拍拍贷风控比赛方案

团队： *shu_brothers*

队长： 段石石
成员： 李君浩
叶成

SHU_BROTHERS 队伍介绍

队长：段石石 *1*号店精准化算法工程师

个人技术博客：*http://hacker.duanshishi.com*

队员：李君浩 上海大学通信与信息工程学院硕士

队员：叶成　　同济大学硕士

特征基本处理一

clean_data.py：

1，category变量除了UserInfo_2,UserInfo_4直接做factorize，因为在tree类的model，不需要做dummies处理；

2，UserInfo_2,UserInfo_4是城市信息，不用factorize处理，取两列城市并集，然后做映射；

3，从baidu上拉出城市的经纬度信息，这样可以找出对应城市的经纬度信息，能够解决一些在经纬度上相关联的数据问题；

4，增加字段UserInfo_2_4_01，为0表示UserInfo_2与UserInfo_4相等，反之，为1；

5，使用从city_ratio.py生产个UserInfo_2的target为0的数量以及占比和UserInfo_4中target为0的数量与占比；

特征基本处理二

create_features.py 增加log和user update数据：
1,登录的次数、频率、时间区间
2,用户更改信息的次数
3,增加用户修改信息如修改qq或者是否有车，则在对应位置置1，增加约55维二值变量

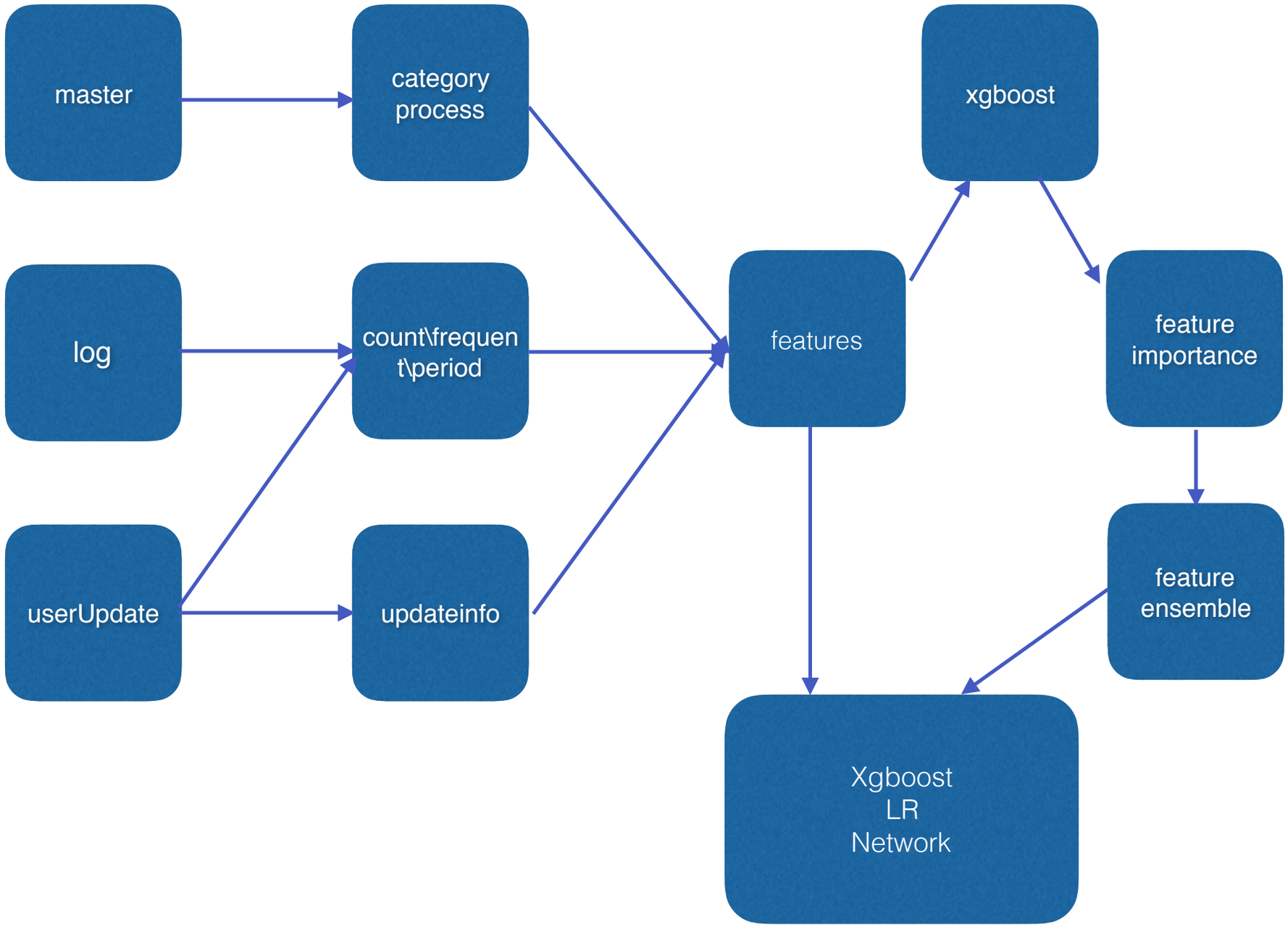
模型选择

1，lr_model.py:Logistic Regression模型来预测，效果不好，放弃；
2，network.py: 用keras封装的神经网络模型，机器太差调参跑不了，放弃;
3，xgboostClassifier.py: 将xgboost封装为sklearn pipeline支持的类，方便调参,且时间成本相对nn较低，效果在初赛也比较好，故选择Xgboost作为model。

代码基本解释：

1，xgb_model.py： 查看对应的cv分数，初步判断num_boost为eta的一些初步取值范围；
2，xgb_feature_selection.py： 通过xgboost的feature importance观察那些feature的重要性更高，然后对那一类特征做基本的特征组合处理（feature_ensemble.py）；
3，xgb1_20_2.py RandomSearch 寻找xgboost最优参数
4，xgb1_20_2_pos_weight.py 与xgb1_20_2.py效果一样，考虑不平衡样本数量，多了scale_pos_weight，但是发现效果一般
5，feature_ensemble.py 从xgb_feature_selection.py中选择一批importance比较高的特征值，然后做组合特征计算。

模型训练基本流程


master
category process
xgboost
log
count\frequent\period
features
feature importance
userUpdate
updateinfo
feature ensemble
Xgboost
LR
Network

总结

1，特征工程耗费时间太多，机器不给力，很多方案没有有效验证；
2，模型调参应该先根据xgb.cv进行粗调，盲目调参太费时间；
3，特征选择策略对最终结果影响很大，这里耗费时间太多太多，单机每次尝试太费时间；
4，使用多模型做ensemble处理能够有效防止单模型的带来的随机性问题，一般都能提高AUC
5，以为是24号完成数据提交即可，前面只是一直在测试数据方案，最终数据提交没有等数据cv跑出来后做ensemble，人工定的几个结果做的ensemble，且没有考虑local CV的score来做ensemble的weight。