机构 单量(单库) 机构名 贷后 类型 走势

11 2216240 你我贷 千万 int12,3 线上

15 434683 泰丰租赁 没有

3089 213870 买呗 12w pdl 已经下线

13 186165 没有

3410 153038 同信云盾 20w pdl 线上

1. 138923 51贷 36w pdl 已经下线

1 125394 小木鱼

50 100895 掌富贷

3066 73138 信而富

2890 69605 杭州闪银

4440 63326 金蛋科技

3490 45290 资赫网络

1000 35916 奋青邦

金盘201903月有贷后的pdl订单为评估对象 标签pd10 数据量6.7w

1 地址映射特征

train\_rate: 0.31550826559503153 test\_rate: 0.31827562485668426

训练上的auc：0.992203 训练上的KS：0.911816

验证集上的auc：0.729246 测试上的KS：0.364928

测试集上的auc：0.712250 跨时间上的KS：0.326322

columns feature\_importances

11 addr\_avg\_apply\_area 0.034977

10 addr\_avg\_apply\_country 0.034352

23 addr\_avg\_loanamount\_area 0.034173

22 addr\_avg\_loanamount\_country 0.033352

21 addr\_avg\_loanamount\_city 0.029418

9 addr\_avg\_apply\_city 0.028638

15 addr\_avg\_approve\_area 0.028326

2 地址LDA特征

Topic=50 最优

训练上的auc：0.970964 训练上的KS：0.794075

验证集上的auc：0.642000 测试上的KS：0.210972

测试集上的auc：0.639619 跨时间上的KS：0.200637

Pd10 = 30% 贷后表现很差，可能原因：乙方推送部分贷后数据有误，

但是在这样的数据上这些特征表现很好，之前在秒啦首贷数据上测试地址LDA特征

基本无效，究其原因，秒啦首贷pd10的占比大概是6-7%左右，本身已经做的很好，在这个基础上地址特征效用不高，但是从整体PDL上来看 地址LDA特征有效

---------------------------------------------------------------------------------------------------------------------------------

同信云盾数据 201905上半月 8w pd10 32% pd0 34%

此机构地址数据比较规整，详细

地址LDA数据

Topic=50

train\_rate: 0.32430696090713435 test\_rate: 0.3280997956769545

训练上的auc：0.950790 训练上的KS：0.754096

验证集上的auc：0.780555 测试上的KS：0.434766

测试集上的auc：0.795400 跨时间上的KS：0.458559

训练上的auc：0.987143 训练上的KS：0.873071

验证集上的auc：0.779963 测试上的KS：0.451536

测试集上的auc：0.783270 跨时间上的KS：0.448378

Tppic=65 通过困惑度确定的最佳topic

训练上的auc：0.991782 训练上的KS：0.894888

验证集上的auc：0.798141 测试上的KS：0.475973

测试集上的auc：0.795351 跨时间上的KS：0.474398

Topic=80

训练上的auc：0.987266 训练上的KS：0.870780

验证集上的auc：0.768173 测试上的KS：0.434567

测试集上的auc：0.779730 跨时间上的KS：0.439844

Topic=110

训练上的auc：0.961176 训练上的KS：0.757080

验证集上的auc：0.754120 测试上的KS：0.382872

测试集上的auc：0.747014 跨时间上的KS：0.372104

地址历史映射特征 topic=50

train\_rate: 0.3261861167325174 test\_rate: 0.3218087966448005

训练上的auc：0.953914 训练上的KS：0.745680

验证集上的auc：0.812028 测试上的KS：0.495452

测试集上的auc：0.806906 跨时间上的KS：0.477302

columns feature\_importances

11 addr\_avg\_apply\_area 0.048461

23 addr\_avg\_loanamount\_area 0.046355

15 addr\_avg\_approve\_area 0.039683

10 addr\_avg\_apply\_country 0.038365

22 addr\_avg\_loanamount\_country 0.038103

7 addr\_overdue\_rate\_area 0.036456

14 addr\_avg\_approve\_country 0.031504

6 addr\_overdue\_rate\_country 0.029251

19 addr\_avg\_overdue\_area 0.028575

排名靠前的特征全是和area,country相关的特征，说明粒度越细效果越好

最后将lda 和映射特征结合

train\_rate: 0.3261861167325174 test\_rate: 0.3218087966448005

训练上的auc：0.995569 训练上的KS：0.928347

验证集上的auc：0.850875 测试上的KS：0.574703

测试集上的auc：0.846126 跨时间上的KS：0.567284

同信云盾数据 201905全月 17w pd10 31% pd0 33%

Topic=50 lightgbm 测试集取45%

train\_rate: 0.31334198394362334 test\_rate: 0.3118956136053018

训练上的auc：0.964247 训练上的KS：0.775687

验证集上的auc：0.765961 测试上的KS：0.401788

测试集上的auc：0.767938 跨时间上的KS：0.399261

利用17w做词袋 用8w数据训练的LDA 50做子模型

训练上的auc：0.965738 训练上的KS：0.778200

验证集上的auc：0.779753 测试上的KS：0.420376

测试集上的auc：0.781213 跨时间上的KS：0.420834

说明8w数据训练的模型更好

Topic=60 通过困惑度确定的最佳topic

训练上的auc：0.960668 训练上的KS：0.764451

验证集上的auc：0.774349 测试上的KS：0.410512

测试集上的auc：0.770681 跨时间上的KS：0.406283

利用17w做词袋 用8w数据训练的LDA 60做子模型

训练上的auc：0.966751 训练上的KS：0.783146

验证集上的auc：0.784923 测试上的KS：0.425633

测试集上的auc：0.781321 跨时间上的KS：0.425419

Topic=65 利用17w做词袋 用8w数据训练的LDA 65做子模型

训练上的auc：0.968963 训练上的KS：0.789380

验证集上的auc：0.784894 测试上的KS：0.431282

测试集上的auc：0.783002 跨时间上的KS：0.434701

Topic=80

训练上的auc：0.966974 训练上的KS：0.780310

验证集上的auc：0.751145 测试上的KS：0.381343

测试集上的auc：0.757691 跨时间上的KS：0.389102