数据挖掘项目报告——个人信用评估

1. 项目介绍

在银行贷款行为中，银行最关心的就是贷款人是否会违约从而造成贷款损失。而避免贷款被发放给可能违约的贷款人的主要手段就是对贷款人进行信用评估。

信用评估是银行对借款人信用情况进行评估的一种活动。银行贷款的最基本条件是信用，信用好就容易取得银行贷款支持，信用差就难以取得银行贷款支持。而借款人信用是由多种因素构成的，包括借款人资产负债状况、经营管理水平、产品经济效益及市场发展趋势等等。为了对借款人信用状况有一个统一的基本的正确的估价，以便正确掌握银行贷款，就必须对借款人信用状况进行评估。

（添加进入贷款违约预测的重要性）

赛题以金融风控中的个人信贷为背景，要求选手根据贷款申请人的数据信息预测其是否有违约的可能，以此判断是否通过此项贷款

1. （介绍实际生活中，发放贷款前关于贷款违约预测流程，以及一般会参考贷款人哪些方面的资料）

信用评估包括个人信用评分、企业信用评级和职业信用评价等。其中个人信用评分指信用评估机构利用信用评分模型对消费者个人信用信息进行量化分析，以分值形式表述。除此之外，还有个人综合信用评分，指通过使用科学严谨的分析方法，综合考察影响个人及其家庭的内在和外在的主客观环境，并对其履行各种经济承诺的能力进行全面的判断和评估。

一般针对个人信用评分时，会主要参考诸如个人基本信息、银行信用信息、个人缴费信息、个人资本状况四类变量，其中，银行信用信息所占权重最大，接近50%，其余三类变量所占权重大致相当。

1. 数据集

（介绍数据集来源）

（简略介绍数据集内容）

1. 数据处理
2. 数据各项特征含义

| Field | Description |
| --- | --- |
| id | 为贷款清单分配的唯一信用证标识 |
| loanAmnt | 贷款金额 |
| term | 贷款期限（year） |
| interestRate | 贷款利率 |
| installment | 分期付款金额 |
| grade | 贷款等级 |
| subGrade | 贷款等级之子级 |
| employmentTitle | 就业职称 |
| employmentLength | 就业年限（年） |
| homeOwnership | 借款人在登记时提供的房屋所有权状况 |
| annualIncome | 年收入 |
| verificationStatus | 验证状态 |
| issueDate | 贷款发放的月份 |
| purpose | 借款人在贷款申请时的贷款用途类别 |
| postCode | 借款人在贷款申请中提供的邮政编码的前3位数字 |
| regionCode | 地区编码 |
| dti | 债务收入比 |
| delinquency\_2years | 借款人过去2年信用档案中逾期30天以上的违约事件数 |
| ficoRangeLow | 借款人在贷款发放时的fico所属的下限范围 |
| ficoRangeHigh | 借款人在贷款发放时的fico所属的上限范围 |
| openAcc | 借款人信用档案中未结信用额度的数量 |
| pubRec | 贬损公共记录的数量 |
| pubRecBankruptcies | 公开记录清除的数量 |
| revolBal | 信贷周转余额合计 |
| revolUtil | 循环额度利用率，或借款人使用的相对于所有可用循环信贷的信贷金额 |
| totalAcc | 借款人信用档案中当前的信用额度总数 |
| initialListStatus | 贷款的初始列表状态 |
| applicationType | 表明贷款是个人申请还是与两个共同借款人的联合申请 |
| earliesCreditLine | 借款人最早报告的信用额度开立的月份 |
| title | 借款人提供的贷款名称 |
| policyCode | 公开可用的策略\_代码=1新产品不公开可用的策略\_代码=2 |
| n系列匿名特征 | 匿名特征n0-n14，为一些贷款人行为计数特征的处理 |

1. 数据缺失项的处理

[贷款违约预测2-完整的数据分析 - 简书 (jianshu.com)](https://www.jianshu.com/p/b797b043e9b6)

