**申请评分卡开发（改进）**

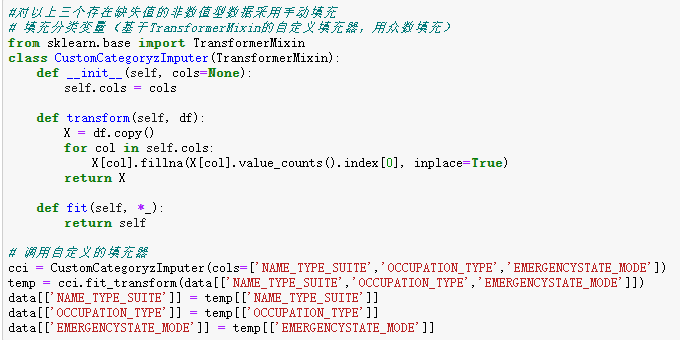
**K近邻、逻辑回归、集成学习、梯度提升决策树、随机森林**

**成员：陈广威 梁锦伦**

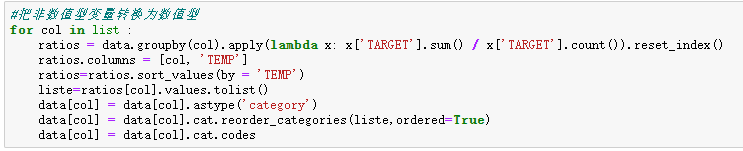
本报告为基于逻辑回归的申请评分卡模型的改进尝试，使用的数据为application\_data.经过逻辑回归模型数据处理后的application\_data\_cleaning数据，采用了决策树模型，以对申请评分卡进行另类模型的尝试。

一．对存在样本缺失的非数值型特征使用“众数”填充

非数值型特征有：NAME\_CONTRACT\_TYPE、CODE\_GENDER、FLAG\_OWN\_CAR、FLAG\_OWN\_REALTY、NAME\_TYPE\_SUITE、NAME\_INCOME\_TYPE、NAME\_EDUCATION\_TYPE、NAME\_FAMILY\_STATUS、NAME\_HOUSING\_TYPE、OCCUPATION\_TYPE、WEEKDAY\_APPR\_PROCESS\_START、ORGANIZATION\_TYPE、EMERGENCYSTATE\_MODE，其中NAME\_TYPE\_SUITE、OCCUPATION\_TYPE和EMERGENCYSTATE\_MODE存在缺失。编写自定义非数值型特征“众数”填充器并对其进行填充：

