**目的**

探索哪些人行衍生变量对预测申请授信客户的逾期风险较为重要

**数据描述**

1. 原收集人行数据有10779条记录，每条记录由83个字段组成。据松生反馈里面有很多查询失败的数据，过滤掉这些数据后还剩5473条记录。
2. 原收集标签（逾期>30天）数据中1306条，阴性数据1059条，阳性数据247条。去重还剩813条，没有出现同一个用户有两个label的情况。
3. 人行数据和标签数据的交集共有565条，即有565个放款的客户有对应人行数据，因此用于分析的数据大小为565\*84（83个特征+1个标签）。

**特征工程**

1. “人行学历”用NONE取代空值

2. “人行婚姻”用NONE取代空值

3. “人行职业”用NONE取代空值

4. “人行贷记卡当前还款率>9的账户数占比”的意思是当月还款与最小应还之比>9的账户数与客户持有贷记卡总数的比例，特征的表示形式为”n/m”，将其转换为小数，并把分子抽提出来作为新的特征

5. 以下这些特征需要做这个变换来获取真实数值：取整((原始数值-987) \* 100)\*10：

人行贷记卡单月最高逾期总额, 人行当前贷记卡已用额度，人行贷记卡近6个月平均使用额度，人行贷记卡授信总额，人行贷记卡最大额度，人行贷记卡最小额度，人行贷款单月最高逾期总额，人行贷款近6个月平均应还款额，人行车贷月供，人行房贷本月应还，人行推测贷款月应还款额，人行公积金缴存基数，人行贷款当前逾期金额

6. 因以下这些特征存在重尾现象，对其做对数转换：

人行贷记卡单月最高逾期总额, 人行当前贷记卡已用额度, 人行贷记卡近6个月平均使用额度, 人行贷记卡授信总额, 人行贷记卡最大额度, 人行贷记卡最小额度, 人行贷款单月最高逾期总额, 人行贷款近6个月平均应还款额, 人行贷记卡当前逾期金额, 人行贷记卡本月应还, 人行贷记卡本月实还, 人行贷记卡隐含分期, 人行贷记卡最低还款, 人行车贷月供, 人行房贷本月应还, 人行推测贷款月应还款额

7. 以下特征的值全为0，做丢弃处理

人行资产处置笔数, 人行是否有欠税记录, 人行是否有强制执行记录, 人行是否有行政处罚记录, 人行未结清房贷笔数

8. 人行贷记卡综合缴款评等（2年）/人行贷记卡综合缴款评等（5年）/人行贷款综合缴款评等（2年）/人行贷款综合缴款评等（5年）/近2年贷记卡逾期月份数/近5年贷款逾期月份数/近5年贷记卡逾期月份数/近2年贷款逾期月份数，这些字段原来的值是这种模式的：/1//////7///////////////，表示近两年或五年每个月的还款情况（每个字符一个月），我将其转换为近2年或5年来的逾期月份占比和逾期月份数

9. 以下字段没有意义，丢弃处理：'id', 'customer\_unique\_key', 'serial\_no', 'create\_time', 'edit\_time', 'serial\_no.1'

特征工程后的数据大小是567\*79

**建模**

使用CatBoostClassifier建模，并输出特征的重要性。因为模型受随机性影响，输出的特征重要性也不一样，因此训练模型10次，取每次输出的特征重要性平均值作为特征的重要性评估指标。

模型参数：depth=3, loss\_function='Logloss', eval\_metric='AUC', colsample\_bylevel=0.5, early  
\_stopping=10

模型表现：[0.7223404255, 0.6723404255, 0.7042553191, 0.6957446809

, 0.7659574468, 0.6808510638, 0.7106382979, 0.7, 0.7446808511, 0.6957446809],

平均值: 0.71