1. 离散特征的处理
2. 分类特征的编码
3. 时间特征的处理
4. 模型选择
5. GBDT

GBDT (Gradient Boosting Decision Tree) 是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT不仅在工业界应用广泛，通常被用于多分类、点击率预测、搜索排序等任务；在各种数据挖掘竞赛中也是致命武器，据统计Kaggle上的比赛有一半以上的冠军方案都是基于GBDT。

1. XGBoost

GBDT (Gradient Boosting Decision Tree) 是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT不仅在工业界应用广泛，通常被用于多分类、点击率预测、搜索排序等任务；在各种数据挖掘竞赛中也是致命武器，据统计Kaggle上的比赛有一半以上的冠军方案都是基于GBDT。

1. LGBMRegressor

[深入理解LightGBM - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/99069186)

LightGBM（Light Gradient Boosting Machine）是一个实现GBDT算法的框架，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据等优点。

为了避免上述XGBoost的缺陷，并且能够在不损害准确率的条件下加快GBDT模型的训练速度，lightGBM在传统的GBDT算法上进行了如下优化：

基于Histogram的决策树算法。

单边梯度采样 Gradient-based One-Side Sampling(GOSS)：使用GOSS可以减少大量只具有小梯度的数据实例，这样在计算信息增益的时候只利用剩下的具有高梯度的数据就可以了，相比XGBoost遍历所有特征值节省了不少时间和空间上的开销。

互斥特征捆绑 Exclusive Feature Bundling(EFB)：使用EFB可以将许多互斥的特征绑定为一个特征，这样达到了降维的目的。

带深度限制的Leaf-wise的叶子生长策略：大多数GBDT工具使用低效的按层生长 (level-wise) 的决策树生长策略，因为它不加区分的对待同一层的叶子，带来了很多没必要的开销。实际上很多叶子的分裂增益较低，没必要进行搜索和分裂。LightGBM使用了带有深度限制的按叶子生长 (leaf-wise) 算法。

直接支持类别特征(Categorical Feature)

支持高效并行

Cache命中率优化

1. 实验效果分析
2. 第一次试验

直接对所有特征进行回归计算

实验结果：

localAuth is :

0.7249421218515957

Prediction auc is :

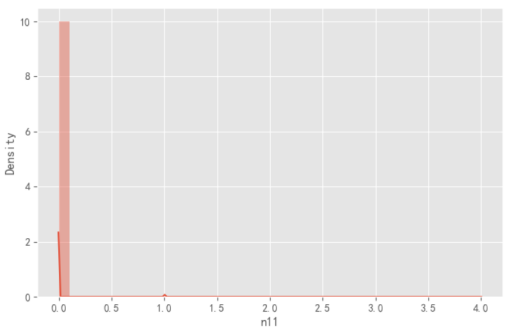
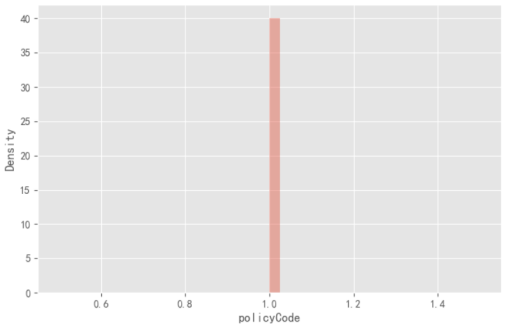
0.7241445091503432

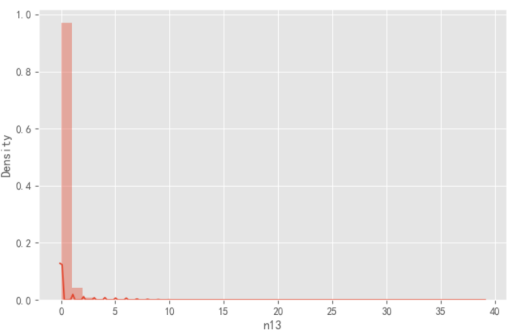
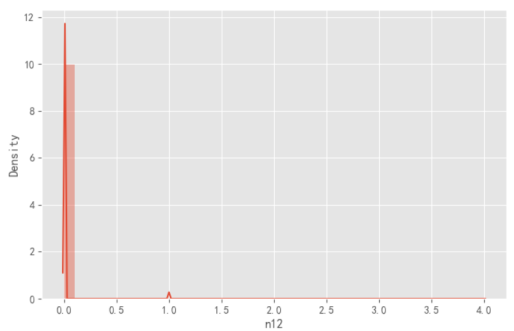
1. 第二次试验

