2019 厦门国际银行“数创金融杯”数据建模大赛
算法说明文档

队伍名称： 连扔正面 102 次

队伍成员： 曹威，卢俊杰，郑升圆

队伍排名： 初赛 a 榜：第 11 名

初赛 b 榜：第 3 名

目录

1 赛题分析及整体方案设计

1.1 赛题分析与理解

信用风险是金融监管机构重点关注的风险，关乎金融系统运行的稳定。在实际业务开展和模型构建过程中，面临着高维稀疏特征以及样本不平衡等各种问题，如何应用机器学习等数据挖掘方法提高信用风险的评估和预测能力，是各家金融机构积极探索的方向。本次竞赛提供实际业务场景中的信贷数据作为建模的对象，通过数据挖掘手段进行预测分析。

本题目的在于通过给出的数据来预测用户是否会有违规的情况，给出的特征包括用户基本属性信息的特征，借贷相关特征以及用户征信相关信息等。

1.2 思路

本赛题是一个典型的数据包含数值型、类别型以及多值离散型特征的问题，解决思路主要是利用决策树对数据进行处理，构造训练集进行训练，通过整合三方面的特征信息进行用户行为和用户信用度的预测，进一步判断用户是否违约。

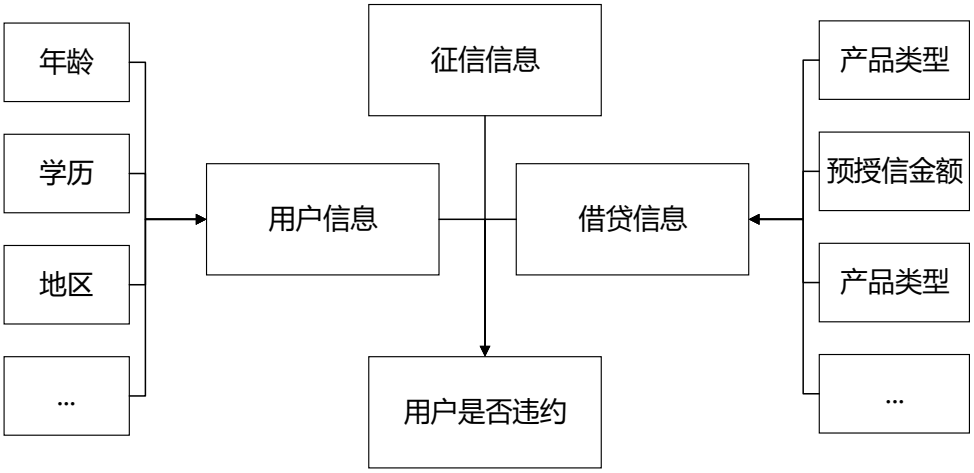

图 1 整体信息结构图

1.3 难点

(1) 多值离散特征的处理；

(2) 存在大量高维稀疏特征字段，且其意义不明，如何对其进行合理的处理是本赛题的一个关键难点所在；

(3) 数据正负样本比超过 100:1，存在严重不平衡的问题 ；

(4) 原始数据特征高达 100 多维，且存在许多涉密特征，因而对于原始特征的理解以及筛选对于减少特征维度以及特征工程均至关重要。

(5) 新旧数据部分特征分布不一致，如何在保证分布一致的情况下使用尽可能多的数据信息成为挑战

1.4 整体方案设计

在建模过程中，考虑到正负样本数严重失衡且数据量较小，为使得模型具有更强的表现力以及更好的泛化能力。我们在建模过程中采用了三个 gdbt 模型，为了使得模型融合效果最佳，对三个模型分别进行验证集扰动、特征扰动和模型扰动，单模型效果通过调参和特征选择来保证单模型最优，最后根据其预测结果进行加权融合。

