信用风险建模与机器学习

比赛背景：

违约损失率（Loss Given Default, LGD）是信用风险评估和巴塞尔协议框架下银行最低资本计算中的重要内容。LGD的定义为贷款违约后不能回收的部分所占违约暴露的比例。巴塞尔协议要求LGD需要在贷款审批时进行估计，且由于LGD受产品、抵押、环境等因素影响较大，对LGD的估计成为世界范围内银行风险管理的重要课题。基于此背景，本次比赛将对LGD进行建模预测，包括但不限于使用数学、统计学、机器学习等领域的模型与算法。

赛题描述

本次比赛包含两个数据集：训练集和测试集。在训练集中，Label为二元目标变量，Label为1表明LGD<1，Label为0表明LGD=1。每条样本有唯一独立id。其他字段为备选预测变量。字段释义请参见数据字典。 现要求预测目标变量，自选模型并估计参数，在测试集上进行验证。测试集与训练集数据结构和字段名称相同。建模过程中自选方法进行数据校验、数据清洗、特征工程、参数估计、模型拟合等步骤。 模型效果以在测试集上的AUC作为唯一评判标准并按照AUC排序名次计算初赛成绩。 建模语言环境R、SAS、Python、MATLAB任选。

评分规则：

比赛成绩由三部分构成：初赛成绩、代码成绩、答辩成绩。

最终成绩 = 40%初赛成绩 + 20%代码成绩 + 40%答辩成绩

背景分析

信贷风险的概念是银行业监管资本模型中的一个重要课题。信用风险是指当银行的客户开始拖欠贷款时（即，没有偿还贷款或没有偿还贷款），银行的贷款组合所遭受的损失。这些贷款可以是住房贷款、信用卡、汽车贷款、个人贷款、公司贷款等（即抵押贷款、循环信贷额度、零售贷款、整体销售贷款）。

此次比赛所给出的目标变量是违约损失率（Loss given default）是金融机构预测借款人违约造成的预期损失的重要计算方法，其指的是银行或其他金融机构在借款人违约时损失的金额，以违约时总风险敞口的百分比表示。金融机构的总违约率是在使用累计损失和风险敞口对所有未偿贷款进行审查后计算的。计算公式如下：

$$LGD = \frac{PD \times EAD}{EL}$$

其中，EL是预期损失（ Expected Loss），PD是违约概率（Probability of Default），EAD是违约风险敞口（Exposure at Default）。

根据F-IRB（基于基金会内部评级），银行计算自己的PD风险参数，而其他风险参数，如LGD和EAD，则由国家银行监管机构（即澳大利亚审慎监管局、美国联邦储备委员会/货币监理署、英国审慎监管局）提供，零售风险除外。巴塞尔协议A-IRB（基于高级内部评级）允许银行根据某些监管准则计算自己的所有风险参数。更多详情　请参阅国际清算银行（BIS）。

数据介绍

此次比赛一共有三个数据文件：

dictionary.xlsx为数据字典[1] LGD training.csv为训练数据，样本量为75046，变量为73个。 LGD test.csv为测试数据 ，样本量为19842，变量为72个。

探索性数据分析（EDA）

数据预处理

变量类型识别与转换

通过将数值型变量和分类变量的前5行分别打印出来，可以发现**revol_util**应为带有百分号的原因被划分为分类变量，若将之作为数值型变量处理，则会影响模型预测效果，因此，此处通过去掉百分号将**revol_util**处理为数值型变量。

	grade	home_ownership	purpose	revol_util
0	C	RENT	debt_consolidation	56.60%
1	B	MORTGAGE	debt_consolidation	61%
2	C	OWN	debt_consolidation	26.30%
3	C	RENT	debt_consolidation	38.20%
4	C	RENT	debt_consolidation	51.60%

异常值检测与处理

在数据挖掘中，异常检测是指识别出与大多数数据显著不同的个别数据。通常异常点会对预测有较大的影响，因为预测问题通产关注的是整体样本的性质，而异常点的生成机制与整体样本完全不一致，如果算法对异常点敏感，那么生成的模型并不能对整体样本有一个较好的表达，从而预测也会不准确。