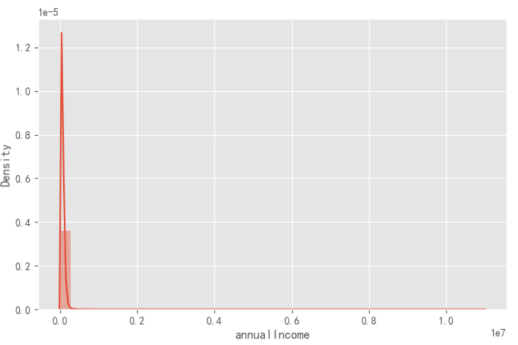
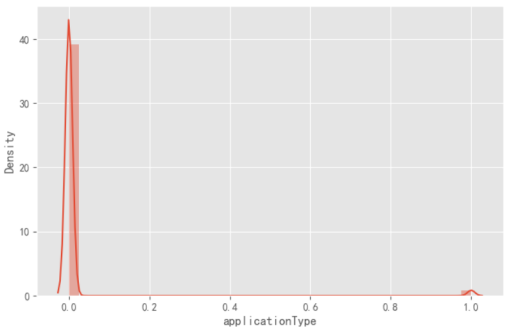
（降维：对具有强线性关系的多对特征选择性保留） + pca降维

（1）检查数据分布

policyCode, annualIncome, applicationType, n11, n12, n13数据上没有变化，删去

