A．项目描述
为 home credit 公司提供综合评估客户还款能力的解决方案，为了评估那些之前没有过借款或信用记录的潜在客户的还款能力，home credit 公司使用包括电信记录、交易记录在内的大量信息来预测客户的还款能力。

2.数据描述
提供了 7 个 csv 文件的数据
application_train/test.csv：主表，分为训练集和测试集，客户的主要个人信息
bureau.csv：客户在信用咨询公司留存的所有在其他机构贷款记录
bureau_balance.csv：上述贷款记录的月度余额情况
POS_CASH_balance.csv：客户在 home credit 贷款的现金和 pos 机月度消费情况
credit_card_balance.csv：客户在 home credit 持有信用卡月度余额情况
previous_application.csv：客户在 home credit 之前的贷款申请记录
installments_payments.csv：客户在 home credit 之前的贷款还款记录

3.评估标准
使用 ROC 曲线下的面积（AUC）进行评估。

B．模型实现
1.数据处理
将每个样本进行信息整合与连接（preprocess），以 SK_ID_CURR 为主码
a.对性别、是否有车等 bool 型数据，改为 0/1
b.对多值数据用 get_dummies 转化为 one hot 编码
c.对连续值数据添加了诸如人均收入＝总收入 / 家庭成员一类的自定义数据

2.学习算法
lightgbm

3.调整超参数
准备用贝叶斯优化算法，但是我的电脑被它跑死了 emm 所以放弃了

C.结果
在整个训练集上 AUC 0.748

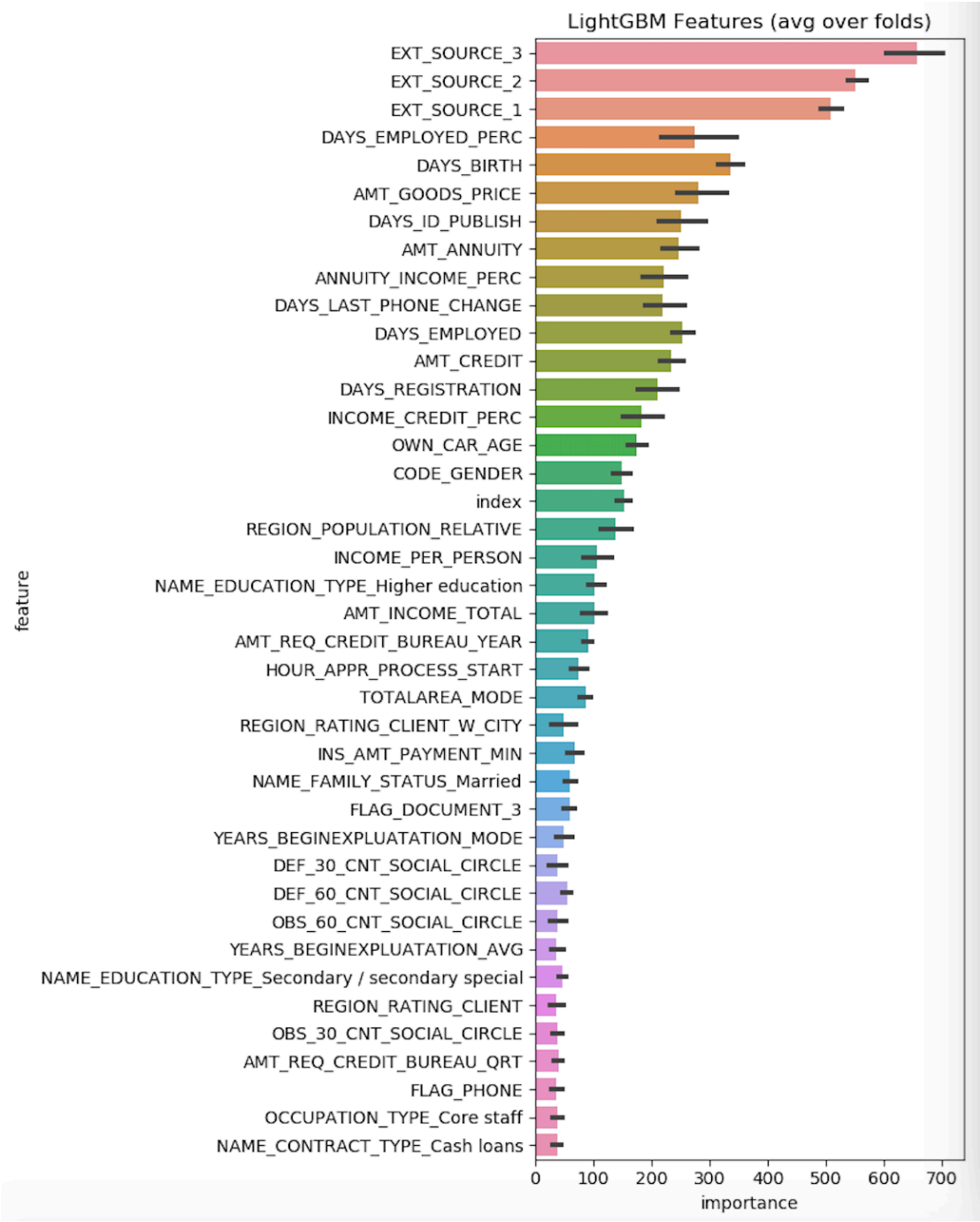
```
Full AUC score 0.747656
1.py:304: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  test_df['TARGET'] = sub_preds
Run LightGBM with kfold - done in 116s
Full model run - done in 123s
```

kaggle 分数评判 0.742

Submission and Description	Public Score
submission_kernel01.csv a minute ago by Mingjun Liu add submission details	0.742

feature importance



D.可能存在的问题和改进

1.参数调整，贝叶斯优化算法是可行的（但是我实在不想做了。。
2.可视化上还可以改进，很遗憾没有画出 ROC 曲线，因为需要对代码结构进行大幅度更改。。。是我开始没考虑到这个，但是总的来说可视化不是关键。
3.结果应该可以更好，可能应该到达 0.78-0.8 之间，大概是因为我没有调参数，再加上自定义数据还有可以改进的地方