异常值的检测方法通常有以下两种：

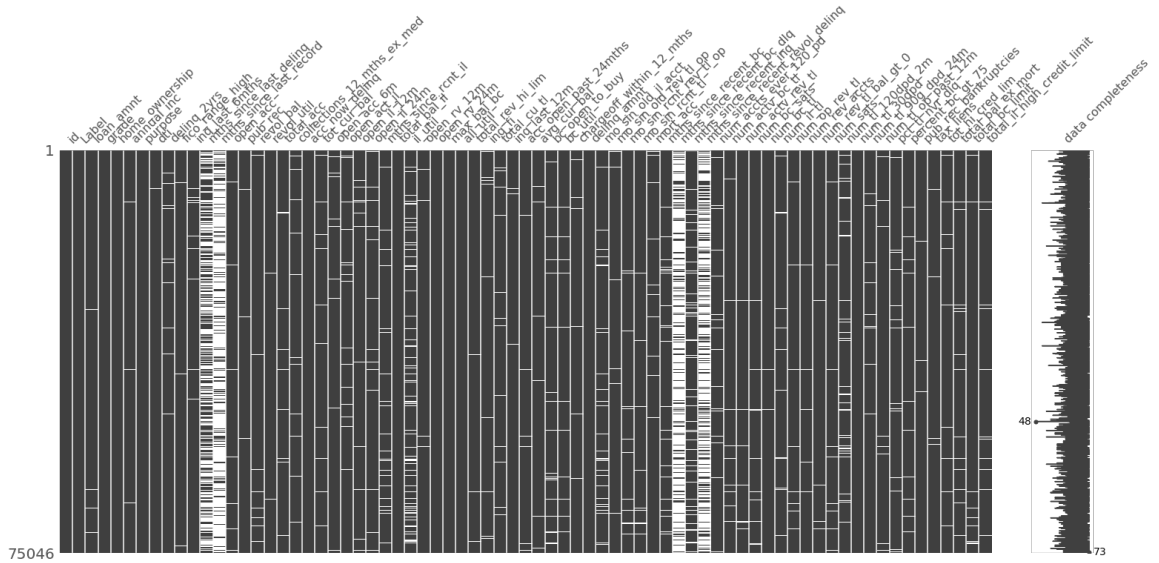
1、3σ原则：通过计算出变量的标准化分数（$z-score=\frac{x-\mu}{\sigma}$），正负3$\partial$的概率是99.7%，那么距离平均值3$\partial$之外的值出现的概率为$P(\frac{|x-u|}{\sigma}>3)<=0.003$，属于小概率事件，因此可以认为z-score绝对值大于3的样本值是属于异常值的。

2、箱线图：利用箱线图的四分位距（IQR）对异常值进行检测，IQR是上四分位与下四分位的差值，一般把超过四分位距1.5倍距离的样本值定义为异常值。

此处采用3σ原则来检测异常值，并将异常值处理为缺失值，以在后面的步骤中处理。

缺失值处理

首先通过画出训练集的无效矩阵数据密集图：

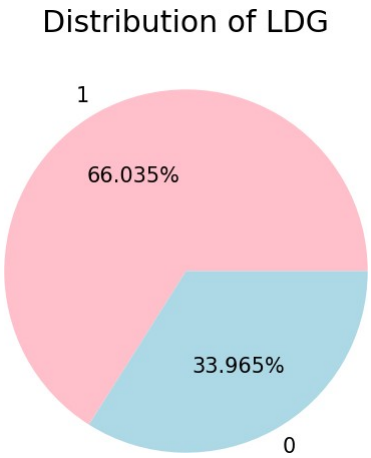


从上图可以直观的看出**mths_since_last_record**、**mths_since_recent_bc_dlq**、**mths_since_recent_revol_delinq**与**mths_since_last_delinq**这四个变量含有极高占比的缺失值，因此，可以直接将这四个变量删除。对剩下的变量的缺失值，则需要进行填充，此处将数值型变量的缺失值用其中位数填充（当变量的分布是非对称时，中位数往往比均值更能代表该变量的整体性质），定性变量则用众数填充。此时，由 (75046, 73)缩减为 (75046, 69)，测试集的维度由(19842, 72)缩减为(19842, 68)。

特征分析

目标变量观察

Label是一个二元目标变量，**Label**为1表明**LGD**<1，即违约损失率小于1，**Label**为0表明**LGD**=1，即违约损失率为1，也就是完全违约。相关研究表明，**LGD**概率分布通常是呈现双峰分布，即债务人的违约时的损失率要么比较低，在0附近，要么就会比较高，在1附近。