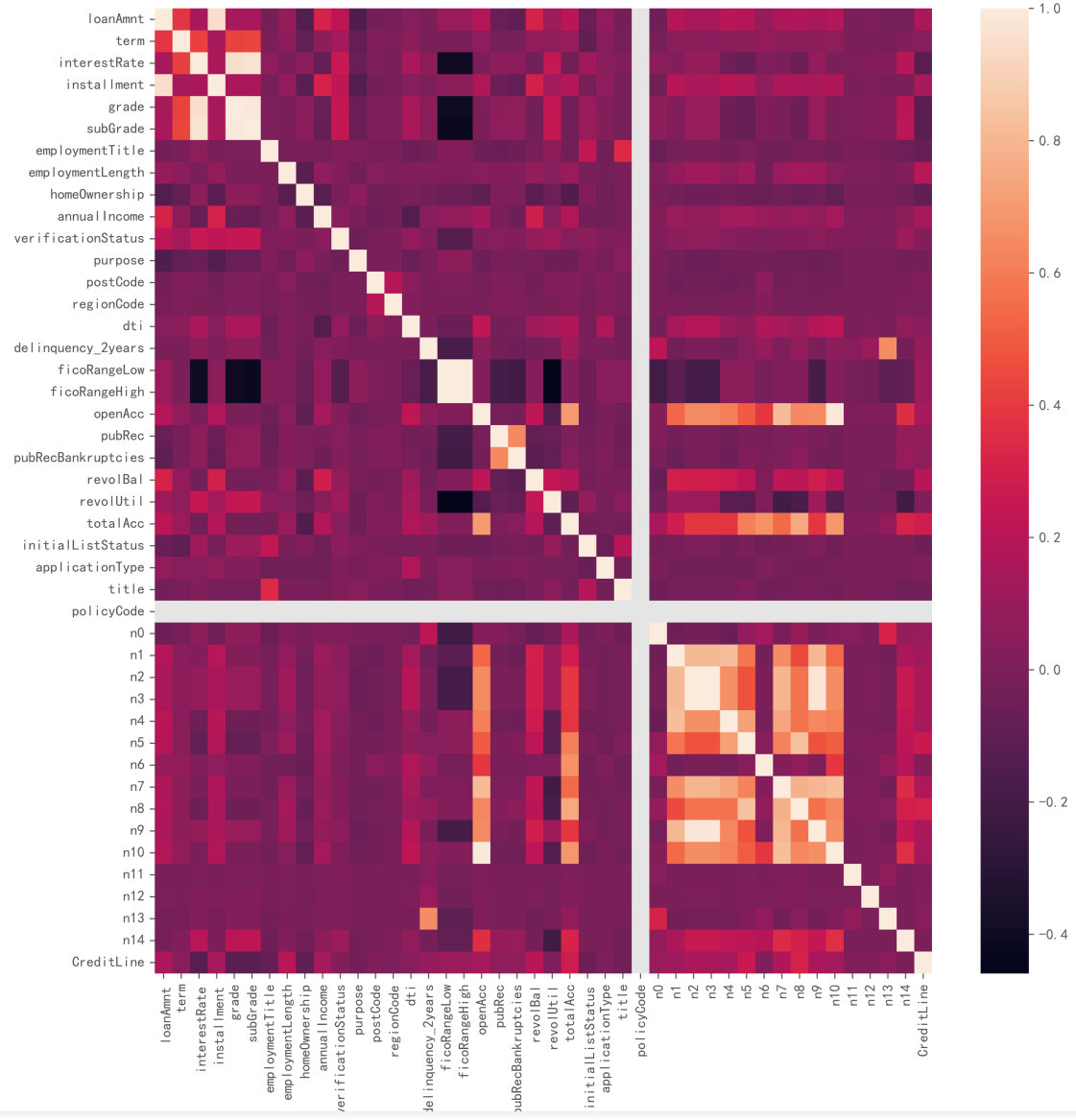






1. 分析特征间线性关系

先画出所有特征的热力图：

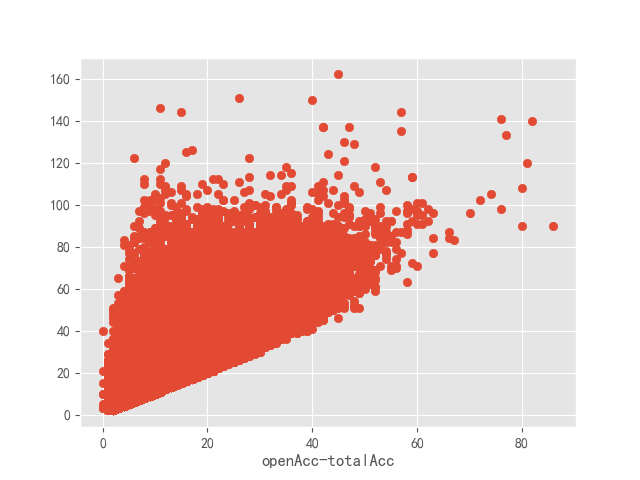
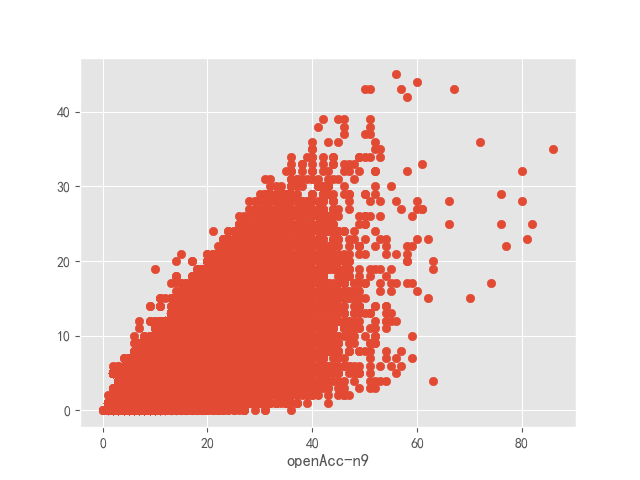
