方向方法

主阶段-1

- 主要就是快速复现别人的代码，尽快刷成绩。对于赛题和数据以及模型没有深刻的理解。排名尚可，但都是别人玩过的了。

主阶段-2

计划 & 分析

- 问题：
  - xgboost的0.786的模型是否还有改进空间?
  - 特征数量过少吧?
  - GBDT对于过多的特征貌似影响不大，宁多误少吧?
- 情况：
  - 现在top选手：
    - 第一名开源部分代码，只有特征提取的java代码，而且是2017-10-30最后一次提交
    - 第二名和第十一名交流后使用了模型融合；改进第十一名的代码，苟进前二
    - 第六名下面的使用了多模型训练，提交不知道单模型还是多模型
    - 第十一名开源代码，代码问题多；特征过多106个，模型过多6个；模型融合。直接复现代价太大
    - 根据陈浩阳测试0.81的特征结果：gbdt和xgb分别一个0.7954，一个0.80180
  - 过去情况：
    - top1的auc只有0.78228，且两次的数据和场景基本相似，现在我的auc已经超过这个，所以过去的方案未必有参考价值了
    - 第一名 GBDT给的权重大概是XGBoost的二倍
- 想法：
  - 样本：
    - 线上线下都用
  - 特征：
    - 宁多误少的原则 尝试
  - 模型：
    - GBDT 单模型效果大部分选手说较好
    - XGBoost 加强调参
    - 模型融合 & 模型一定要训练的互补
    - xgboost、lightgbm、catboost的auc和权重不好说；看论文大概就是catboost调好了最好但有局限性，lightgbm和xgboost相近。看错了xgb的auc极限并不是0.79左右
- 步骤：
  1. 从头做起，代码自己手撸（最快大概可能需要5天）
  - 认识数据
  - 数据可视化
  - 数据预处理
  - 特征提取 & 特征要多
  - 尝试多模型，效果预计依次递减：

        - GBDT
        - xgboost
        - lightgbm
        - catboost
        - RandomForestClassifier
        - ExtraTreesClassifier
    - 集成学习：还是就6个模型==》融合
    - 网格搜索