第二届融360“天机”金融风控大数据竞赛

CADV - DU - DA

目录

综述

- 本次大赛以用户个人资料，消费信息，关系信息和标签信息为基础，预测用户是否会在未来进行二次贷款。
- 大赛以AUC为评测指标，评测的用户数目为12261名，带标签用户26000名。
- 团队以两个方向入手
  - 数据驱动模型
  - 业务驱动模型
- 两个方向的模型融合，最终AUC排名第一！

比赛实时排行榜

初赛结束时间：

2016.10.8 24:00

您最近一次预测得分为0.6224（2016-10-08），历史最高预测得分为0.6224，目前排名第1位

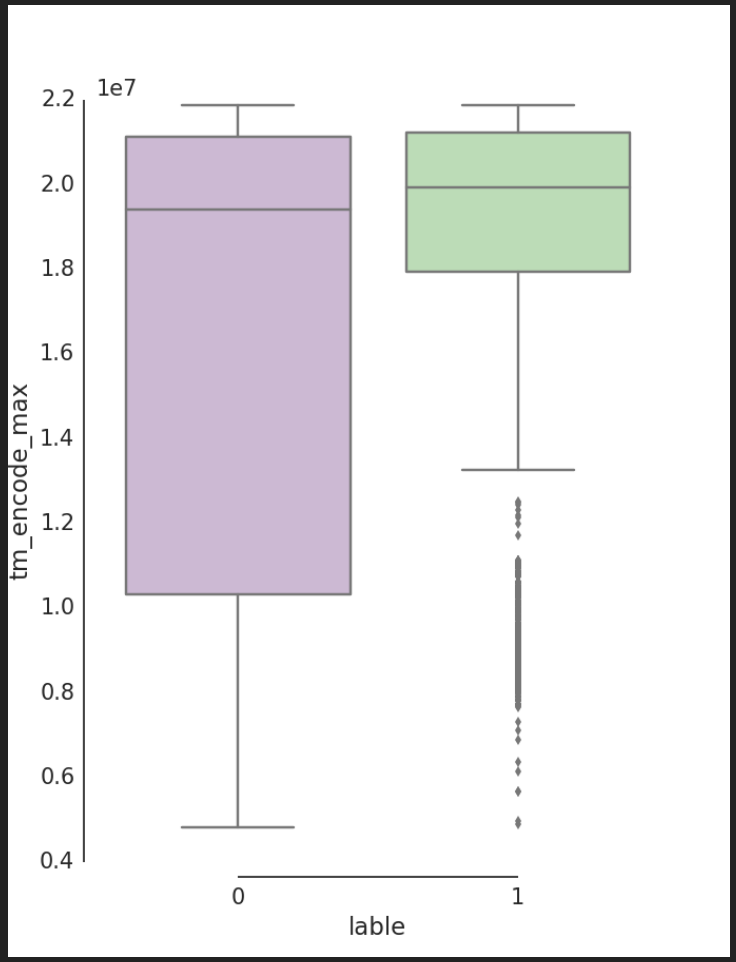
排名	预测得分	参赛队	赛区	所在学校
1	0.6224	CADV – Du – Da	广州赛区	UWA
1	0.6224	青岛黄渤	上海赛区	青岛大学
3	0.6212	微博搞笑排行榜	广州赛区	中山大学
4	0.6209	华南tfboys	广州赛区	中山大学
5	0.6195	Fengari-Kele-Pot	北京赛区	北京大学