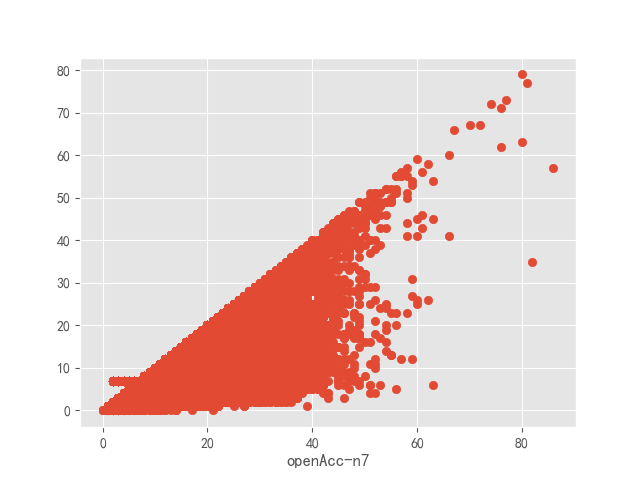
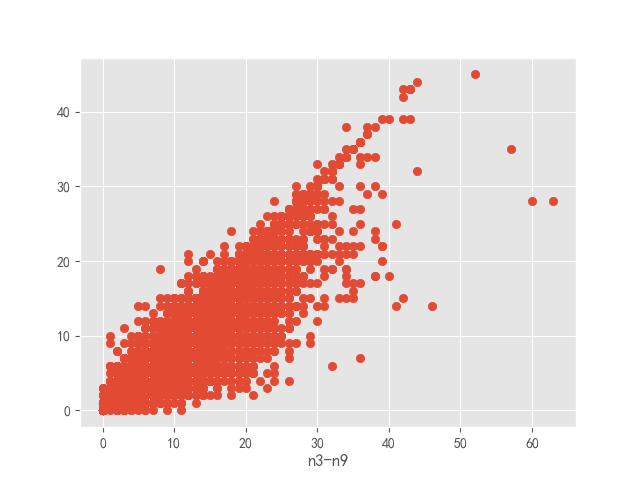
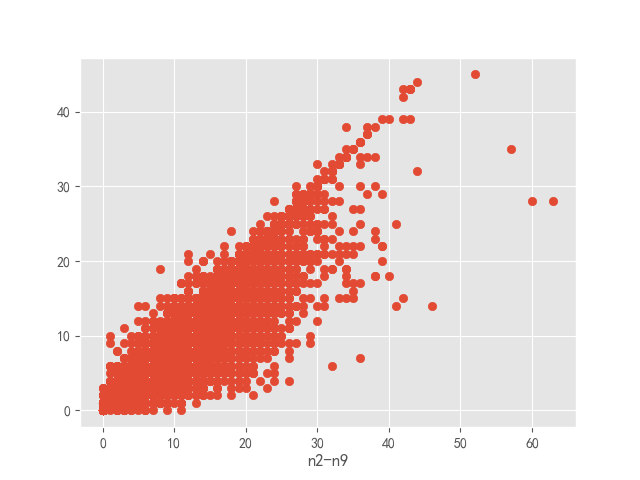
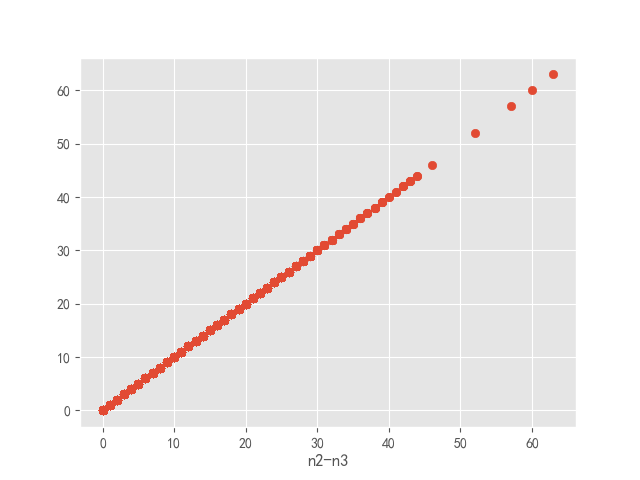