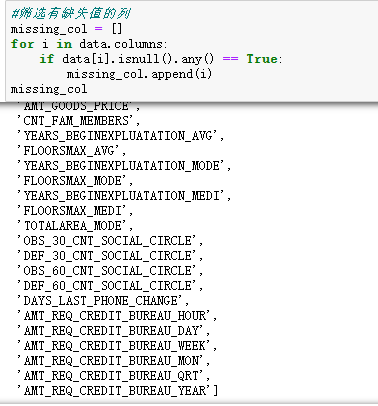


填充之后把非数值型特征转变为数值型特征：使用自定义函数将其中的分类型取值进行数字转变。

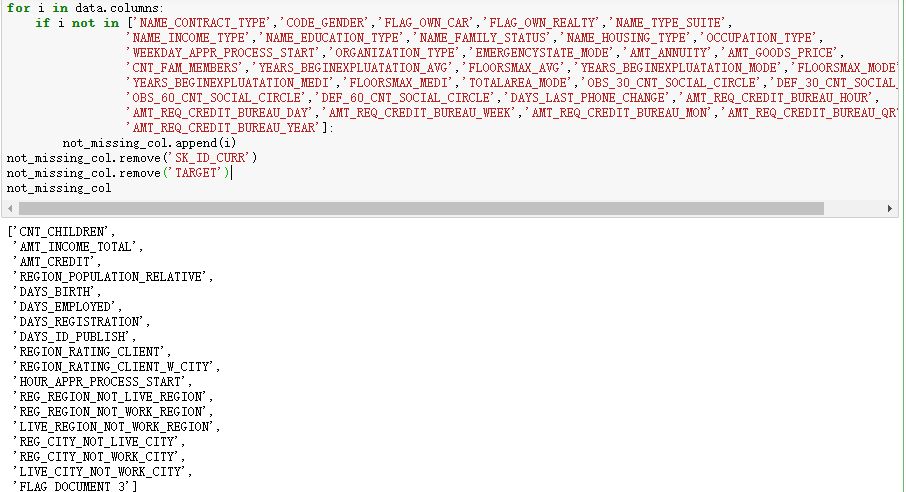


二．使用随机森林的填充数值型特征缺失值

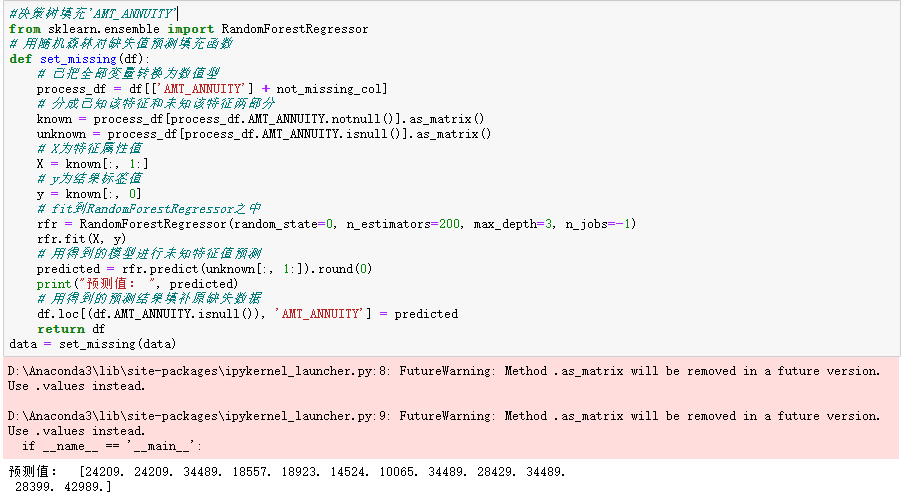
筛选出存在缺失值的特征：AMT\_ANNUITY、AMT\_GOODS\_PRICE、CNT\_FAM\_MEMBERS、YEARS\_BEGINEXPLUATATION\_AVG、FLOORSMAX\_AVG、YEARS\_BEGINEXPLUATATION\_MODE、FLOORSMAX\_MODE、YEARS\_BEGINEXPLUATATION\_MEDI、FLOORSMAX\_MEDI、TOTALAREA\_MODE、OBS\_30\_CNT\_SOCIAL\_CIRCLE、DEF\_30\_CNT\_SOCIAL\_CIRCLE、OBS\_60\_CNT\_SOCIAL\_CIRCLE、DEF\_60\_CNT\_SOCIAL\_CIRCLE、DAYS\_LAST\_PHONE\_CHANGE、AMT\_REQ\_CREDIT\_BUREAU\_HOUR、AMT\_REQ\_CREDIT\_BUREAU\_DAY、AMT\_REQ\_CREDIT\_BUREAU\_WEEK、AMT\_REQ\_CREDIT\_BUREAU\_MON、AMT\_REQ\_CREDIT\_BUREAU\_QRT、AMT\_REQ\_CREDIT\_BUREAU\_YEAR。



筛选出未缺失的特征：CNT\_CHILDREN、AMT\_INCOME\_TOTAL、AMT\_CREDIT、REGION\_POPULATION\_RELATIVE、DAYS\_BIRTH、DAYS\_EMPLOYED、DAYS\_REGISTRATION、DAYS\_ID\_PUBLISH、REGION\_RATING\_CLIENT、REGION\_RATING\_CLIENT\_W\_CITY、HOUR\_APPR\_PROCESS\_START、REG\_REGION\_NOT\_LIVE\_REGION、REG\_REGION\_NOT\_WORK\_REGION、LIVE\_REGION\_NOT\_WORK\_REGION、REG\_CITY\_NOT\_LIVE\_CITY、REG\_CITY\_NOT\_WORK\_CITY、LIVE\_CITY\_NOT\_WORK\_CITY、FLAG\_DOCUMENT\_3。

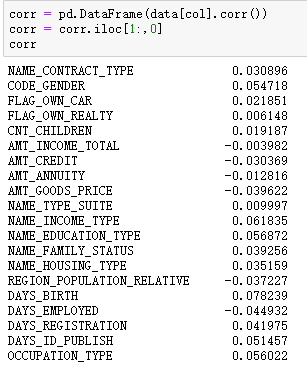


由之前逻辑回归的数据处理过程可以知道：特征之间存在一定的关联。决策树填充可以考虑到特征之间的关联关系。因此使用决策树根据特征之间的关联填充缺失值。依次对存在缺失的特征进行填充，以“AMT\_ANNUITY”为例：