用户信息统计

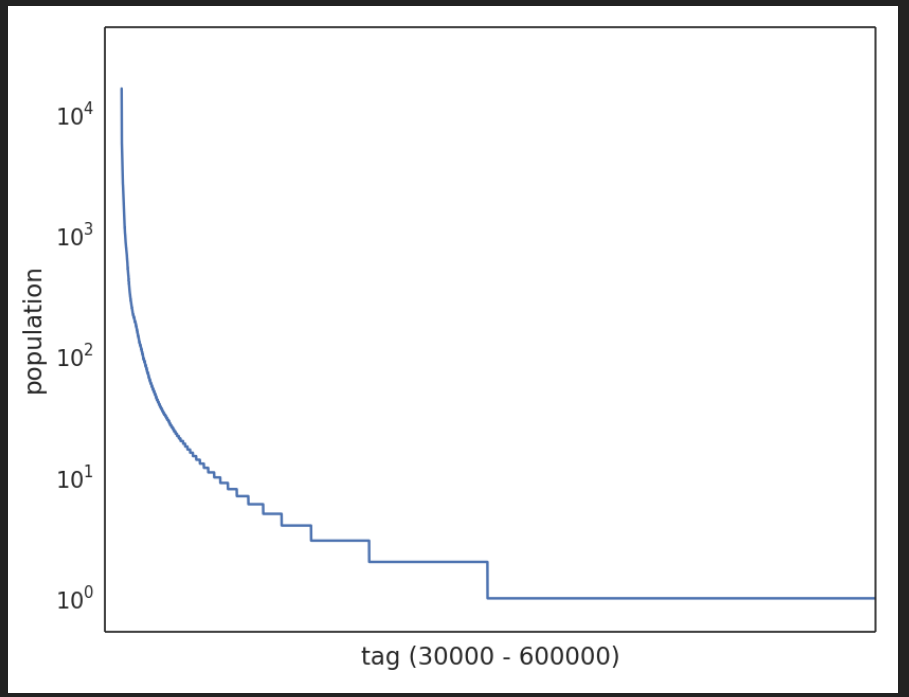
38621名用户部分信息概述

	mean	std	min	25%	50%	75%	max
age	28.224	5.483	0	24	27	31	98
business_type	0.444	2.171	0	0	0	0	16
education	3.325	0.720	1	3	3	4	4
expect_quota	51762.428	5338536.859	0	5000	10000	20000	999999999
occupation	2.646	0.985	1	2	2	4	5
pay_type	0.506	1.279	0	0	0	0	5
product_id	1.240	0.427	1	1	1	1	2

用户个别重要信息描述



记录时间分布



37359个标签拥有用户数量

用户信息缺失描述

	训练用户信息缺失比例	预测用户信息缺失比例	缺失用户数 训练 / 预测
用户特征信息	~ 0%	~ 0%	100%
消费信息	38.7%	41.9%	90%
标签信息	52%	64%	80%
关系1信息	12.0%	28.7%	40%
关系2信息	71.2%	76.1%	90%
关系集合信息	9.5%	26.7%	40%
消费信息+标签信息	68.9%	77.3%	90%
标签信息 + 关系	53.1%	66.1%	80%
消费信息 + 关系	44.0%	56.6%	80%
消费信息+标签信息 + 关系	69.6%	78.5%	90%