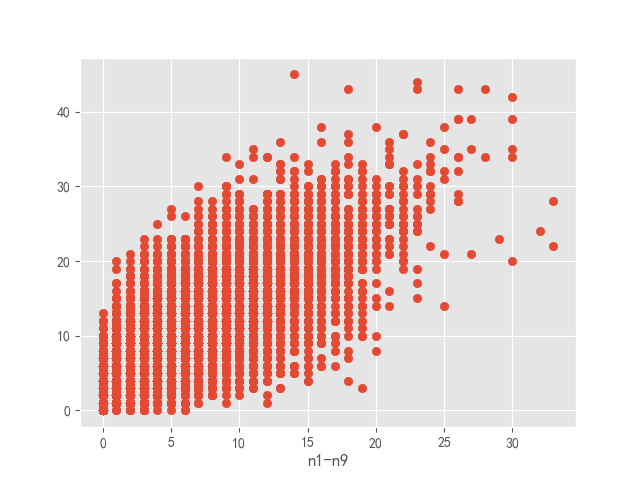
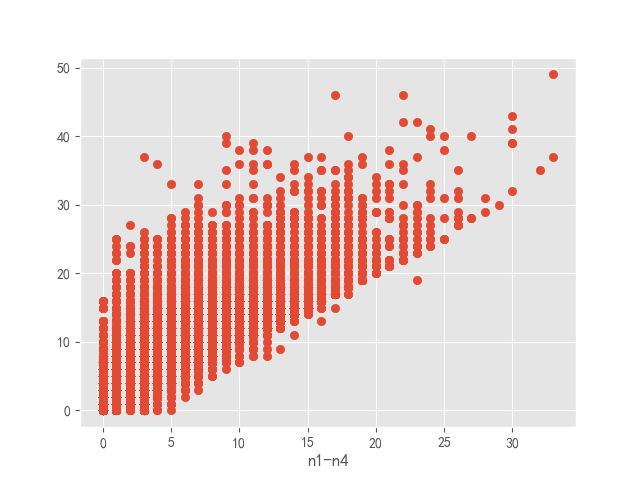
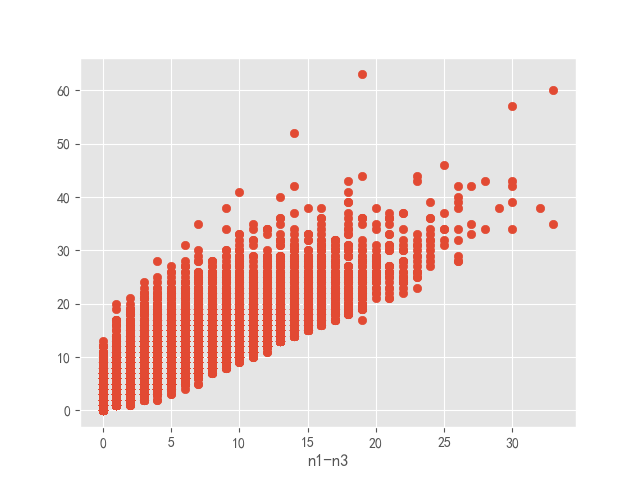
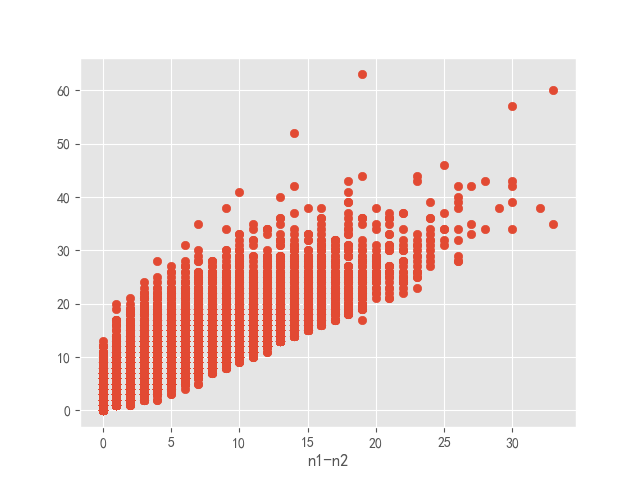
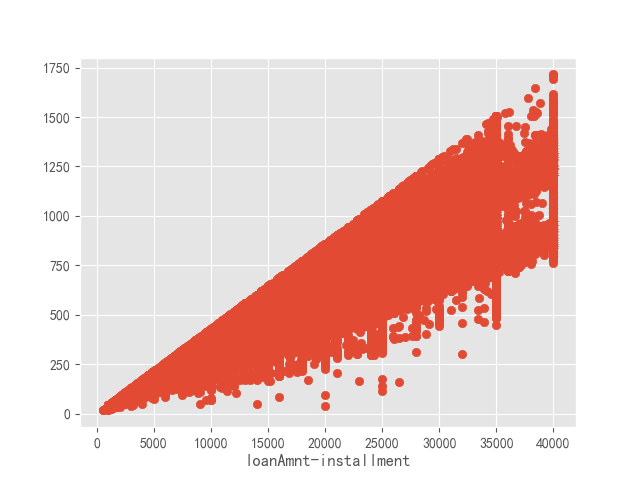