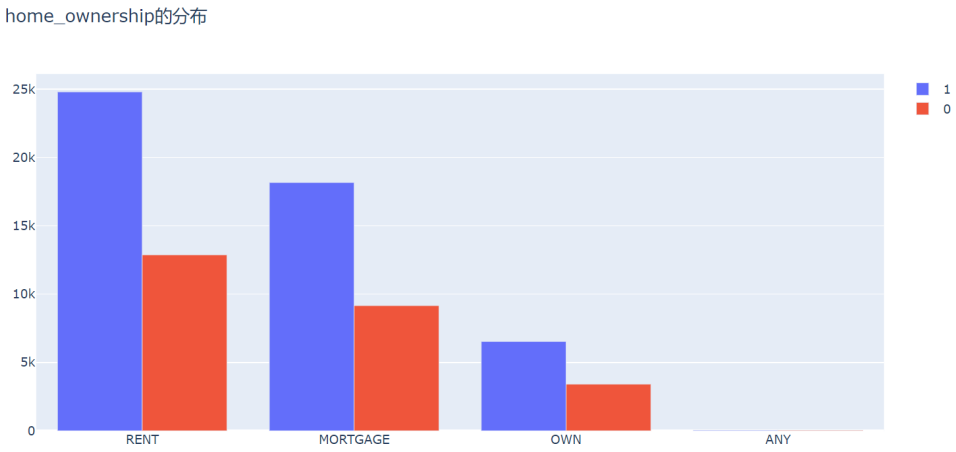


由上图可以知道正负样本比约为66:34，样本相对比较均匀。

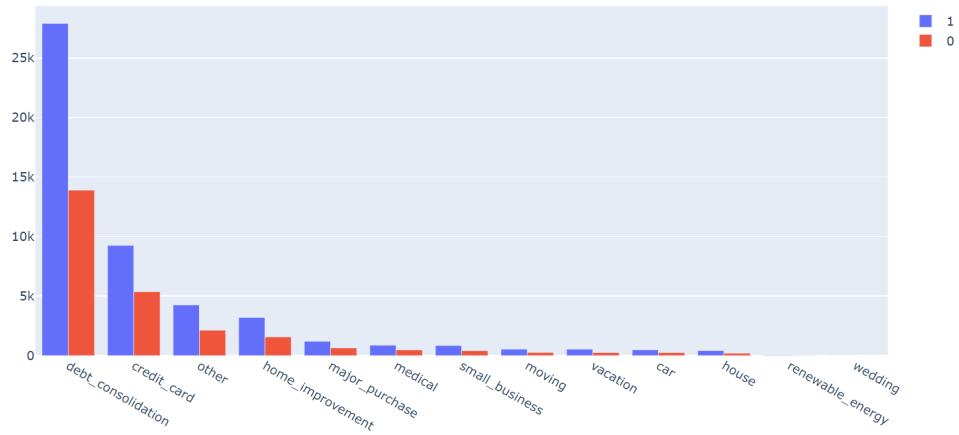
特征变量分析

分类变量

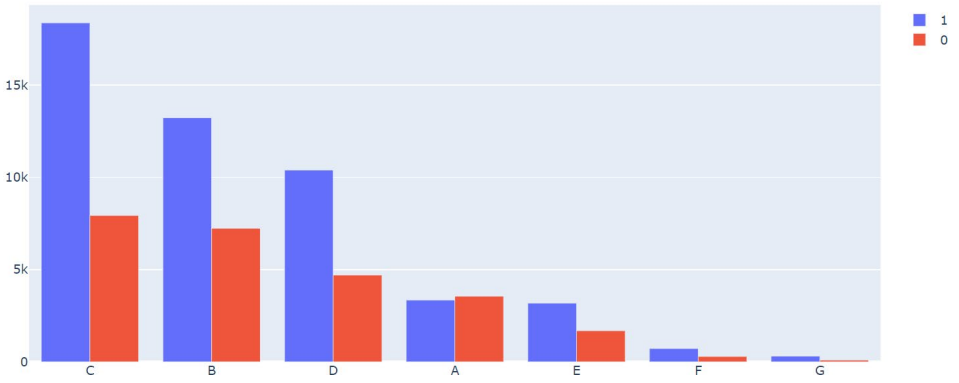
首先，观察分类变量的分布情况，并对分类过多、样本量较少且分布相似的变量值进行合并。



purpose的分布



grade的分布



通过观察上图可知，**home_ownership**的ANY类样本量极少，无法代表整体分布，但又不能直接对样本进行删除，因此只能将之合并进样本量最多的RENT类；**purpose**中分类十分多且变量值为'home_improvement','major_purchase', 'medical', 'small_business', 'moving', 'vacation', 'car', 'house', 'renewable_energy'和'wedding'的样本量较少，可以将这些样本合并到other类中；**grade**的分类虽然不多，但是可以将样本量较少的'E'、'F'、'G'合并为'low'类。