建模思路

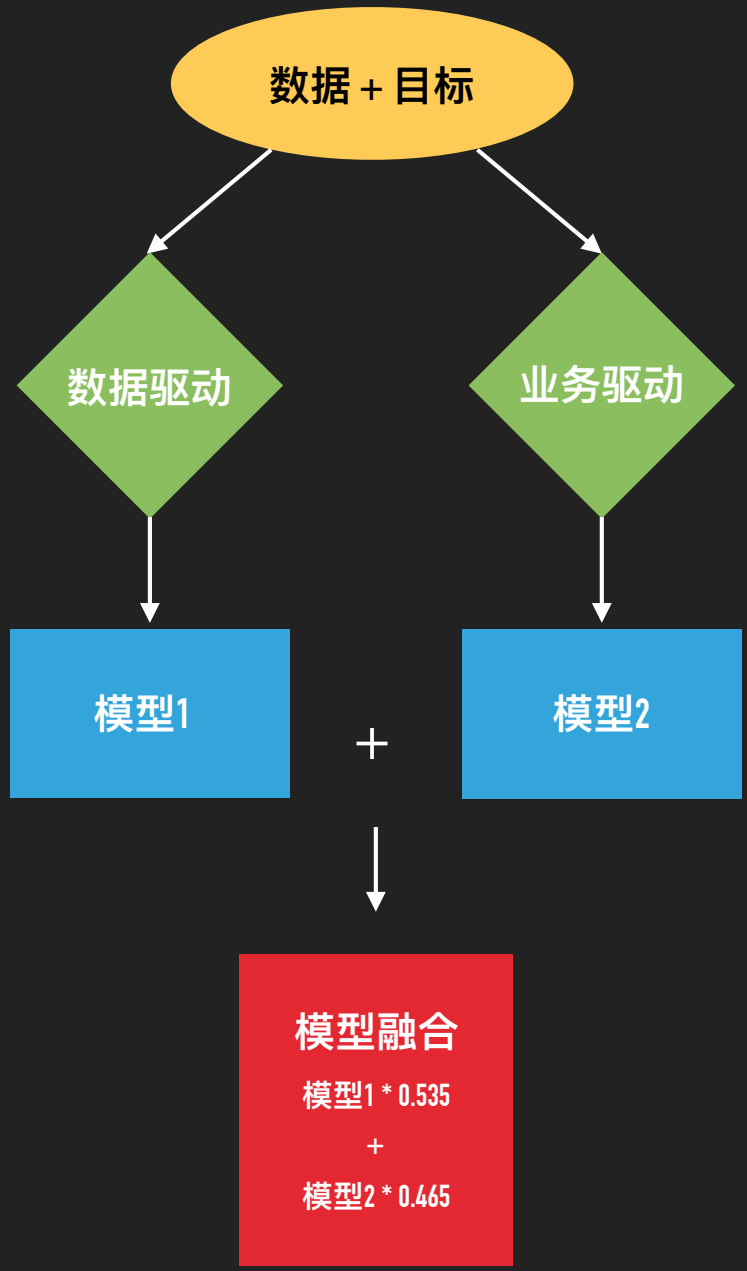
建模思路（1/2）

- 数据挖掘比赛套路！
  - 统计分析
  - 特征工程
  - 单模型构建
  - 模型融合
    - 模型间差别越大且各自效果好，融合效果最佳
- 奥卡姆的小剃刀
  - 模型效果相当时，优先选择**复杂度较低**