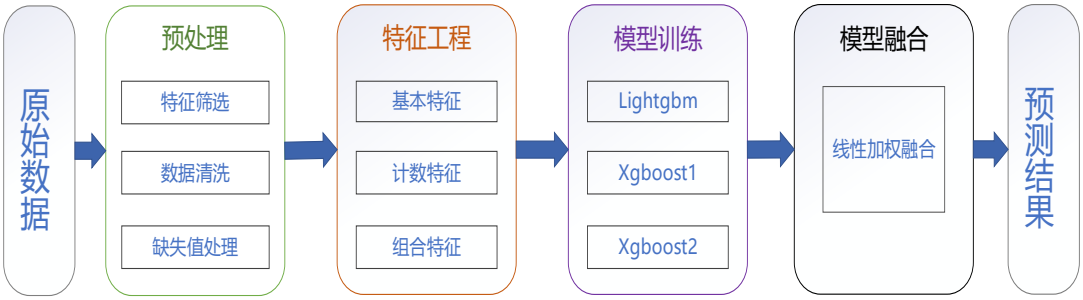

图 2　整体方案流程图

2 数据观察分析及预处理

2.1 特征理解

根据官方描述，数据集共分为三类特征，考虑到实际场景意义，我们做出以下分析：

表 1：相关特征理解分析

特征名称	业务理解与分析
certValiBegin certValidStop	certValidStop - certValiBegin：身份证有效期(单位 s)； 编码：（certValidStop - certValiBegin）/365/24/60/60 的结果大多数为 10，5，20，有极少数的异常值(异常值表示含义可能为长期有效)； 信息挖掘：根据身份证有效期，不同年龄段的人身份证有效期不同，据此，我们可以挖掘出申请人大致所处年龄段的信息。
cretID dist residentAddr	certID（证件号）：身份证前六位，dist（地区）：根据户口本或身份证上的地区进行编码得到的六位数字，residentAddr（居住地）：在申请相关借贷产品时填写的居住地； 编码：均使用身份证前六位的编码方式进行编码，均为 6 位数字，其中前 1、2 位数字:所在省份的代码；第 3、4 位数字:所在城市的代码；第 5、6 位数字代表所在区县的代码； 信息挖掘：根据这三个特征，我们可以挖掘出用户是否是在出生地申请相关借贷产品，是否搬过家，是否是在外地城市，是城市家庭还是农村家庭等信息。
bankcard	bankcard：放款卡号 信息挖掘：放款卡号存在缺失值（非-999），根据分析，网络放款一般不需要银行卡号，直接将贷款金额打入用户账号（一般为手机号注册账号）。而且，根据放款卡号，我们可以提取出银行卡的前六位，用来判断是那个银行，还可以判断出是普通的储蓄卡还是信用卡。
linkRela	linkRela：联系人关系 信息挖掘：必然存在一项为配偶或夫妻，可以通过此特征和 age，gender 特征，可以推断出一个人是否已婚。
Weekday， setuphour	信息挖掘：通过申请时间发现，业务的产生不限制时段，可以判定此数据应该包含线上借贷业务。

2.2 数据预处理

(1) 异常值处理

对于一般类别特征，通过筛选离群点对异常值进行过滤；对于特殊特征，根据实际意义进行调整，如观察分析后可以发现，isNew=0 和 isNew=1 的数据中部分特征编码方式不一样，其中 residentAddr 特征 isNew=0 中比 isNew=1 中大 300000，因而对 isNew=0 中的不为-999 的 residentAddr 特征减 300000，使得 isNew=0 和 isNew=1 的 residentAddr 特征编码方式相同。

(2) 缺失值处理

除了进行数据填充之外，考虑到实际场景中数据的缺失可能反应的用户信息，我们对用户的缺失信息进行了统计，比如 bankCard，residentAddr，highestEdu，linkRela 等字段，构造出的特征符合实际意义，在线上线下保持一致并且有提升。

(3) 针对高维稀疏片段

对于“用户征信相关信息”，x_0—x_78,以及 ncloseCreditCard，unpayIndvLoan，unpayOtherLoan，unpayNormalLoan，5yearBadloan 总共 84 个特征，其中 x_0—x_78，根据计算其两两特征之间的相关性，设置固定的阈值，以滤除相关性较大的特征。本程序中，阈值设置为 1，即滤除相关性为 1 的特征，减少数据的冗余。最后保留的特征为 46 个，减少近一半的特征，同时也进一步说明，只有相关性为 1 的特征之间存在大量冗余，相关性较大但是不为 1 的，还有会保留一些信息。并且在实际意义不明的情况下，再利用统计特征替换掉单一特征规避特征稀疏片段对于预测结果的影响。

(4) 关于采样

题目给出的数据存在严重不平衡的问题，正负样本比超过 100:1,在此基础上直接带入模型会存在预测结果偏差的问题，在此基础上考虑了下采样和按 isNew 特征进行采样的策

略，通过对数据的观察分析，lmt 是对结果影响最大的特征之一，我们对于 lmt 分布出现差别的部分进行下采样，发现在线上线下分数都有提升。

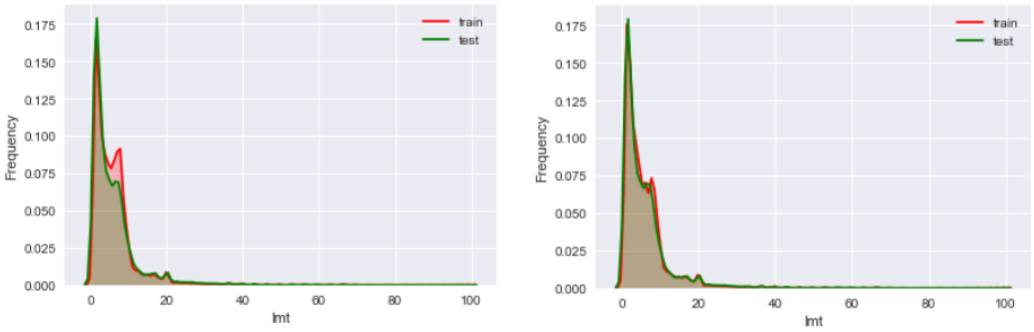

图 3 采样前后的 lmt 分布

2.3 数据分析

根据赛题中所给的原始特征以及对原始特征相应的理解，进行了以下的分析:

(1) 对 loan Product 即贷款类型的理解。从直观理解，不同的人会选择不同的贷款类型，因而可能会导致不同的失信率，在本赛题中的表现如图 4 所示，产品一和产品二的分布在训练集和测试集中有较大差距，不同的产品对应着不同的失信率，产品 1 失信率最高，约为 0.012150;产品 2 的失信率次之,约为 0.005681;产品 3 的失信率最低,约为 0.004387。也就是说 loan Product 这一特征对于失信率预测有较强的影响作用。