IV特征选取

IV的全称是Information Value，即信息量，其可用于衡量变量对目标变量的预测能力。变量的IV就越大，即对目标变量的预测能力就越强，它就越应该进入到入模型中。计算IV前，需要了解WOE，全称是“Weight of Evidence”，即证据权重，它是对是对原始自变量的一种编码形式。其计算公式如下：

$$WOE_i = \ln[\frac{p(G_i)}{p(B_i)}] = \ln[\frac{\#\text{ of } g_i/\#\text{ of } G}{\#\text{ of } b_i/\#\text{ of } B}] = \ln[\frac{\#\text{ of } g_i/\#\text{ of } b_i}{\#\text{ of } G/\#\text{ of } B}]$$

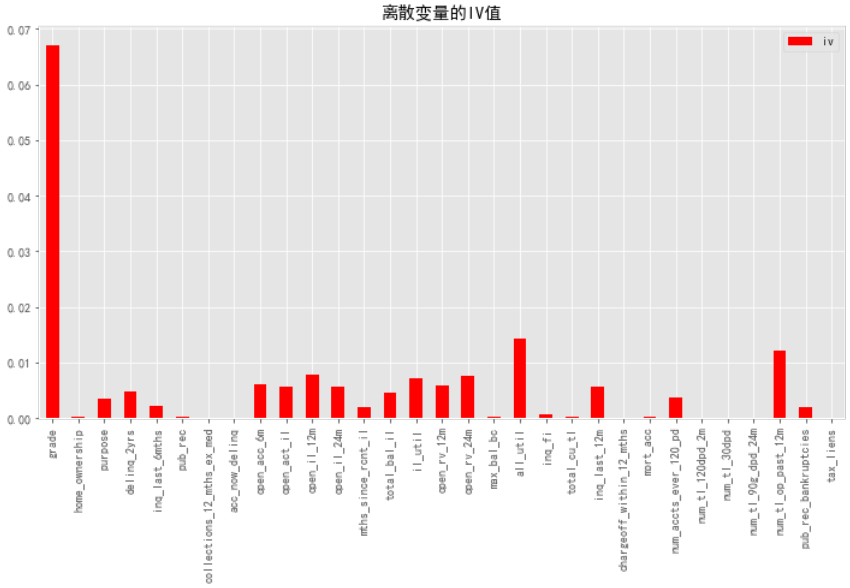
其中，$p(G_i)$是第i组中的LGD为1的客户占所有样本中所有LGD为1的客户的比例，$p(B_i)$是第i组中LGD<1的客户占所有样本中所有LGD<1的客户的比例。 $\#\text{ of } g_i$是第i组中LGD为1的客户的数量，$\#\text{ of } b_i$是第i组中LGD<1的客户的数量，$\#\text{ of } G$是所有样本中所有LGD为1的客户的数量，$\#\text{ of } B$是所有样本中所有LGD<1的客户的数量。

WOE_i表示的实际上是“第i个分组中LGD为1的客户占所有LGD为1的客户的比例”和“第i个分组中LGD<1的客户占所有LGD<1的客户的比例”的差异，因此，WOE的绝对值越大，这种差异越大。每一个分组i都会对应一个信息量IV_i，其计算公式为：