考思向逆

建模思路（2/2）

- 不同的思考模式 - 保证模型差异性大
- 简单的模型 - 保证模型解释度高
- 简单的融合方法
- 最好的效果！



模型解析

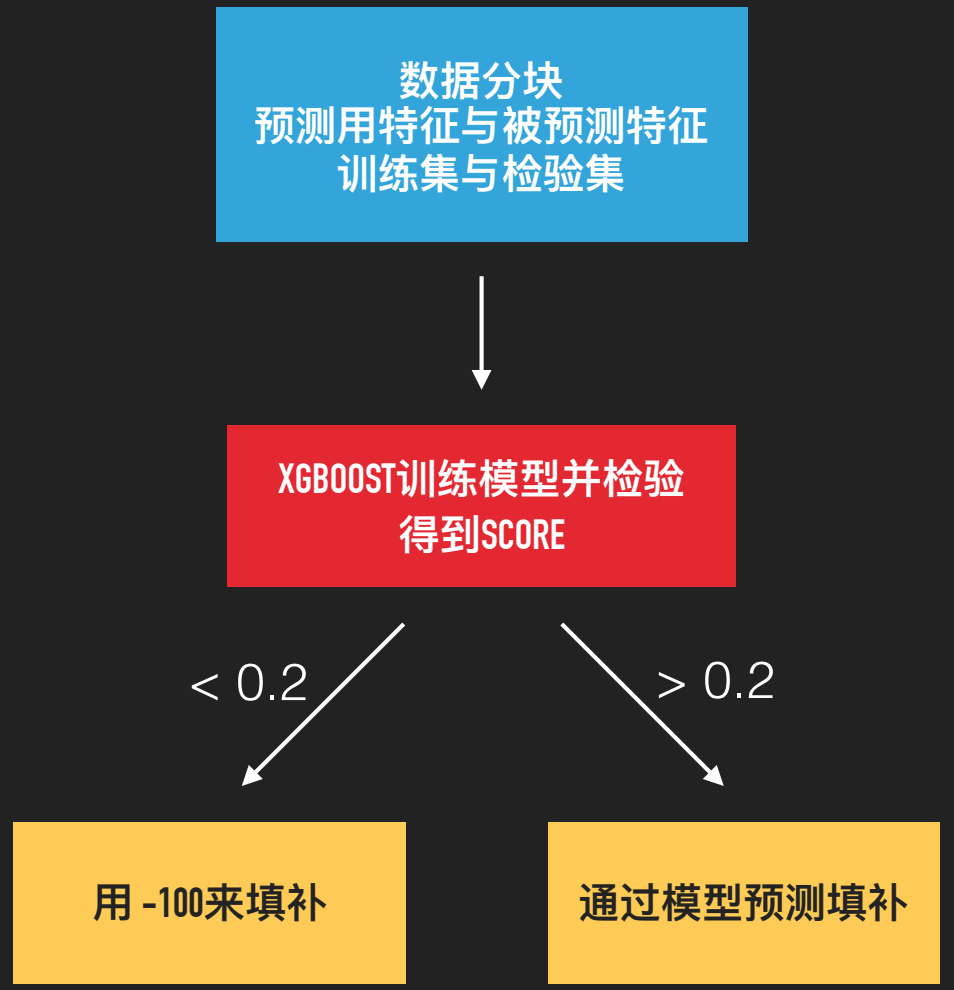
模型解析（1/10）

- 数据驱动模型（xgboost算法 - 39维特征 - rank：5）
  - 缺失值处理
  - 2级特征构建
  - 特征组合筛选（*）
- 业务理解模型（random forest算法 - 17维特征 - rank：11）
  - 数据预处理
  - 特征工程（*）
  - 模型构建
- 模型融合