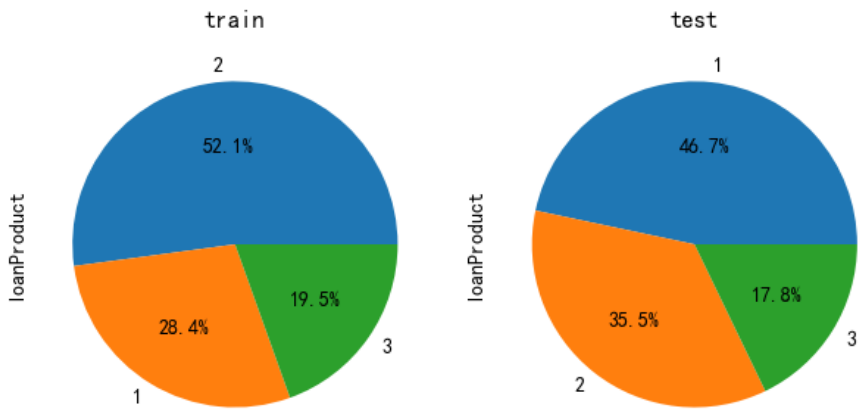

图 4 训练集和测试集 loanProduct 分布

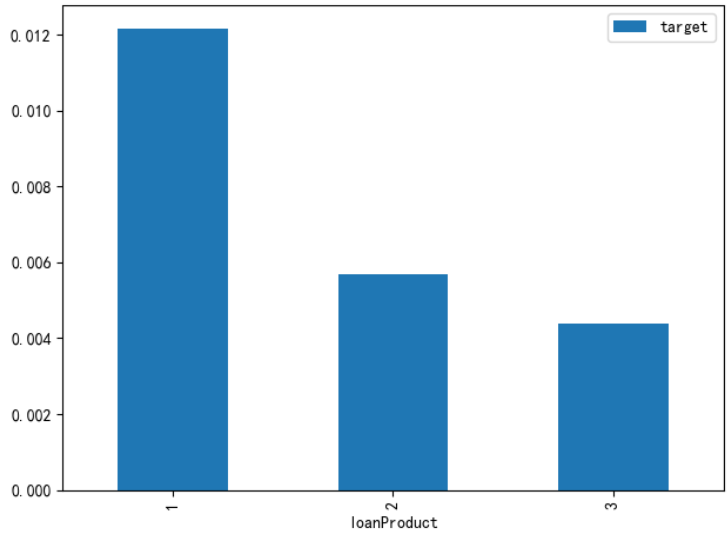

图 5 不同 loanProduct 失信率

(2) 对 basicLevel 的理解，不同的 basic Level 往往代表了用户的不同信用等级，因而不同的 basiclevel 会影响用户的失信率，其具体的表现如图 6、图 7 所示。可以看到训练集和测试集中的分布有较大差别，basiclevel 为 1、2、3 的失信率较低，略微有差别，basiclevel 为 4 的失信率很高，但其在训练集和测试集中样本均很少，分别为 48 和 6。

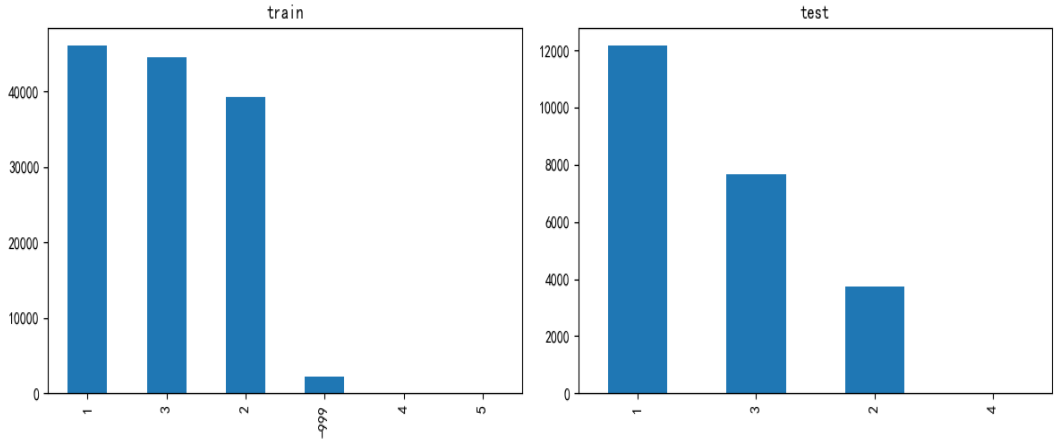

图 6 训练集和测试集 basicLevel 分布

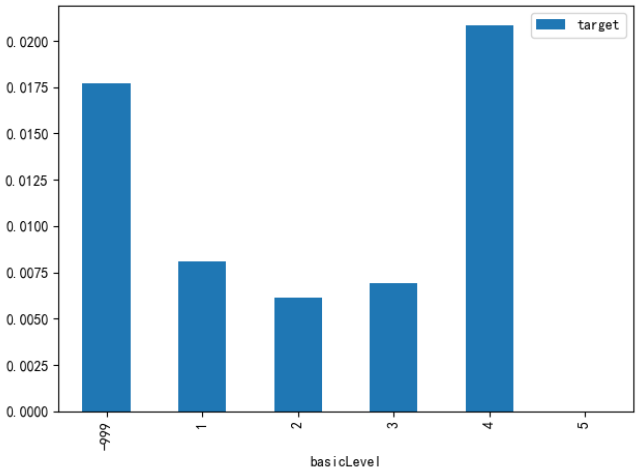

图 7 不同 basicLevel 失信率

(3) 对 gender 的理解，不同的性别可能会有不同的失信率，其具体分析结果如图 8、图 9 所示。可以发现训练集和测试集的 gender 分布大致一致，且不同的 gender 类型，失信率差别甚微，因而 gender 特征对于失信率影响不大。但可以考虑 gender 与年龄或者工作等其他特征的组合影响。