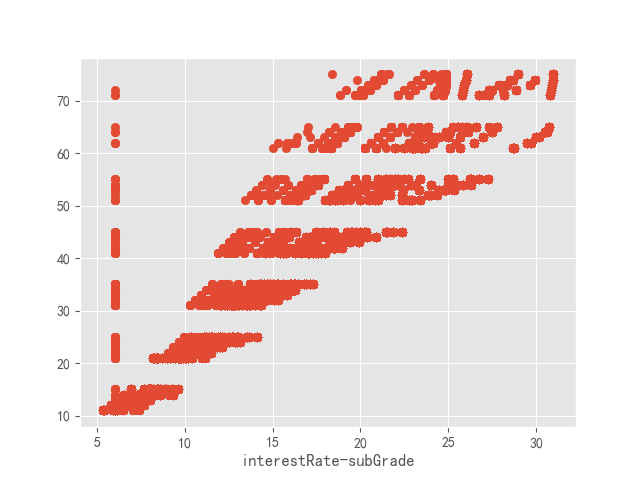
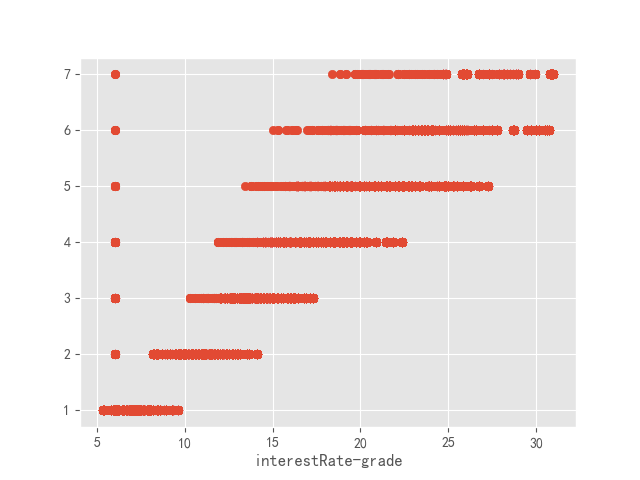
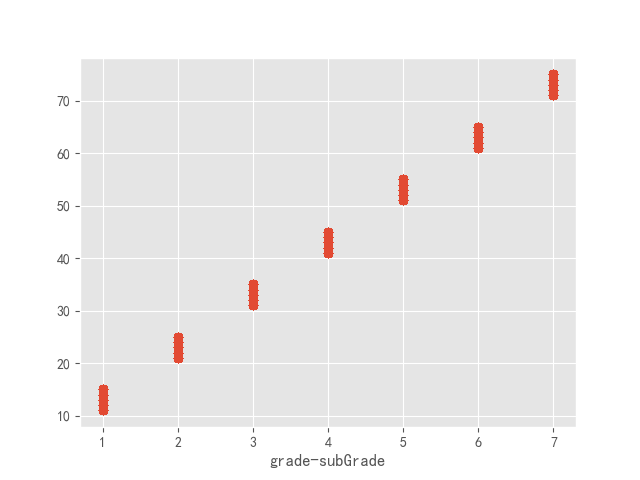
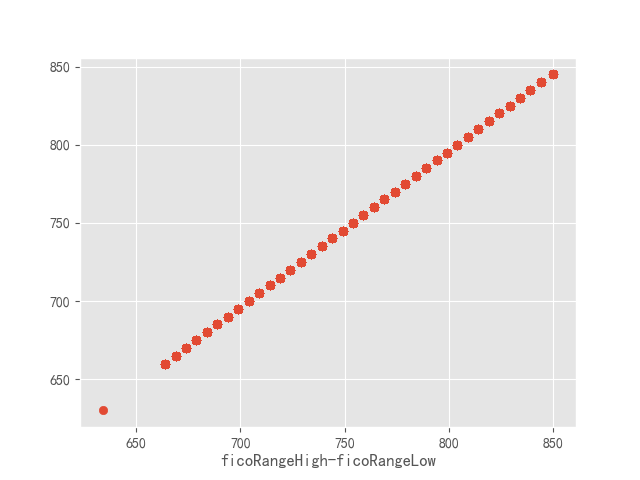