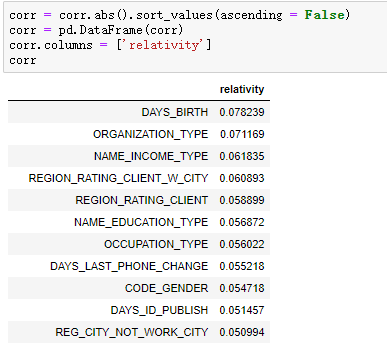


三．根据相关性选择模型变量

移除ID特征之后，计算其余特征与TARGET之间的相关性：



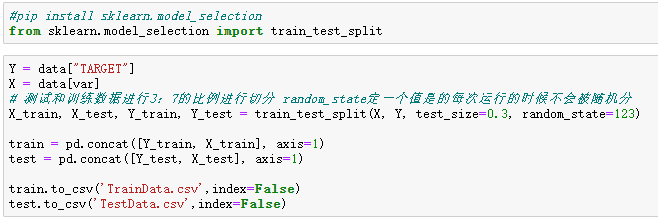
将相关性的绝对值按降序排列：



根据所有特征与TARGET之间的相关性大小判断，选择与TARGET相关性大于0.05的特征初步作为模型拟合特征。

四．数据切分

将307511条样本数据按3：7的比例分为测试集和训练集：

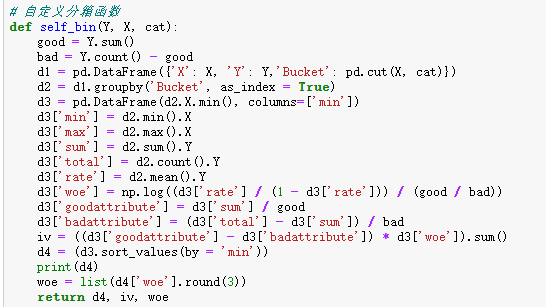


五．特征分箱

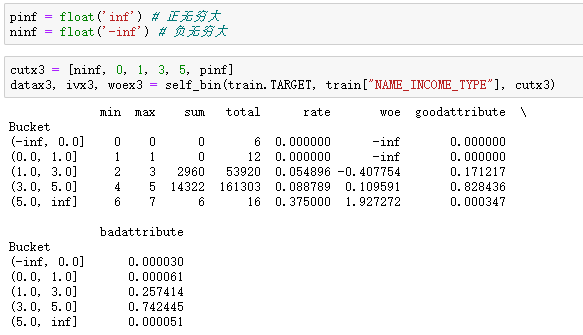
定义对特征自动分箱的函数（返回值含IV、WOE）：



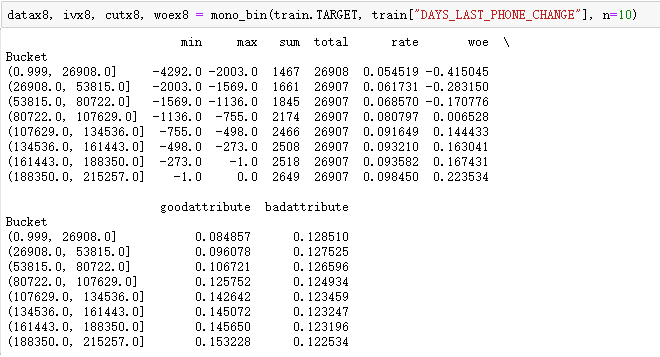
定义自定义特征分箱函数（返回值含IV、WOE）：



对第三步骤中初步选择的11个模型特征使用mono\_bin进行自动分箱或者使用self\_bin进行自定义分箱。以NAME\_INCOME\_TYPE使用self\_bin和DAYS\_LAST\_PHONE\_CHANGE使用mono\_bin为例：



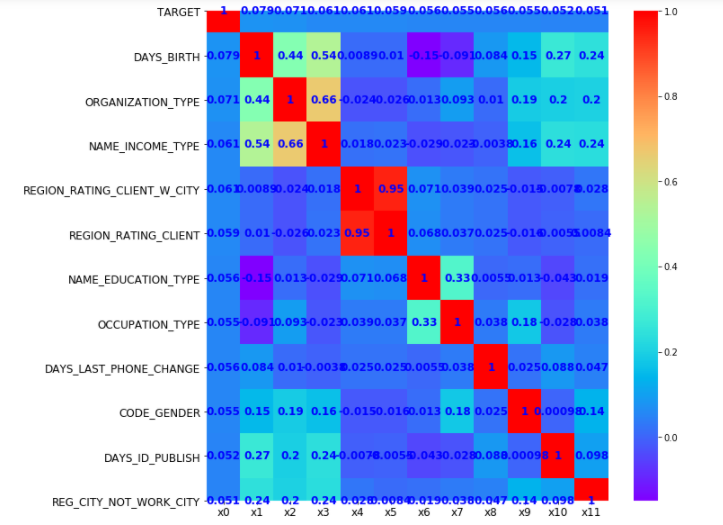
NAME\_INCOME\_TYPE分为5个箱子，，输出中包含每个箱子的最小值、最大值、和、计数、WOE值等。