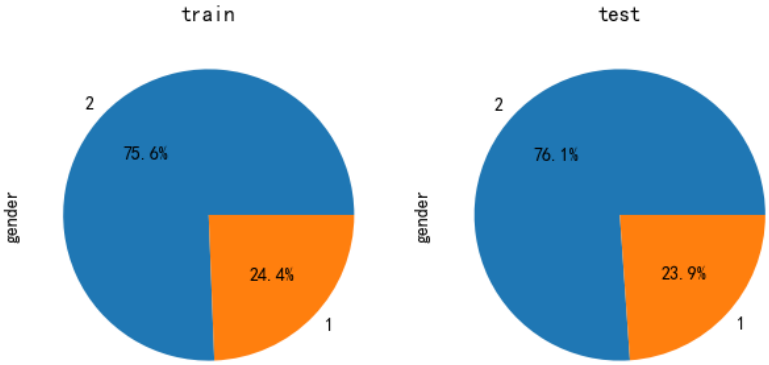

图 8 训练集和测试集 gender 分布

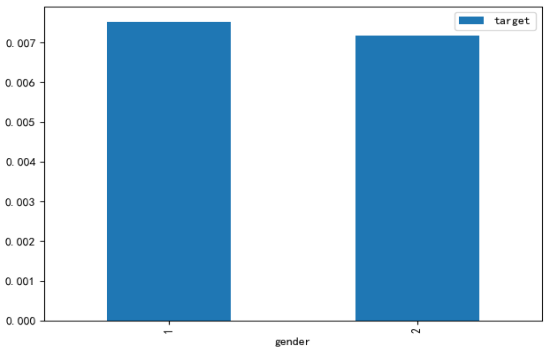

图 9 不同 gender 失信率

(4) 对 age 的理解，不同的年龄整体教育水平不同，以及不同年龄的消费观念都不一样，因而可能会有不同的失信率，其具体分析结果如图 10、图 11 所示。发现训练集和测试集中年龄分布类似，且年龄较大群体的失信率总体高于年龄较小群体。可以考虑对年龄做分箱或是其他相应的处理。

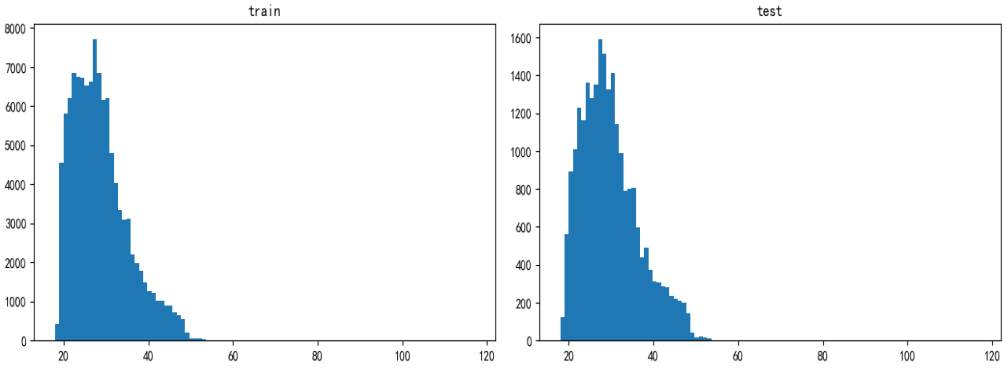

图 10 训练集和测试集 age 分布

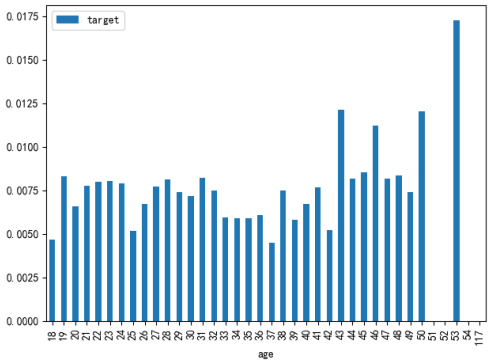

图 11 不同 age 失信率

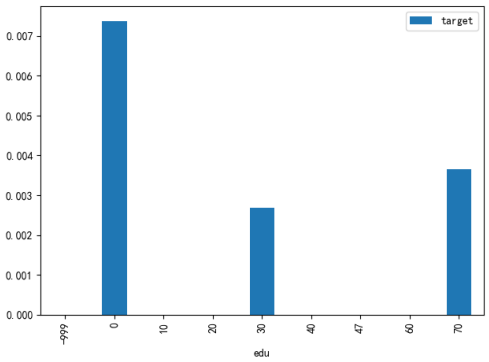

图 12 不同 edu 水平的失信率

(5) 对 edu 的理解，不同教育水平人群的消费理念以及其守信观念均不同，因而 edu 这一特征可能会影响失信率，其具体分析如图 12 所示。可以发现只有学历为 0、30 和 70 的出现失信情况，特别需要注意的是学历为 0 的数据超过了总数据的 90%以上。