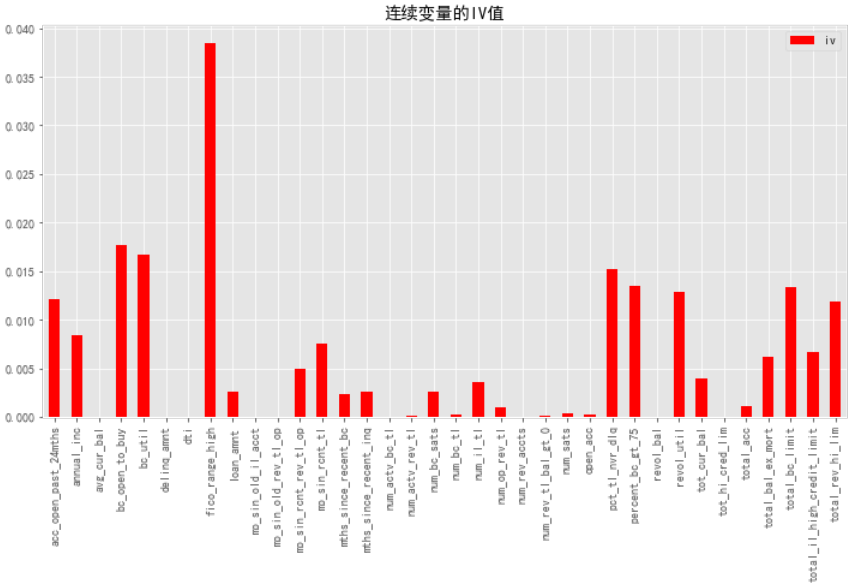
$$IV_i = WOE_i(p(G_i) - p(B_i))$$

获取每一个分组的信息量后，可以计算出该变量总的IV： $IV = \sum IV_i$。

因为离散型变量的每一个独立的变量值可以作为一个分组，因此，可以直接计算出离散变量的IV，其IV分布如下：

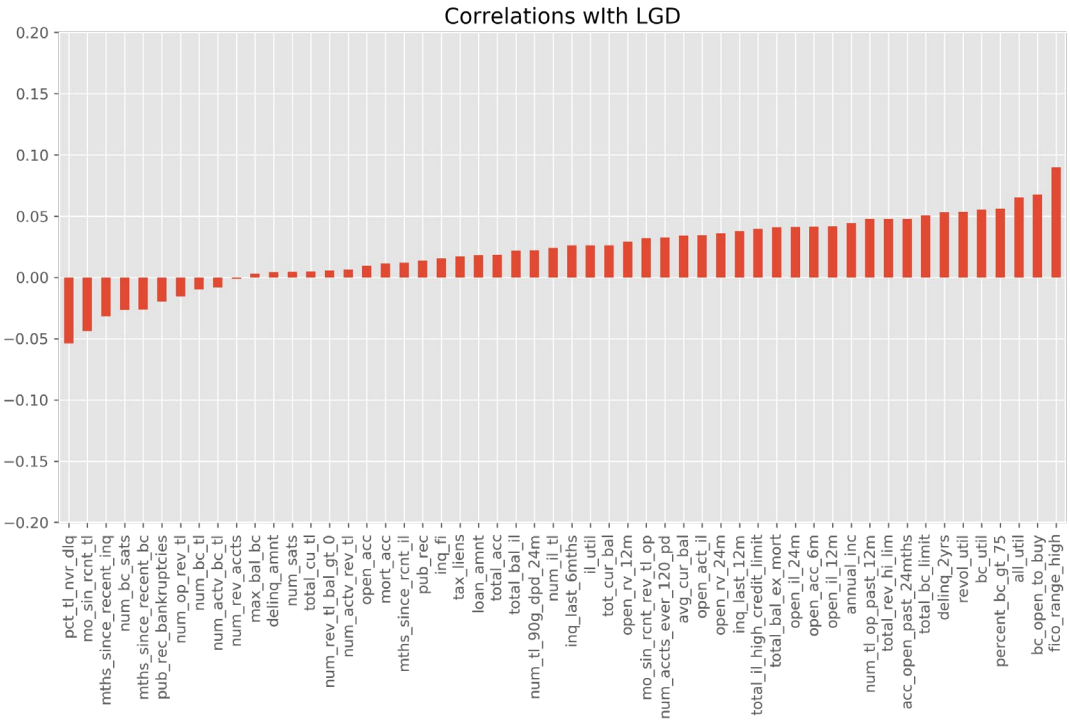


而关于连续型变量则需要将其离散化，此处采用的是最优分箱处理法并计算对应的IV，并将连续变量进行WOE 转换，其IV分布如下：



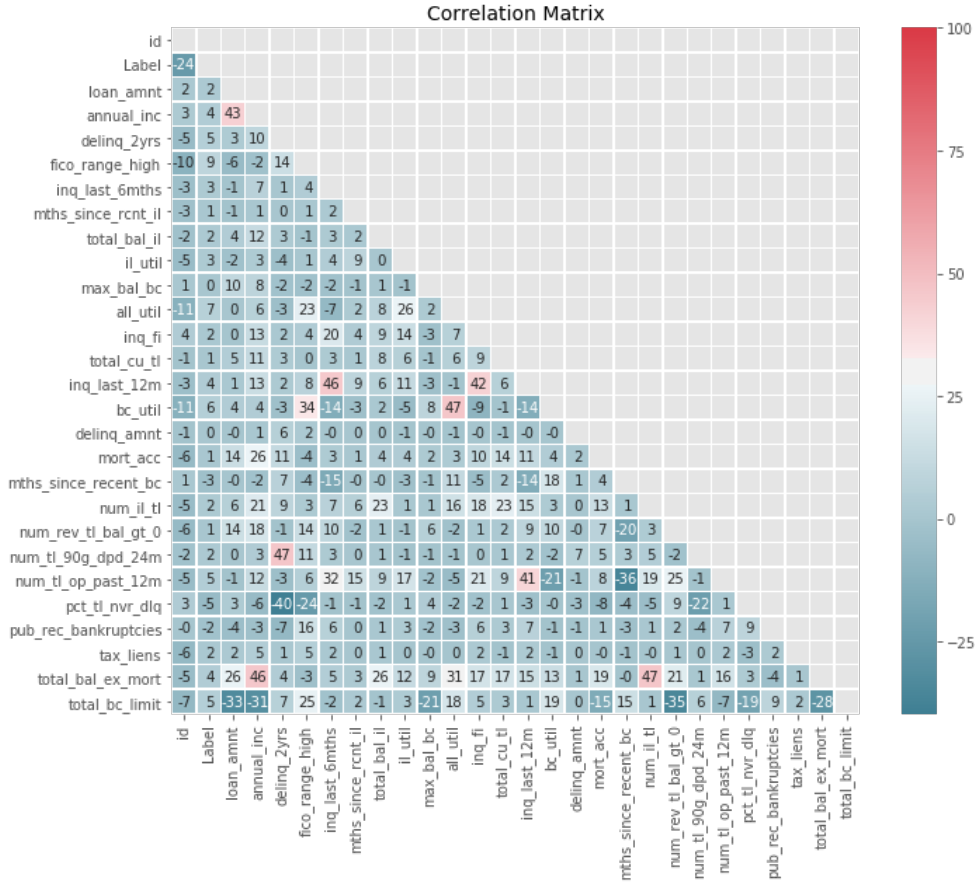
因为数据质量的原因，此处首先选择剔除IV为0的变量，由此，训练集的维度由 (75046, 69)缩减为(75046, 55)，测试集的维度由(19842, 68)缩减为(19842, 54)，均剔除了14个特征变量。