图中policyCode, annualIncome, applicationType, n11, n12, n13项的异常是由于其有效数据过少而造成的，由于在上一步中决定删去这些特征，因此此处不再进行讨论。

打印出相关系数较高的服务对：

|  |  |  |
| --- | --- | --- |
| feature\_x | feature\_y | corr |
| loanAmnt | installment | 0.9533692932239131 |
| interestRate | grade | 0.9532685750717439 |
| interestRate | subGrade | 0.9708469479718348 |
| grade | subGrade | 0.9939067441072746 |
| ficoRangeLow | ficoRangeHigh | 0.9999999288414679 |
| openAcc | n7 | 0.817523495370809 |
| openAcc | n10 | 0.9838900341270191 |
| n1 | n2 | 0.8081573447236511 |
| n1 | n3 | 0.801243568608452 |
| n1 | n4 | 0.8266510426867069 |
| n1 | n9 | 0.801243568608452 |
| n2 | n3 | 1.0 |
| n2 | n9 | 0.9820448979716367 |
| n3 | n9 | 0.838420270694751 |
| n5 | n8 | 0.838420270694751 |
| n7 | n10 | 0.826807072189599 |

继续分析其各自间相关性，绘制其分布关系如下：



分析后决定，loanAmnt-installment特征均保留；interestRate-grade-subgrade中删除grade；n1-n2-n3-n4-n9均保留；n2-n3均保留；openAcc-totalAcc-n7-n9中均保留。

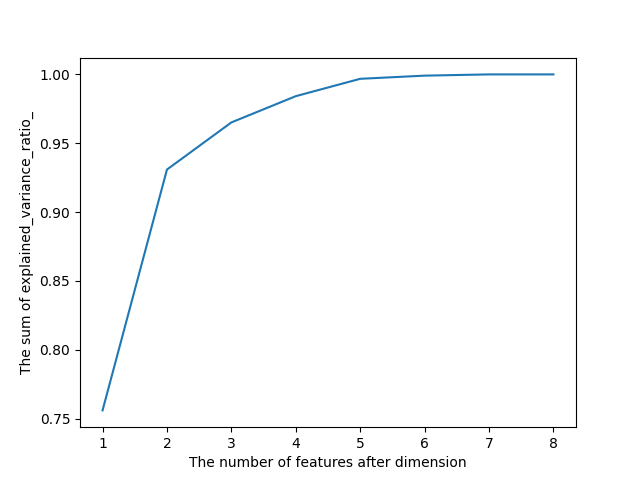
1. pca降维

[【机器学习】降维——PCA（非常详细） - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/77151308)

上一步中发现了具有强相关关系的三组特征["openAcc", "totalAcc", "n1", "n2", "n3", "n4", "n7", "n9"], ["loanAmnt", "installment"], ["interestRate", "subGrade"]

准备使用pca对这三组数据进行降维处理。其中第二三组均降至一维即可。而对于第一组特征的降维目标参数需要斟酌。

以下是对第一组数据降维后能保留的信息，由图可知，在仅保留3个维度的特征前提下，仍然能保证整体信息丢失较少。



pca参数[(31条消息) PCA中n\_components的设置\_林下月光的博客-CSDN博客](https://blog.csdn.net/weixin_41857483/article/details/109604239)

实验结果：

localAuth is :

0.7268395109024496

Prediction auc is :

0.7256975890664609

1. 第三次试验

调整模型参数