(6) 对 job 的理解,不同的工作行业往往代表了不同的薪资水平以及不同的消费观念，因而失信率很可能与其从事的工作密切相关。其具体分析如图 13、图 14 所示。观察训练集和测试集分布可以看出其分布差距甚大，在训练集中占最多的 16 未在测试集中出现，同时

不同的工作会影响失信率，但是对于训练集和测试集样本分布不均衡的问题应重点关注。

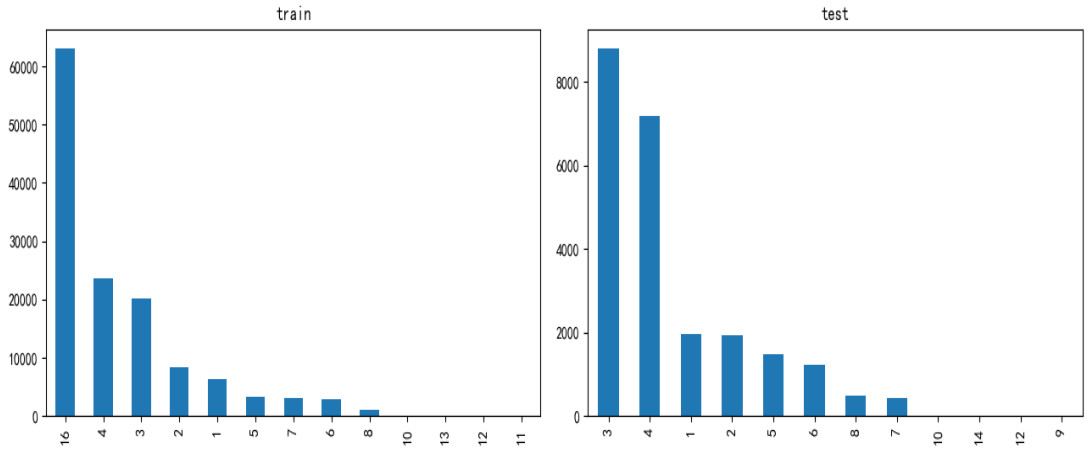

图 13 训练集和测试集 job 分布

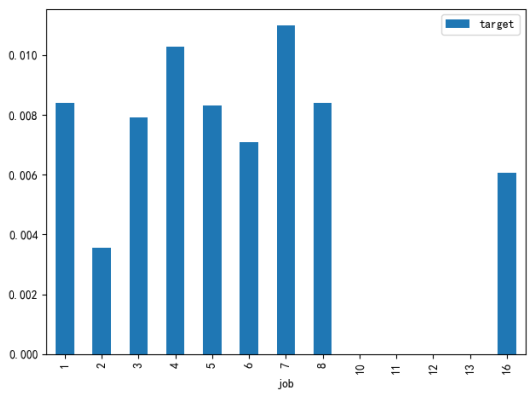

图 14 不同 job 失信率

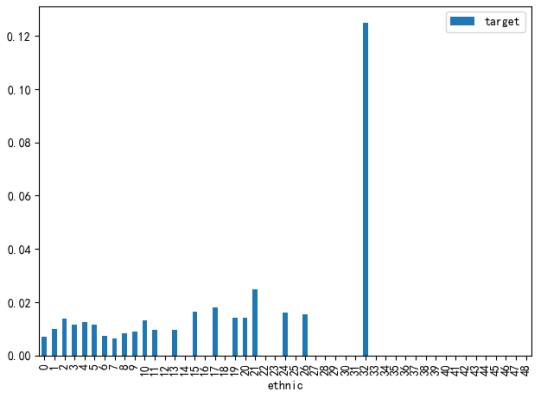

图 15 不同民族失信率

(6) 对 ethnic 的理解，不同的民族可能会影响失信率，其具体分析如图 15 所示。可以发现不同民族具有不同的失信率。但需要注意的是民族 0 的占比超过了总样本的 90%，失信率最高的 32 民族，仅有 8 个样本，因而参考意义较低，在后续处理应注意。

(7) 对 setupHour 的理解，在不同时段申请的人说明了其具有不同的作息习惯。在凌晨或者深夜申请的人说明其可能具有较差的生活习惯，因而其信用程度可能也相应的更差。据此理解，对 setupHour 做了以下分析，如图 16、图 17 所示。可以发现与之前分析类似，在凌晨 4、5、6 及深夜 11 点申请的失信率较高，可将其作为一个重要特征构建新特征。