相关性分析



由相关系数图可以看出，特征变量与目标变量的相关性是比较低的，这表明特征变量与目标变量之间的线性关系是比较弱的，因此通过对目标变量与解释变量间的关系是难以发现业务逻辑的。

但为了提高模型的预测能力，此处还需要对变量间的相关性进行分析，并剔除相关性较高的变量。剔除规则是：当变量X_i与多个变量高度相关时（相关系数大于0.5），剔除X_i，若X_i仅与X_j高度相关时，则比较二者的信息量大小，剔除信息量较小的变量。由此，训练集的维度由 (75046, 55)缩减为 (75046, 30)，测试集的维度由 (19842, 54)缩减为(19842, 29)，训练集与测试集整整缩减了25个特征变量。降维后的训练集的相关系数图如下所示：



独热编码（One-Hot Encoding）

独热编码（One-Hot Encoding），其是使用N位状态寄存器来对N个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候，其中只有一位有效，即对于每一个特征，如果它有m个可能值，那么经过独热编码后，就变成了m个二元特征，并且，这些特征互斥，每次只有一个激活。该方法解决了许多机器学习算法不适用于处理定性数据的问题且能在一定程度上也起到了扩充特征的作用。

上面步骤中将**home_ownership**、**purpose**和**grade**进行合并，因此，再经过独热编码处理后，训练集的维度由(75046, 30)扩展为 (75046, 38)，测试集的维度由(19842, 29)扩展为(19842, 37)，由此均增加了8个二元特征变量。

构建衍生变量

已用授信额度与授信总额度比（risk_utt）：是衡量个人财务是否健康的指标之一，该比例越高，债务人的违约损失率往往也会越高，计算公式如下：

$$\begin{aligned}\text{已用授信额度与授信总额度比（risk\_utt）} &= \frac{\text{所有授信帐户的总授信额度 − 所有授信帐户的平均当前余额 × 授信账户总数}}{\text{所有授信帐户的总授信额度}} \\ &= \frac{\text{tot\_hi\_cred\_lim} - \text{avg\_cur\_bal} \times \text{open\_acc}}{\text{tot\_hi\_cred\_lim}}\end{aligned}$$

计算该衍生变量的信息量得$IV = 0.01825$，结合上面的分析可知，这样的衍生变量对目标变量的预测能较好，因此，该变量的构造是有效的。

曾违约示性：根据借款人曾开立的授信交易中从未逾期交易的占比来判断借款人否有不良信用记录，将有过逾期交易的借款人标记为1，没有的标记为0。

建立模型

建立模型之前，对训练集和测试集进行min-max标准化处理以提升模型的收敛速度和模型的精度，再将训练集按8：2的比例分出一份验证集，从而得到三份数据集：训练集，验证集和测试集。