简单的模型，复杂的构建过程

缺失值处理

- 缺失值为新分支
- 用非缺失值预测缺失值

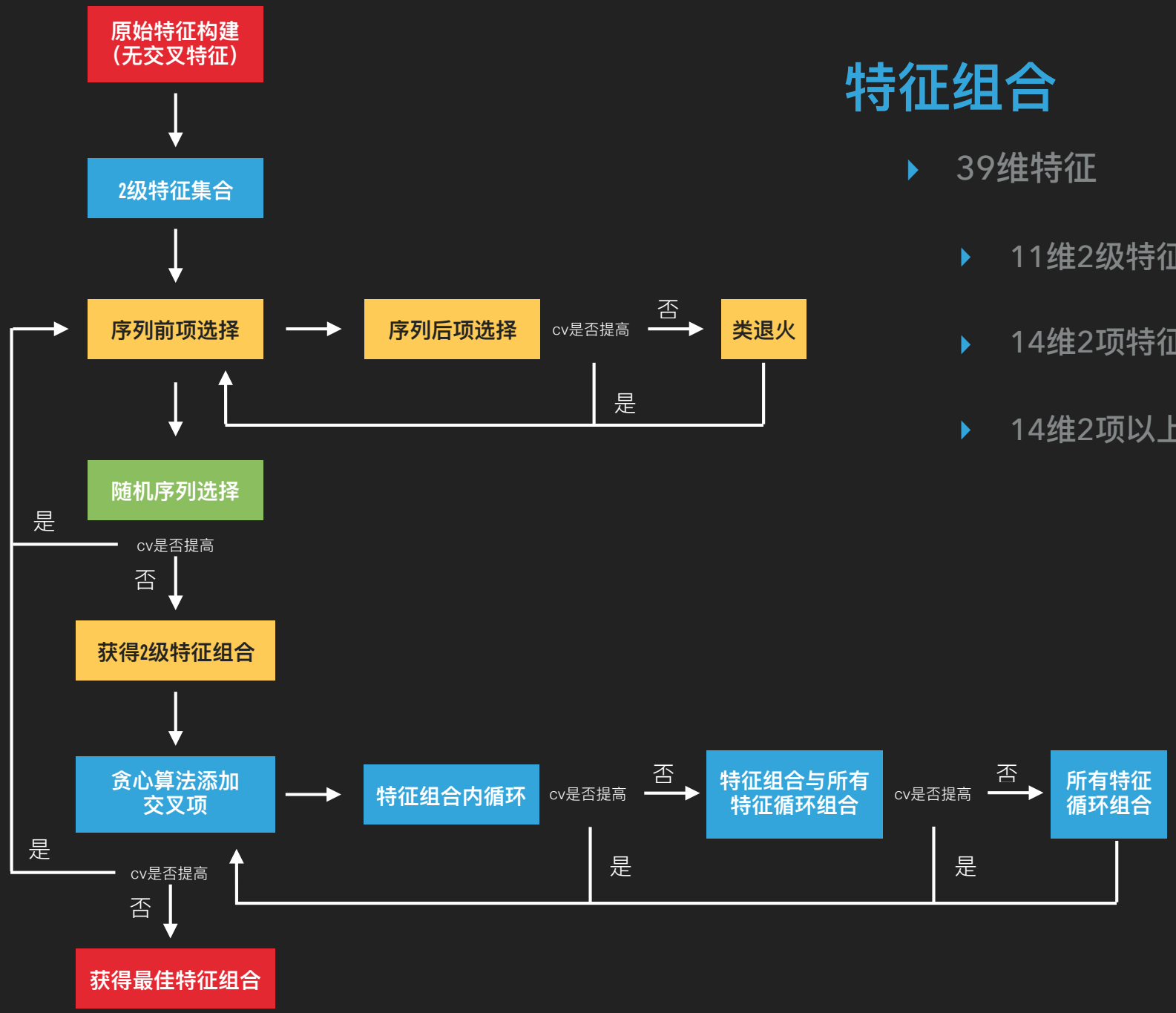


2级特征构建

原始特征拓展

	原始维度	2级特征维度	特征概括
个人特征信息	21	148（21 * 7 + 1）	*_min, *_max,..., info_num
消费信息	26	131（26 * 5 + 1）	*_min, *_max,..., num
标签信息	37359	103（95 + 8）	‘326446’, ..., ‘tag_num’, ‘tag_9’,...
关系信息	1 + 4	7（1 + 3 + 3）	‘relation1_num’, ‘nolabel’, ‘relation2_type’, ...
总2级特征维度			389