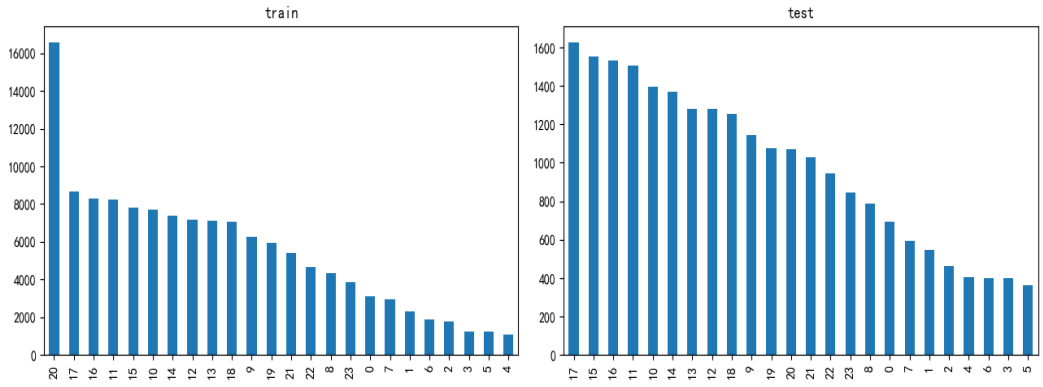

图 16 训练集和测试集中 setupHour 分布

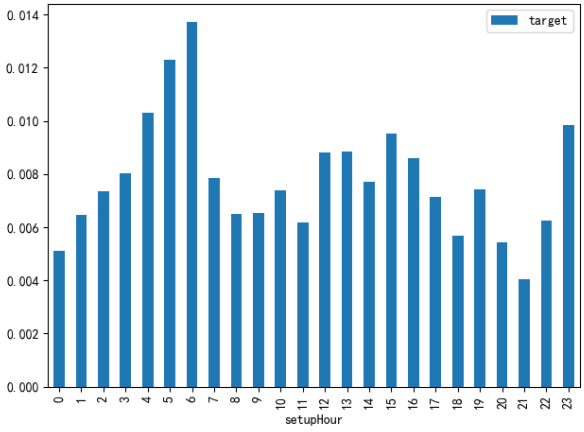

图 17 不同 setupHour 的失信率分布

3 特征工程

3.1 特征构造

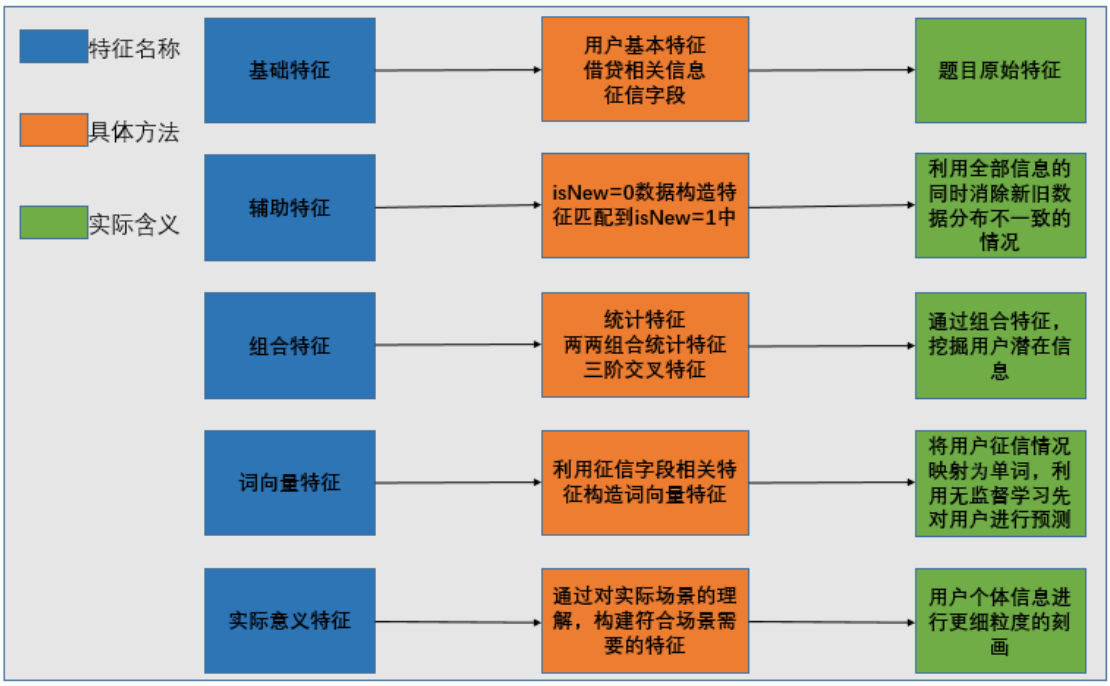

图 18 特征构造示意图

1. 基础特征

包括题目直接给出的用户基本属性特征，借贷相关信息以及征信字段相关信息三个部分。

2.辅助特征

用 certValiBegin 和 certValidStop 挖掘相关特征，提取出身份证有效期中的年，月等信息，再对提取出来的年，月，以及 job，gender，age 等特征，在 isNew=0 中，围绕 target，挖掘其统计特征，由于 isNew=0 的数据远大于 isNew=1 的数据，所以，将 isNew=0 中新得到的特征匹配到 isNew=1 中（包括 train 和 test）。由于引入了目标特征，所以容易造成数据泄露与过拟合，因此在构造特征时，通过线下检验剔除了容易对数据造成过拟合的特征，使匹配得到的特征在包含更多 isNew=0 数据信息的同时，又避免了 isNew=0 和

isNew=1 数据的分布差异。

3. 组合特征

通过加权，组合，排序等方式构造符合实际意义的基础特征组合，比如：

1）统计特征：用 certValiBegin 和 certValidStop 挖掘相关特征，提取出身份证有效期中的年，月等信息，再对提取出来的年，月，以及 job，basicLevel，ethnic 等 16 个特征每个特征的取值进行计数，构造计数特征。这些特征围绕 weekday 构造计数特征，实际业务含义也可以理解为通过工作日与工作时间来挖掘用户的潜在信息。再围绕 loanProduct 构造计数特征，挖掘不同的借贷产品对于不同用户，不同时间，不同地点等的相关信息。再围绕 residentAddr 构造相关的计数特征，统计出一个地区用户的喜好，习惯，消费水平等潜在信息。