DAYS\_LAST\_PHONE\_CHANGE分为了8个箱子，输出中包含每个箱子的最小值、最大值、和、计数、WOE值等。

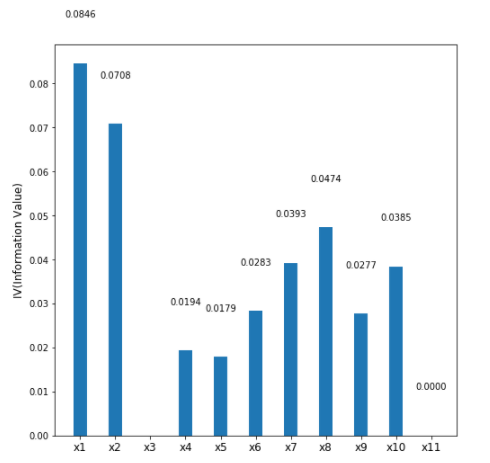
六．相关性分析和IV筛选

分析第三步中初步选择的11个特征之间的相关性，通过相关性热图的形式可视化：



变量之间的相关性都较小，但是 REGION\_RATING\_CLIENT\_W\_CITY 和 REGION\_RATING\_CLIENT 相对来说较大为0.95。

生成11个特征的IV值柱状图：

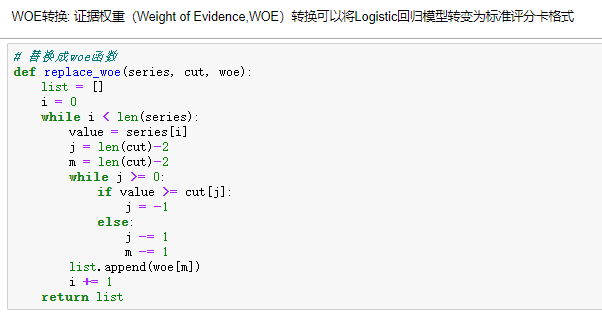


通过IV值判断变量预测能力的标准是：小于0.02：unpredictive，0.02 to 0.1：weak，0.1 to 0.3：medium，0.3 to 0.5：strong，大于0.5：suspicious。

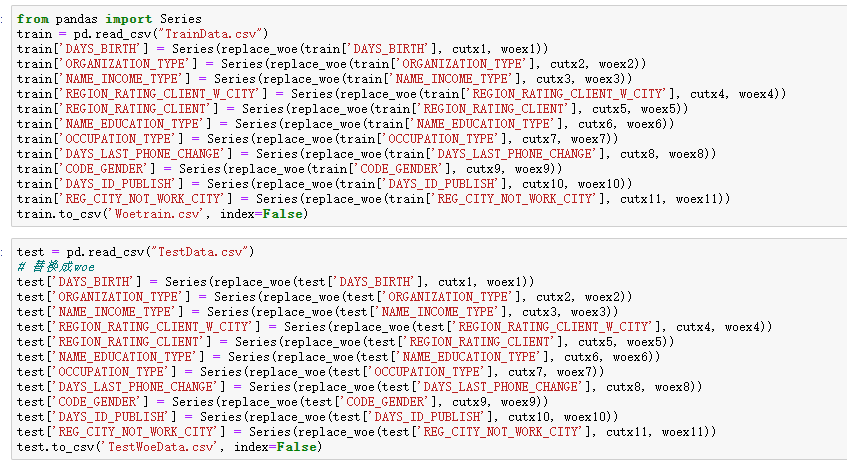
从IV值柱状图可知：NAME\_INCOME\_TYPE、REGION\_RATING\_CLIENT\_W\_CITY、REGION\_RATING\_CLIENT、REG\_CITY\_NOT\_WORK\_CITY变量的IV值明显较低。

