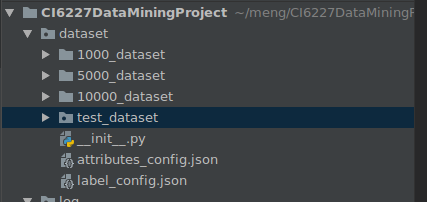
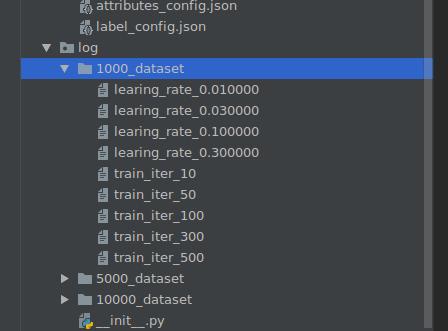
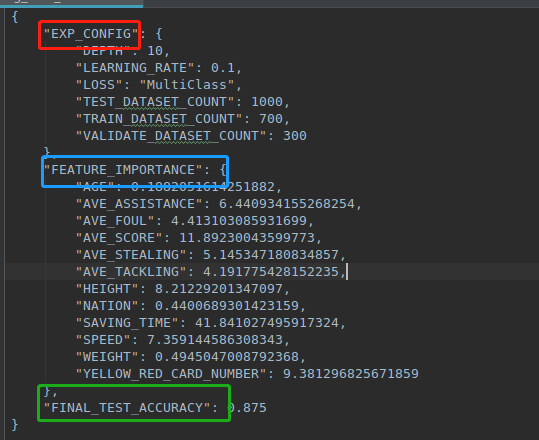
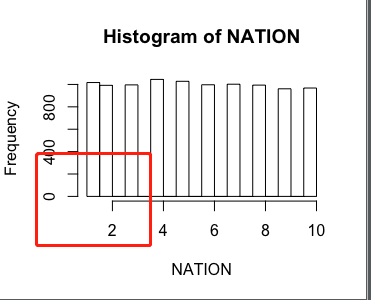
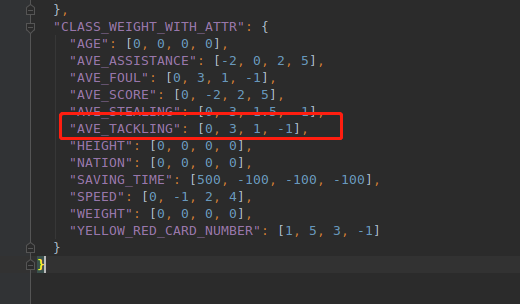
1. 简单的实验说明
   1. 生成的所有dataset在
   2. 1000 5000 10000分别是三个train set，表示了其sample的数量，但因为生成label的原因，其实际数据都稍微小于该值。
   3. 每个dataset下的label\_stat.json记录了各个label的数量，都是相比比较均匀的。Goal keeper的数量相对少。
   4. Test\_dataset是单独跑出来用于测试的数据集
2. Log文件
   1. 
   2. 针对每个数据集，测试了learning rate, trainer iteration次数和depth（是catboost的里的模型的某个参数，表示了gradient boosting里树的深度）的调参。所有的结果都在各自的文件里包括了：
   3. 
   4. 红色是该次实验参数配置
   5. 蓝色为feature importance
   6. 绿色为最后在测试数据集上的准确率
   7. 跑实验室在tune 某一个参数时，其他参数保持默认值不变。
3. 需要画的图：
   1. 数据集上 取10000\_dataset，各个属性值的分布：
      1. 之前的问题：所有histogram为什么左边下表缺失了
      2. 对于连续值的画图，应该是均匀分布的，但是画出来是正太分布，画的是不是CFD图啊？
   2. 对于实验结果：
      1. 对三个调的参数和对应的test acc画折线图，比如learning rate，横轴为learning rate的值，纵轴为test acc，对应三个数据集就有三条折线，画在一张图上即可，标明每条线的label
      2. 对dataset 数量对test acc的影响，取log每个dataset下的： depth\_4, learing\_rate\_0.030000和train\_iter\_300的对应的log，画三条折线图，横坐标为dataset的数据量，纵坐标为test acc。三个文件画出三条折线，放到一个图里即可，标明label
   3. 对feature importance
      1. 取5000\_dataset/learing\_rate\_0.010000的log
      2. 画散点图，横坐标为某个feature的权重的模，就是找到dataset/5000\_dataset/label\_config.json 对应每个feature的长度为4的权重数组：
      3. 取这四个值的平方和开根号作为散点的横坐标，纵坐标为对应的feature importance值。应该是线性正相关的。