2）两两组合统计特征：对于 loanProduct，lmt，basicLevel 等八个特征，统计其两两组合后不同取值的数量，再通过 isNew=0 数据，统计相同的特征匹配到 isNew=1 中（包括 train 和 test），最后计算出两组数据中相同取值的比例大小。

3）三阶交叉特征：通过比较不同特征下在 isNew 的分布情况，发现在 loanProduct 和 residentAddr 两个重要特征上的分布差异很大。对于 loanProduct，类型 1 和 3 的违约率都是上升而 2 却是下降的，而 residentAddr 的分布不同则可能反应了银行地区业务的变化。因此以这两个特征为基础构建了三阶交叉的各种统计特征，线上线下分数都有较大提升。

4. 词向量特征

一般构建特征的目的都是为了寻找相似。在本题中给出了用户的征信特征，这是一组非常稳定的特征。相比于构建各种交叉特征，造成特征冗余，可以利用 nlp 中的 word2vec 的方法构建词向量特征，把每个征信特征当作一个 word，用这种方式构建的特征相当于预先对用户进行了无监督学习，构建出的特征也能较直接的找出相近用户。在模型线下验证阶段，

加入词向量特征后不同随机种子下，模型的结果更加稳定。

5.实际意义特征

通过结合实际风控场景需求，对用户个体信息进行更细粒度的刻画，比如字段 certId，dist 具有相似和强相关性，通过选择 distance 特征作为衡量这两个特征的绝对值误差判断出生地到贷款地的距离特征等。

3.2 特征筛选

特征筛选主要对以下五种特征进行筛选剔除：

1) 具有高 missing-values 百分比的特征

2) 具有高相关性的特征

3) 对模型预测结果无贡献的特征（即 zero importance）

4) 对模型预测结果只有很小贡献的特征（即 low importance）

5) 具有单个值的特征（即数据集中该特征取值的集合只有一个元素）

特征选择结果：具有高 missing-values 百分比的特征数量为 0，具有单个值的特征数量为 0，特征重要性选择结果如图 1 所示，为表示清晰，只选取了重要性前三十的特征作图。

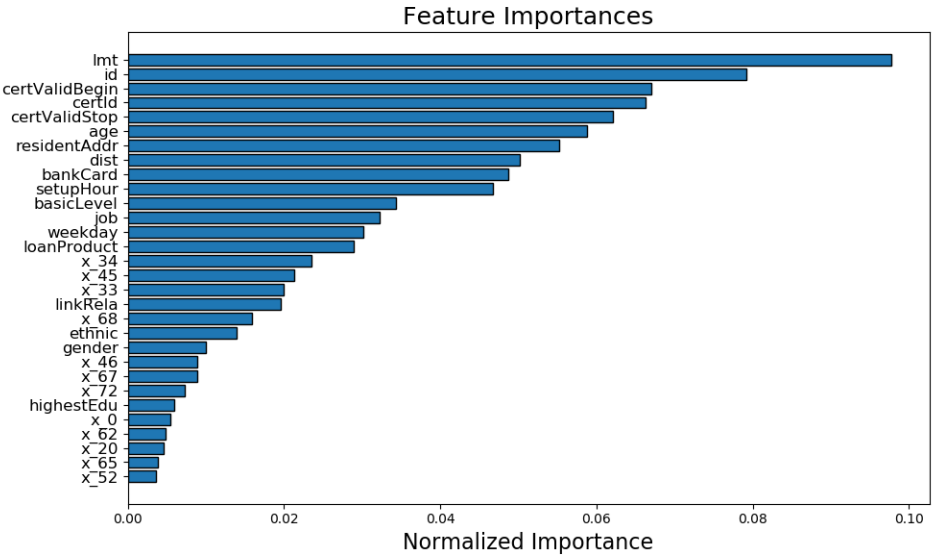

图 19　特征重要性排序

4 模型介绍

为使得模型具有更强的泛化能力，本文共使用了三个模型进行训练，分别进行参数扰动、特征扰动，单模型效果均通过调参和特征选择，保证单模型最优，对三个特征不同的模型按不同比例融合，最终生成模型结果，三个模型具体实现如下：

4.1 用全部数据训练的 xgb1

xgb1 直接将历史数据全部用于训练集，更好的训练模型，在特征选取上面，尽可能滤除新旧数据分布不一致的特征，通过自定义的特征筛选函数，最后保留的特征仅为原始特征数量的 60%左右。测试集的构建则是选取了训练样本中 20% target 为 0 的训练样本及 50%target 为 1 的训练样本作为测试样本，以此来达到平衡数据样本的目的。

4.2 交叉检验的 xgb2 和 lgb

Xgb2 和 lgb 采用不同的特征构造方式，利用交叉检验方式增加结果的稳定性，训练策略为随机选取 5 个种子，每个随机种子下，进行 5 折交叉验证，最后将得到的 25 个结果进行融合。种子数与交叉验证的次数越多，融合得到的结果更健壮。训练策略如下图所示：