七．模型分析

自定义编写WOE转换函数：

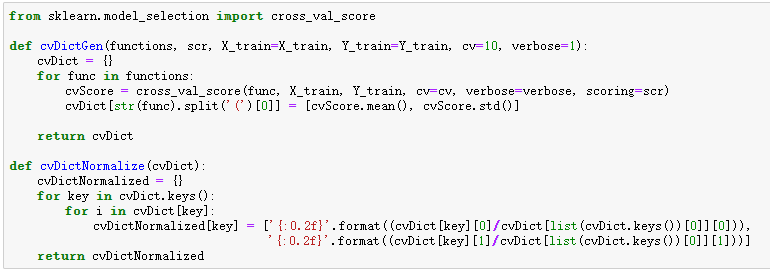


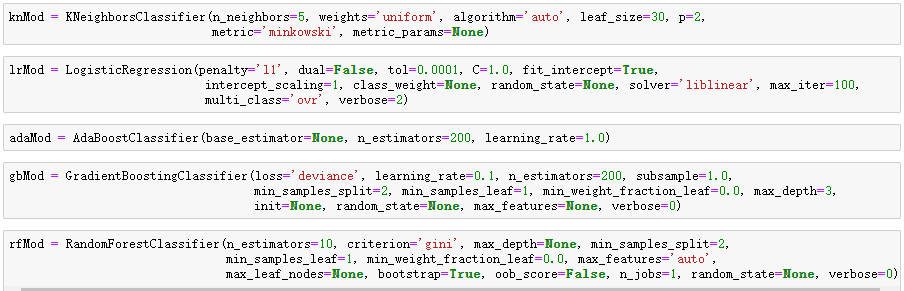
将初步选择的11个特征的训练集和测试集进行WOE转换：

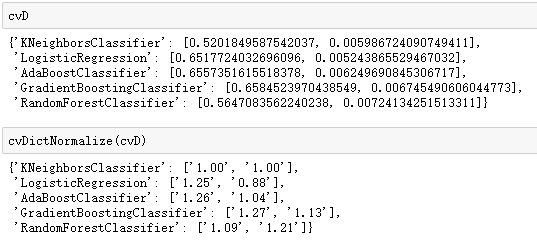


八．模型预测

通过K近邻算法、逻辑回归、集成学习AdaBoostClassifier分类、梯度提升决策树GradientBoostingClassifier、随机森林分类RandomForestClassifier进行模型预测。







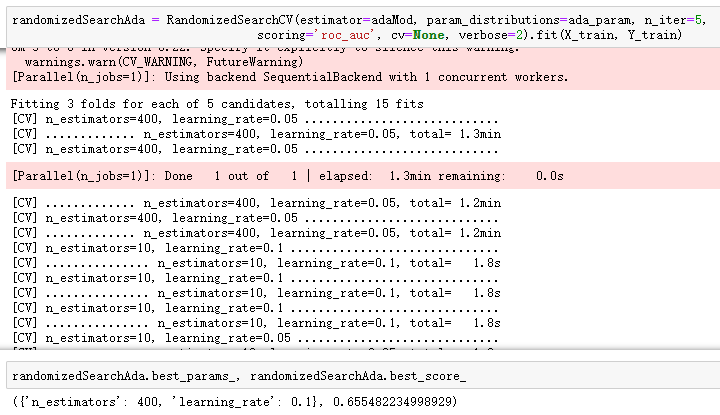
根据模型预测结果可知：LogisticRegression的out-of-the-bag较好。