交叉特征的构建



特征组合

- 39维特征
  - 11维2级特征
  - 14维2项特征交叉
  - 14维2项以上特征交叉

数据预处理

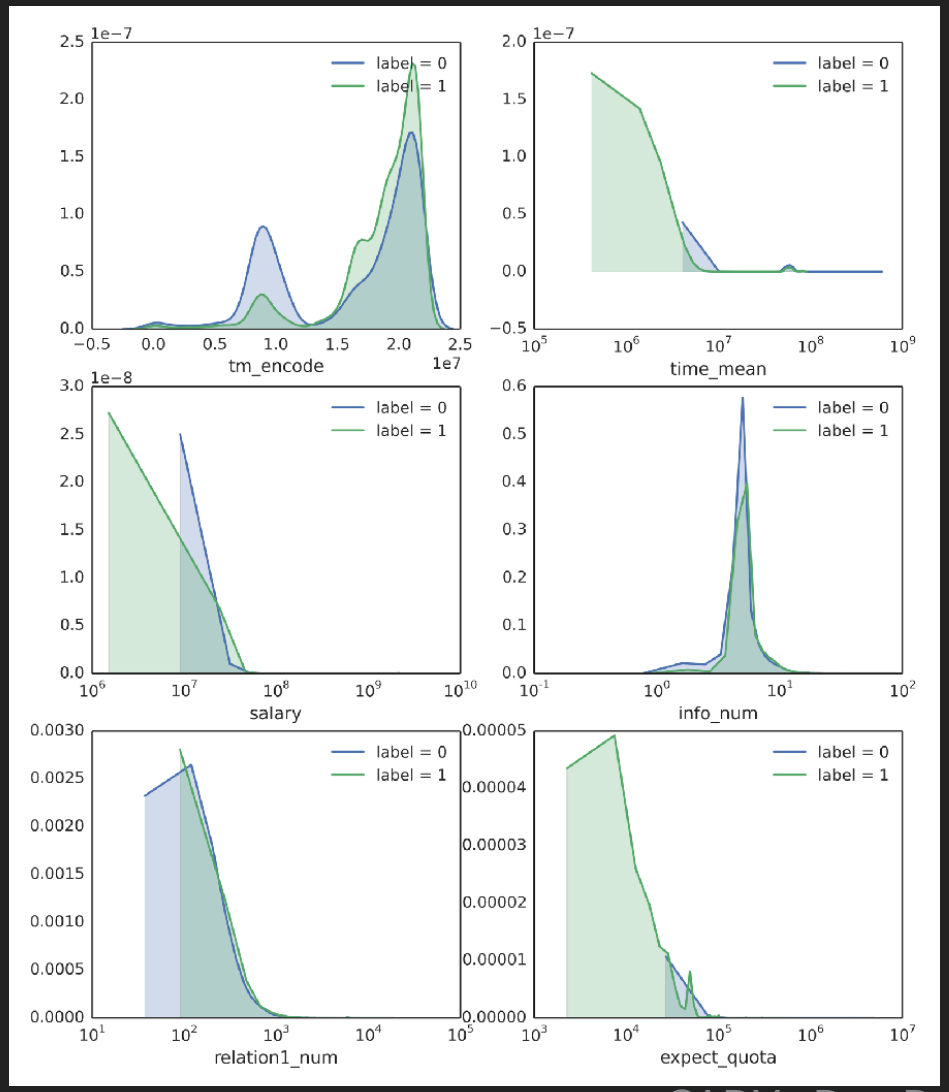
- 用户基本信息表
  - 考虑多数人多次填写表格时**填写正确次数多**，分组后用**众数**来填充。
- relation1表
  - 统计每个user1_id的连接数，并用连接数的**中位数**填充缺失值。
- 消费信息表
  - 0值可以被看做缺失值，统计0的在表中每个变量中的占比，将占比高于30%的变量舍去。将剩余变量的中位数/最大值/最小值计算出作为输入。
- rong_tag / relation2表
  - **缺失比极高**，故将两表舍弃不用
- 异常值检验，剔除