分类模型中比较常用的有逻辑模型、线性判别模型和随机森林。

Logistics Model

考虑n个独立变量的向量$x = (x_1, x_2, \ldots, x_n)$，Logistics 模型则可表示为：

$$\mathbf{P}(y = 1|x) = \Pi(x) = \frac{1}{1 + \exp(-h_\theta(x))}$$

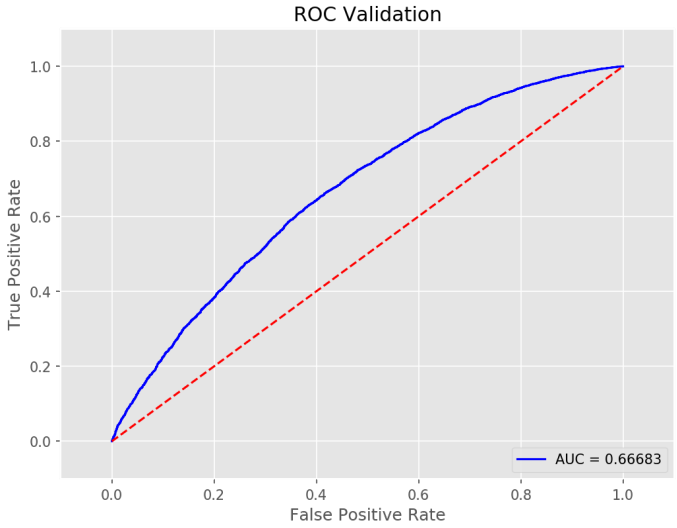
损失函数为：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\text{cost}\left(h_\theta\left(x^{(i)}, y^{(i)}\right)\right),\ \text{其中：}\ \text{cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) \text{ if } y = 1 \\ -\log(1 - h_\theta(x)) \text{ if } y = 0 \end{cases}$$

此处我们采取L2正则化对模型进行拟合，该模型的损失函数具体为：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)} \times \log\left(h_\theta\left(x^{(i)}\right)\right) - \left(1 - y^{(i)}\right) \times \log\left(1 - h_\theta\left(x^{(i)}\right)\right) + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

经拟合可得测试集的AUC为0.66683，结果如下：



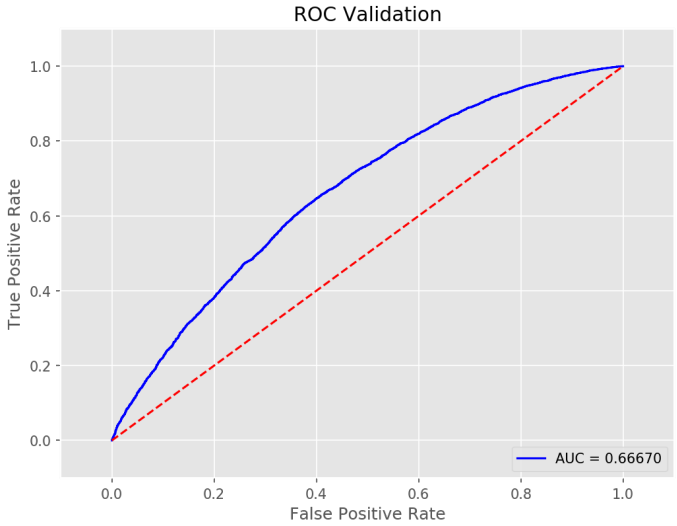
Linear Discriminant Analysis

线性判别分析是一种有监督的算法，主要用于降维和分类预测，其主要思想是：LDA用来寻找两个群体间“最好”的线性组合法则，来最大限 度地区分两个群体（也可用于多分类问题）。

线性判别分析寻找一个 维变量的线性组合（投影方向$\mathbf{a}$），使得两组间投影后的$\mathbf{a}'\bar{\mathbf{y}}_1$和$\mathbf{a}'\bar{\mathbf{y}}_2$ 的“标准化距离”最大。在该问题中，LDA的目标是寻找一个方向向量$\mathbf{a}$使得两类样本中心点尽量分离，即使$t(\mathbf{a})$最大化，$t(\mathbf{a})$计算公式如下：

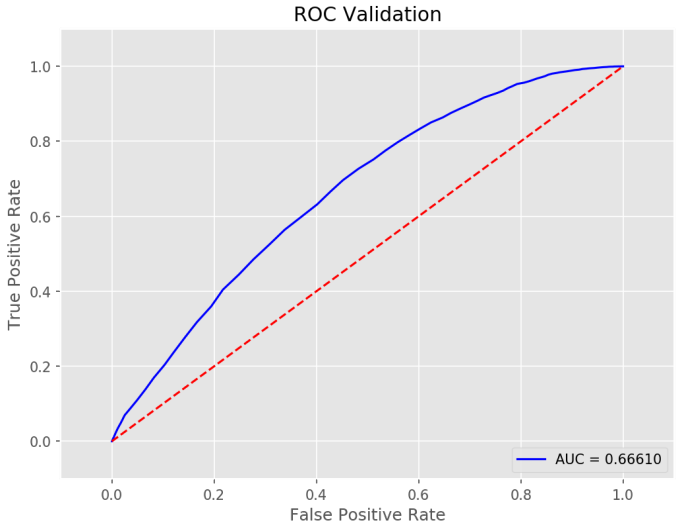
$$t(\mathbf{a}) = \frac{\mathbf{a}'\bar{\mathbf{y}}_1 - \mathbf{a}'\bar{\mathbf{y}}_2}{\sqrt{(1/n_1 + 1/n_2)\,\mathbf{a}'\mathbf{S}_{pl}\mathbf{a}}}$$