九．最优化超参数及模型拟合

使用sklearn.model\_selection中的RandomizedSearchCV最优化AdaBoost超参数和GB模型超参数。

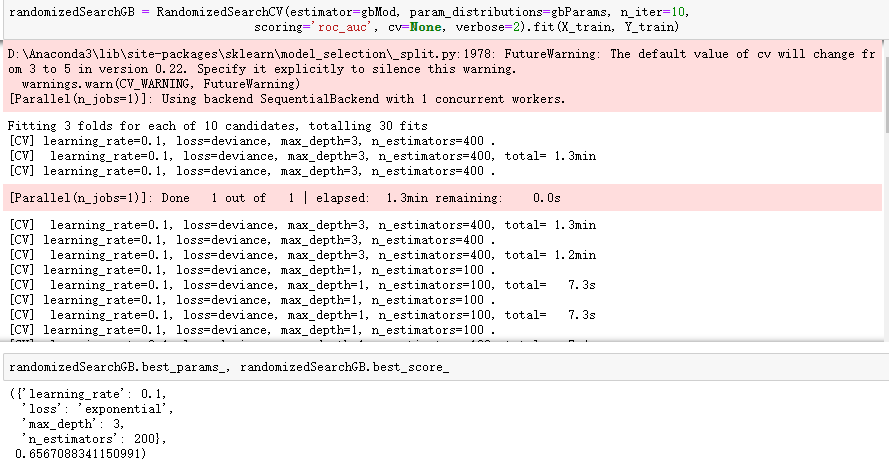
AdaBoost最优超参数：

n\_estimators：400；learning\_rate：0.1。

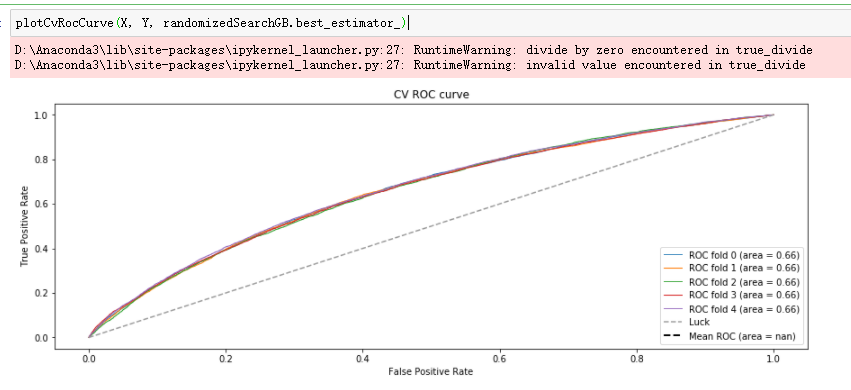


GB模型最优超参数：

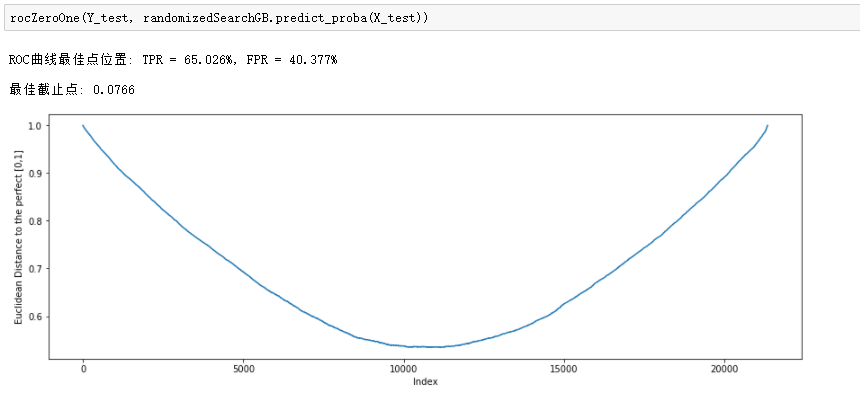
learning\_rate：0.1；loss：exponential；max\_depth：3；n\_estimators：200。