特征工程

- 逐表加入
  - + 用户基本信息表（11个特征，0.5000 - 0.5900）
  - + 用户关系表1　　（1个特征，0.5900 - 0.6070）
  - + 用户消费信息表 （5个特征，0.6070 - 0.6180）
- 特征描述

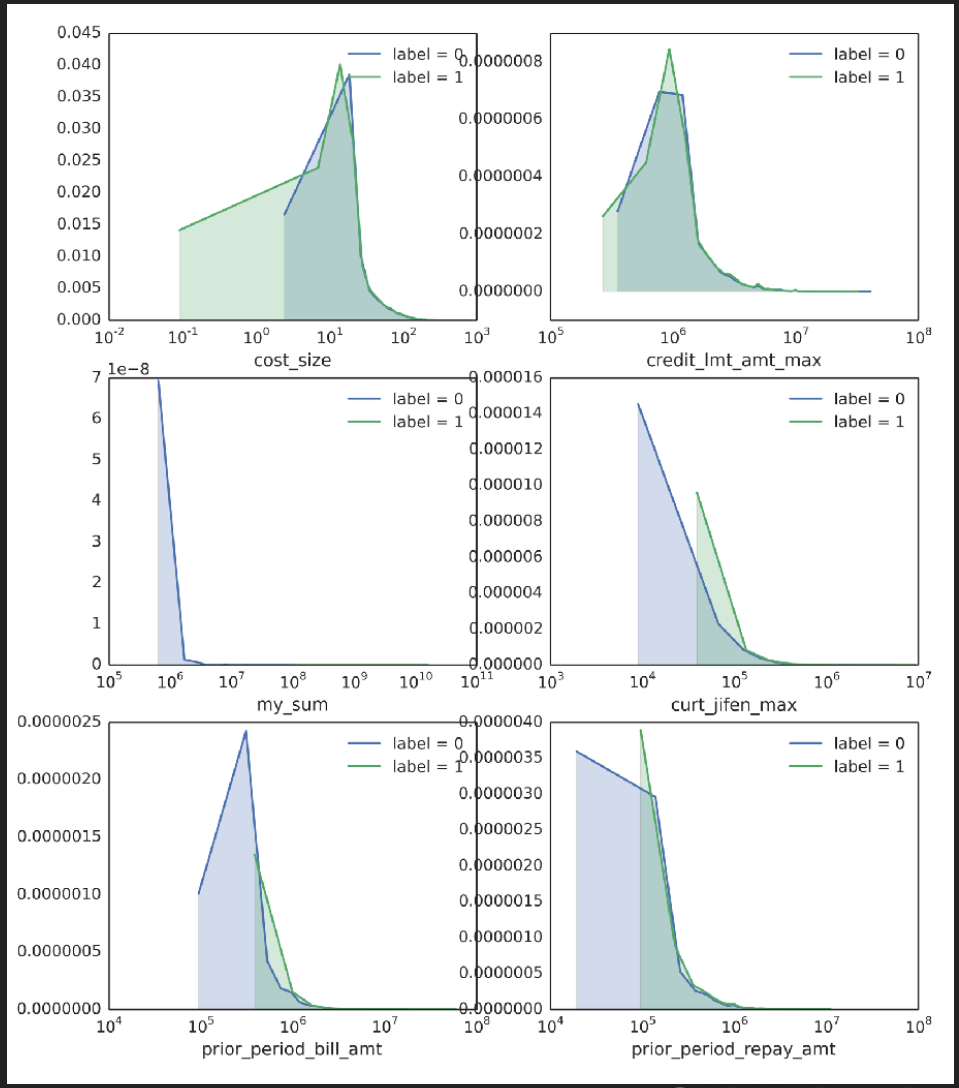
特征描述

特征	描述	使用原因
tm_encode	用户记录时间	- 对于有意向二次贷款的用户，由于个人信息是之前初次贷款的信息，因此在进行二次贷款前必然会对个人信息进行更新，因此更新时间离当前较近的用户有较大可能进行二次贷款。从附录的图A可以很清晰地反映出来。
time_mean	记录时间均值	- 记录时间的均值与记录时间配合，可以了解用户记录时间的离散情况，若离散程度大，则说明多次更新资料的时间间隔大，一定程度反映说明用户二次借贷的意向。
salary	薪水	- 从右图可以看到，是否进行二次贷款的用户薪水分布有较大的差异，为了满足消费需求，薪水较少的用户更倾向于进行二次贷款，这也符合我们日常的理解。
info_num	系统记录条数	- 用户信息更新次数的多少一定程度反应用户对自己资料的考虑与谨慎程度。在需要二次借贷的意向下，如何最大可能地获得贷款是用户需要思考的，而思考的过程必然会让自己的信息不断更新，使资料越来越好，因此记录条数越多，越反映用户二次借贷的倾向。
relation1_num	单个用户连接数	- 从右图中可以看到，二次借贷者的用户连接数普遍比无二次借贷的用户要多，因为在用户考虑二次借贷时，个人信用是需要考虑的重要因素之一，而有越多的联系人，即信用担保人，用户二次借贷的成功率会更高，因此有二次借贷倾向的用户会考虑增加自己的联系人为自己的信用加分。
expect_quota	申请金额	- 从右图可以明显看出，有二次借贷的用户借贷金额分布明显比无二次借贷的金额要低得多。这个原因很可能是因为用户初次贷款没有考虑到自己实际需要的借贷数目，导致需要二次借贷，故采纳该变量进模型中。



特征描述

特征	描述	使用原因
cost_size	可消费数量	- 消费次数越多，一定程度上反映了需要的金钱也更多，因此多二次贷款的需求也可能更加迫切。
credit_lmt_amt_max	信用卡最大额度	- 从右图可以明显看出，有二次贷款的用户信用卡最大额度的分布要比无二次贷款的往下走。若信用卡额度较大，则用户在消费时需要贷款的可能性就减小，因此该变量能反映用户二次贷款的趋势。
my_sum	迫切指数	- 该变量同样体现了用户对于二次贷款的迫切程度。
curt_jifen_max	最大当前积分	- 从右图可以明显看出，当前最大积分的分布差异较为明显，有二次贷款的活跃用户明显积分更多。反过来积分多的用户是贷款的活跃用户。
prior_period_bill_amt	上期账单金额	- 从右图可以明显看出，有二次贷款的用户账单金额明显更高。因此该变量能反映用户二次贷款的倾向。
prior_period_repay_amt	上期还款金额	- 从右图可以明显看出，有二次贷款的用户还款金额明显更高。因此该变量能反映用户二次贷款的倾向。
occupation	职业类型	
marital_status	婚姻状况	- 此信息虽然没有在附录图B中没有显示与二次贷款有明显关系，但由于这是个人信息中与资产状况有明显挂钩的变量，最明显即房屋的信息。另外户口类型与婚姻状况也是反映用钱的方式改变。该系列变量的方差较大，缺失少，线上CV也有效果，因此选择作为本模型的特征。
education	教育状况	
live_info	房屋类型	
local_hk	户口类型	