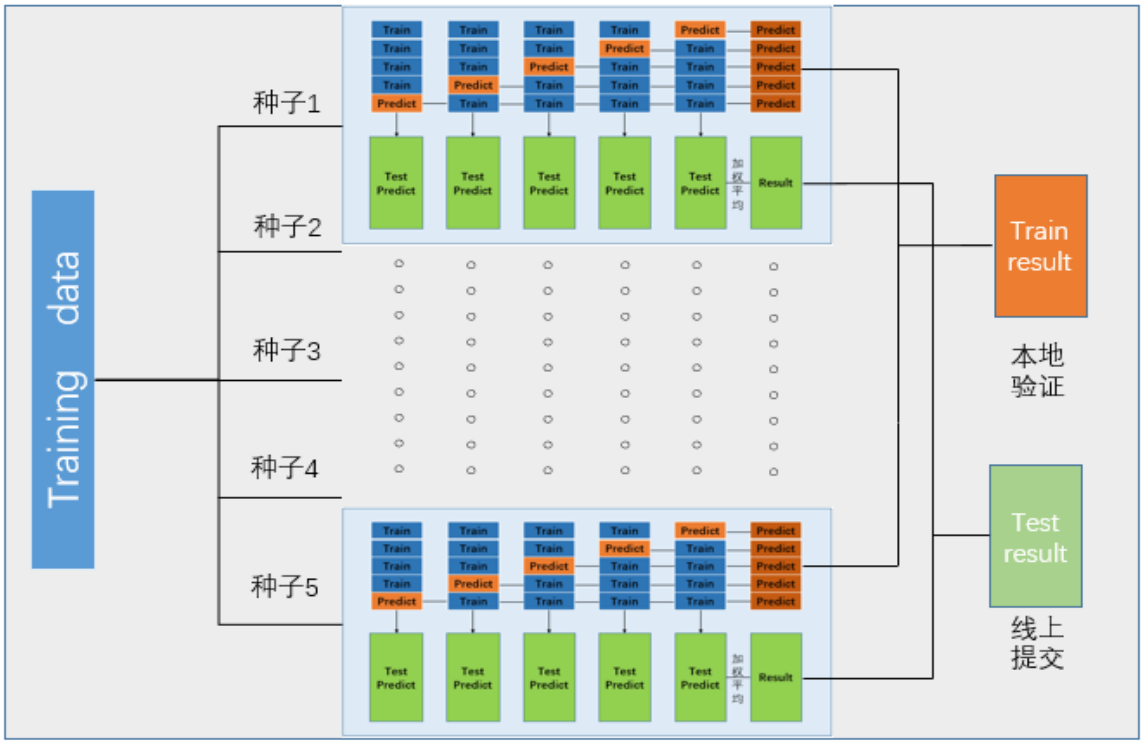

图 20　xgb2 与 lgb 模型结构

4.3. 模型评估

1.模型在 ab 榜中取得了相对稳定的成绩，初赛 11 名，复赛第 3 名，相较于其他复赛队伍，受到数据波动的影响较小，证明了模型有较好的泛化能力。

2. 规避了无意义的构造海量特征，利用有效的特征筛选实现了针对性的特征处理。

3. 为使得模型的泛化及表现能力更强，采用了多个模型综合不同的训练方式进行整合的方法来提高模型的稳定性，并且做了针对性的下采样。

表 2 初赛线上分数变化

模型	线上分数
Xgb2	0.80213
Xgb2+lgb	0.80756
Xgb2+lgb+xgb1	0.81215

5 本赛题创新点及研究展望

5.1 建模过程中的发现与创新点

在建模过程中，经过小组成员的不懈努力。主要有以下创新与发现：

（1）数据分析

1. 在常规基础特征上，根据实际数据分布进行了更加细粒度的调整，比如利用 lmt 的分布进行下采样，利用组合特征提取高维稀疏特征当中可能有用的信息等。

2. 通过 EDA 我们发现 isNew=0 和 isNew=1 的数据某些特征分布不一致。针对该情况，我们采取了如下策略：训练时只采用 isNew=1 的数据，但是由于 isNew=0 的数据量远大于 isNew=1 的数据量，所以 isNew=0 中的数据肯定包含相关信息。因而构造特征时，用 isNew=0 的数据构造了部分特征，匹配到 isNew=1 的数据中，在减小了数据量的同时使用了数据的全部信息，且解决了线下线上分数不一致的情况

(2) 特征工程构建

1. 考虑实际应用场景，针对不同类型的特征进行不同方式的处理，最大程度的实现了数据的精确处理。

2. 通过词向量方法预先对用户进行无监督学习，再送入模型中进行训练，增加了模型的稳定性和特征的可解释性。

(3) 样本不均衡问题的处理

在模型训练时，由于正样本数量过少，所以在每一折交叉验证中，人为添加 isNew=0 数据中 target 为 1 的样本，来调节样本比例极度不平衡的情况。

(4) 模型融合

为使得模型具有更强的表现及泛化能力，采用了多模型融合策略，使用的三个模型在数

据预处理、特征及模型训练均有较大的差距，且单模型表现均较优，最终的融合模型在初赛和复赛中均有较好的表现，体现出了模型强大的泛化能力。

5.2 不足与展望

1.对于影响程度大的特征，可以做更细致的特征分析，而对于意义不明的高维稀疏特征，可以考虑更多的挖掘它们和其他特征的关系来决定是否保留。

2.在正负样本严重不平衡的情况下，可以考虑更多有效的调整方式来平衡样本，使训练集和测试集的数据分布更加均衡，例如探索数据增强来平衡正负样本。

3.可以尝试更多的模型融合方式。例如更适合 auc 评价指标的 rank_avg 融合方式等。