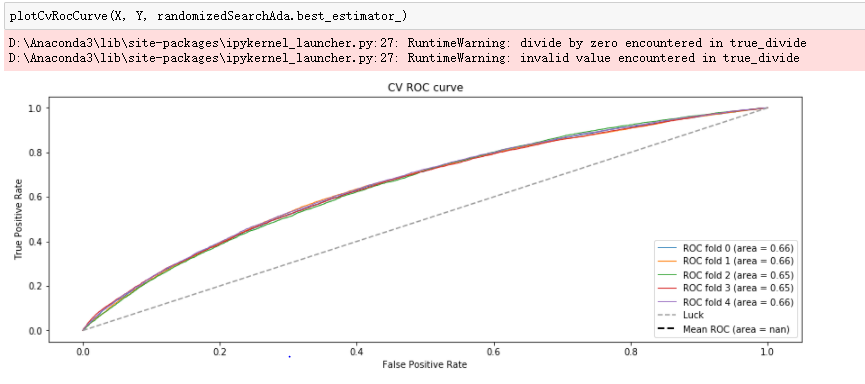
拟合最优超参数下GB梯度提升决策树模型，画出其ROC曲线：



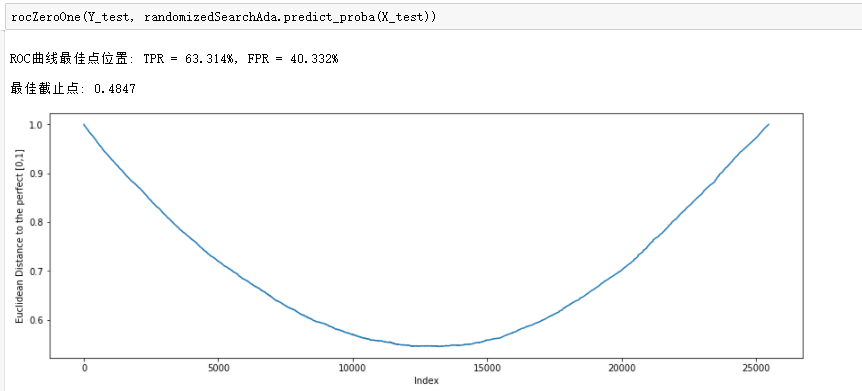
同时得出GB模型ROC曲线最佳截止点：0.0766，最佳点位置：TPR=65.026%，FPR=40.377%。



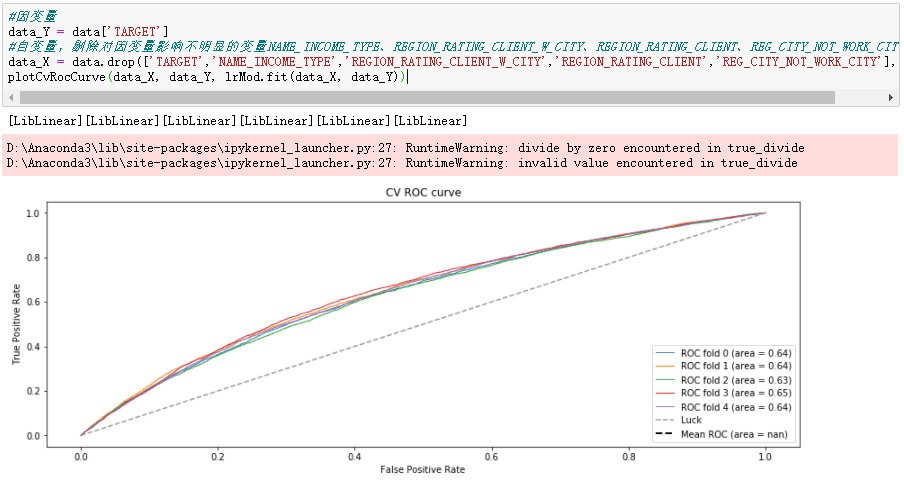
拟合最优超参数下AdaBoost集成学习模型，画出其ROC曲线：



Adaboost模型ROC曲线最佳截止点：0.4847，最佳点位置：TPR=63.314%，FPR=40.332%。

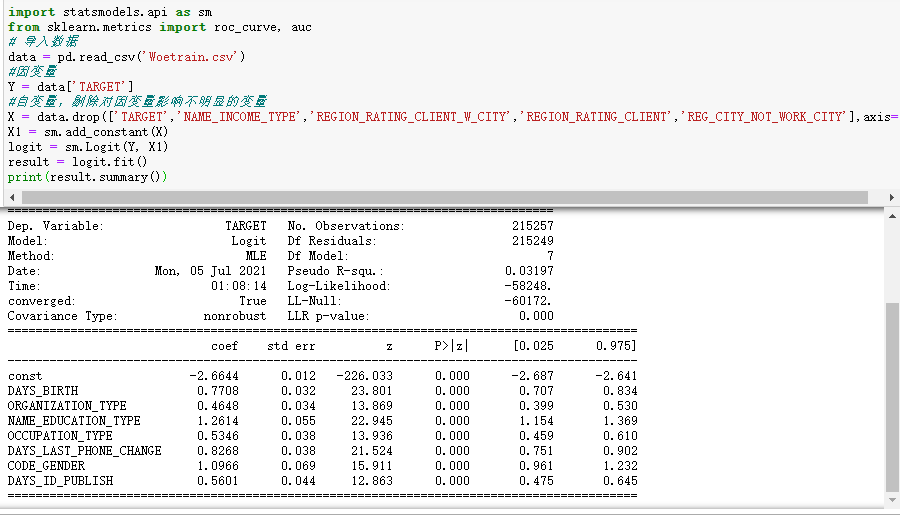


在初步选定的11个特征中，剔除对因变量TARGET影响不明显的变量NAME\_INCOME\_TYPE、REGION\_RATING\_CLIENT\_W\_CITY、REGION\_RATING\_CLIENT、REG\_CITY\_NOT\_WORK\_CITY，通过剩下7个特征对TARGET拟合LR逻辑回归模型，画出其ROC曲线：