交叉检验设计

- 正负样本比例约为11:9，为保证交叉检验中正负样本比例相似，采用分层抽样来划分数据集，这里采用15-交叉检验。

简单模型融合

- 使用采用加权融合的方式 (Gini+Entropy)，具体公式为：

$$Model2 = Gini * 0.51 + Entropy * 0.49$$

业务理解模型输出

加权平均

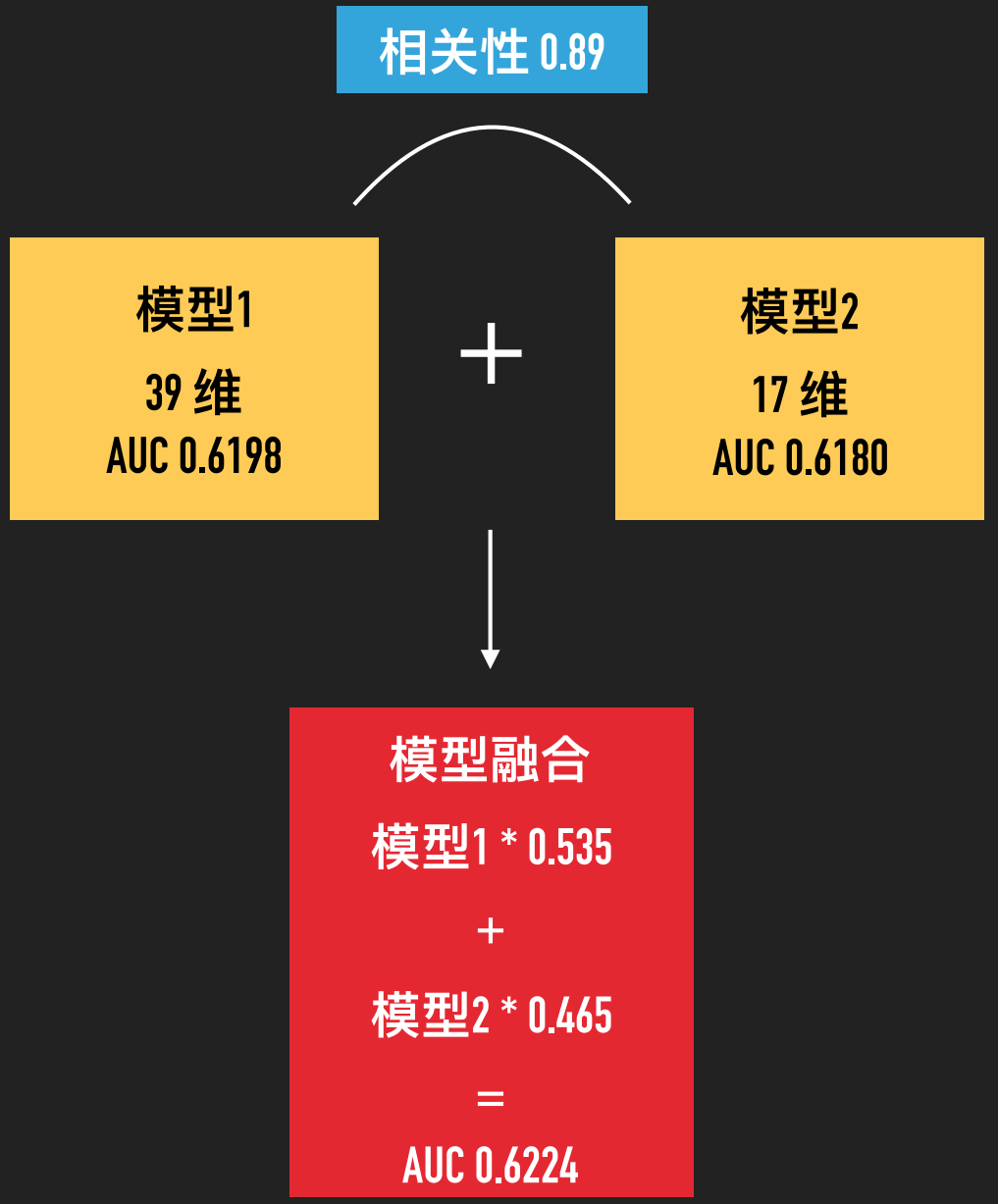
- 依据线上的模型成绩
  - 0.6198
  - 0.6180

采用加权融合的方式，按比例

- 0.535
- 0.465

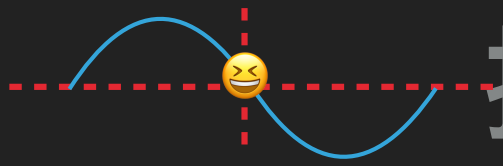
进行融合，最终成绩为0.6224

排名第一



感想融360给我们提供这次宝贵的经验，

希望能将数据挖掘运用到各个领域！



振动波组合