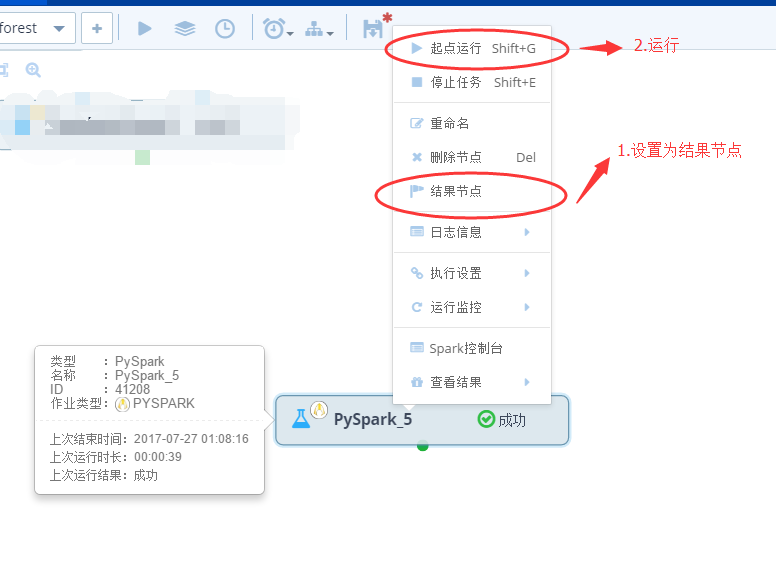
首先，你得看下**官方给的用户手册**，每一个模型的输出是不一样的，你得根据你模型的输出写脚本，如果你不知道你的结果文件长什么样，可以先collect然后print出来看看。我们的这个demo使用的是随机森林，直接上解析结果文件的代码。注意，在我们的demo中，是使用了全部的特征，所以结果文件就是一些“0.0“和”1.0“。

get\_result.py



下面对平台进行操作





当显示运行成功时，就得到最终我们可以提交的结果了。提交结果的方式请见官方提供的用户手册中的4.1节。

**写后语：**

由于表达能力有限，写得又有点儿急，如果有阐述不清楚的地方，还请结合着**官方给的用户手册和demo**看，现在的官方手册已经写的很详细了，其实我们也是按照官方给的demo改的代码。此外，建议大家在用平台前，先线下测试自己的代码是否能跑通，检查一下线下环境和平台环境是否一致。最后，祝大家取得好成绩。