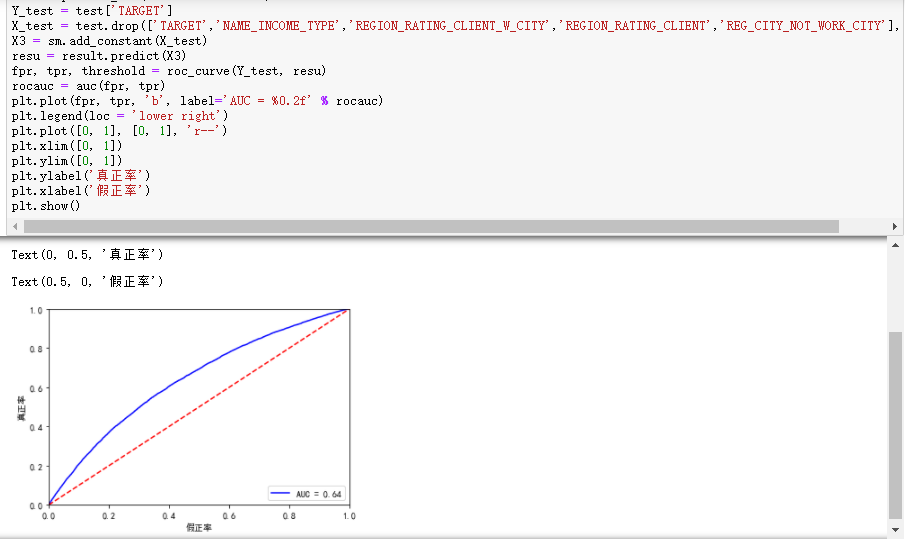
LR逻辑回归模型ROC曲线最佳截止点：0.0815，最佳点位置：TPR=60.793%，FPR=40.245%。

使用WOE转换后的训练集数据，剔除对因变量TARGET影响不明显的变量NAME\_INCOME\_TYPE、REGION\_RATING\_CLIENT\_W\_CITY、REGION\_RATING\_CLIENT、REG\_CITY\_NOT\_WORK\_CITY，通过剩下7个特征对TARGET拟合LR逻辑回归：

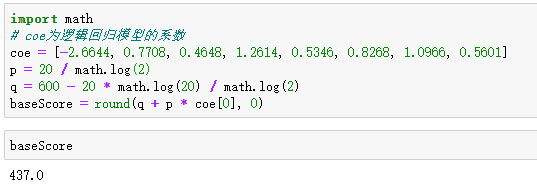


逻辑回归常数项为-2.6644，DAYS\_BIRTH系数为0.7708，ORGANIZATION\_TYPE系数为0.4648，NAME\_EDUCATION\_TYPE系数为1.2614，OCCUPATION\_TYPE系数为0.5346，DAYS\_LAST\_PHONE\_CHANGE系数为0.8268，CODE\_GENDER系数为1.0966，DAYS\_ID\_PUBLISH系数为0.5601。在5%的置信度即95%置信区间，系数是显著的。

使用WOE转换后的测试集数据，剔除对因变量TARGET影响不明显的变量NAME\_INCOME\_TYPE、REGION\_RATING\_CLIENT\_W\_CITY、REGION\_RATING\_CLIENT、REG\_CITY\_NOT\_WORK\_CITY，通过剩下7个特征对TARGET进行LR逻辑回归预测：



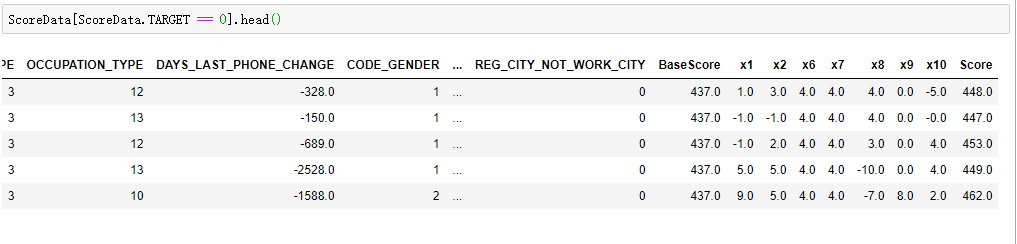
使用逻辑回归计算基础分：



使用测试集数据计算评分，并输出保存：



查看评分数据中TARGET=0即不违约的评分：



上图显示不违约的申请信用评分的前五行，评分均较高。