经拟合可得测试集的AUC为0.66670，结果如下：



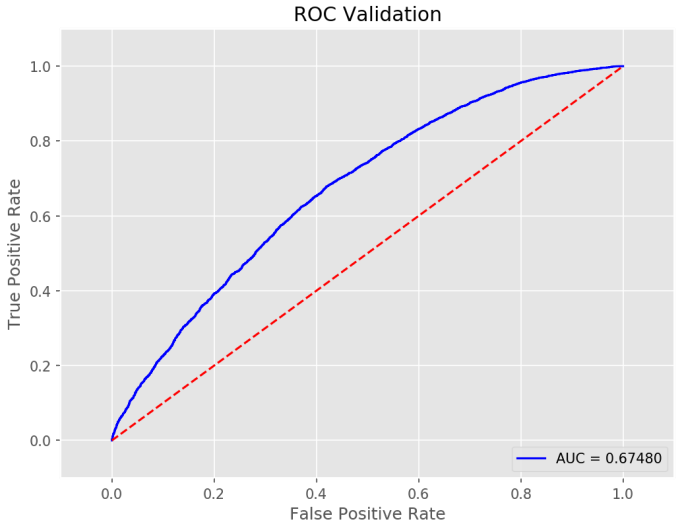
Random Forest Classifier

随机森林是一种通过多棵决策树进行优化决策的算法，其将多棵决策树组合在一起，在略微不同的训练集上训练每个决策树，在每棵树中仅考虑有限数量的特征来拆分节点。随机森林最终的预测也是通过平均每棵树的预测来得到的，经拟合，测试集的AUC为0.66610，结果如下：



Voting classifier

逻辑模型、线性判别分析模型和随机森林模型在测试的表现几乎差不多，因此，采用投票分类器将上面的逻辑模型、线性判别分析模型和随机森林模型集成以提升模型的泛化性预测能力。此处，选用soft模式，将所有模型预测样本为某一类别的概率的平均值作为标准，概率最高的对应的类型为最终的预测结果。经拟合，测试集的AUC为0.67480，结果如下：



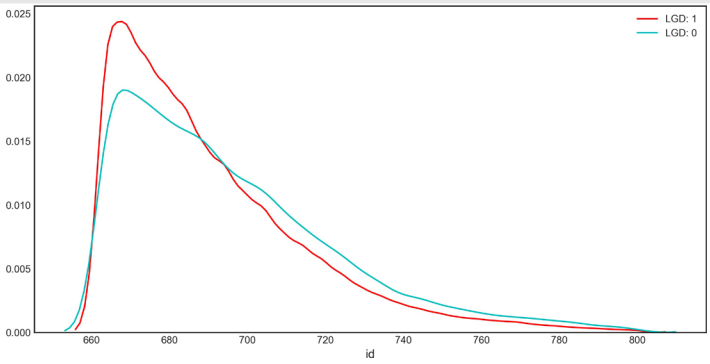
由上可见，相比单个的逻辑模型、线性判别分析模型或随机森林模型，投票分类器这种集成模型在测试集的表现是最好的。

结论

LGD的预测其实是很难的，因为各种难以量化的因素的影响，导致LGD是很难用模型去刻画的，该项目试图通过机器学习的方式来对LGD进行建模与预测，但限于数据质量的问题：比如目标变量LGD应该为一个连续变量，此　处给出的是分类变量，这就将建模预测的难度大大提升。不过，基于对该项目的分析，也可得出一些具有实际意义的结论：

1、用户在申请LC贷款时，其LC评定的贷款等级（grade）能给预测LGD提供较大的较为有效的信息量。由上面grade的分布图可看出，等级最高的用户反而LGD为1的概率最高，这可能是用户获取贷款后受到银行资金使用限制较为宽松的原因导致的。

2、用户的FICO评分同样能为预测LGD提供较大的信息量，FICO分数较低的用户，LGD为1的可能性会更大。


LGD: 1
LGD: 0
0.025
0.020
0.015
0.010
0.005
0.000
660
680
700
720
740
760
780
800
id