**用户申请评分卡开发**

成员：陈广威、梁锦伦

## 项目背景：

评分卡模型又叫做信用评分卡模型，最早由美国信用评分巨头FICO公司于20世纪60年代推出，在信用风险评估以及金融风险控制领域中广泛使用。银行利用评分卡模型对客户的信用历史数据的多个特征进行打分，得到不同等级的信用评分，从而判断客户的优质程度，据此决定是否准予授信以及授信的额度和利率。相较资深从业人员依靠自身的经验设置的专家规则，评分卡模型的使用具有很明显的优点：

1. 判断快速：系统只需要按照评分卡逐项打分，最后通过相应的公式计算出总分，即可准确判断出是否为客户授信以及额度和利率。
2. 客观透明：评分卡模型的标准是统一的，无论是客户还是风险审核人员，都可以通过评分卡一眼看出评分结果和评判依据。
3. 应用范围广：由于评分卡的评分项是客观计算，其得出的分数具有广泛的参考性和适用性。例如，生活中常见的支付宝芝麻信用分，就是依据评分卡模型计算得出。

评分卡模型在银行不同的业务阶段体现的方式和功能也不一样。按照借贷用户的借贷时间，评分卡模型可以划分为以下三种：

1. 贷前：申请评分卡（Application score card），又称为A卡。使用申请评分卡可以更准确地评估申请人的未来表现(违约率)，降低坏帐率，加快(自动化)审批流程, 降低营运成本，还能增加审批决策的客观性和一致性，提高客户满意度
2. 贷中：行为评分卡（Behavior score card），又称为B卡。使用行为评分卡可以建立更好的客户管理策略, 提高赢利，减少好客户的流失，还能对可能拖欠的客户做出预警。
3. 贷后：催收评分卡（Collection score card），又称为C卡。使用催收评分卡可以优化催收策略，提高欠帐的回收率，减少不必要的催收行为，降低营运成本。

## 项目流程：

信用风险评级模型的主要开发流程如下：

（1） 数据获取，包括获取存量客户及潜在客户的数据。存量客户是指已经在证券公司开展相关融资类业务的客户，包括个人客户和机构客户；潜在客户是指未来拟在证券公司开展相关融资类业务的客户，主要包括机构客户，这也是解决证券业样本较少的常用方法，这些潜在机构客户包括上市公司、公开发行债券的发债主体、新三板上市公司、区域股权交易中心挂牌公司、非标融资机构等。

（2） 数据预处理，主要工作包括数据清洗、缺失值处理、异常值处理，主要是为了将获取的原始数据转化为可用作模型开发的格式化数据。

（3） 探索性数据分析，该步骤主要是获取样本总体的大概情况，描述样本总体情况的指标主要有直方图、箱形图等。

（4） 变量选择，该步骤主要是通过统计学的方法，筛选出对违约状态影响最显著的指标。主要有单变量特征选择方法和基于机器学习模型的方法 。

（5） 模型开发，该步骤主要包括变量分段、变量的WOE（证据权重）变换和逻辑回归估算三部分。

（6） 模型评估，该步骤主要是评估模型的区分能力、预测能力、稳定性，并形成模型评估报告，得出模型是否可以使用的结论。

（7） 信用评分，根据逻辑回归的系数和WOE等确定信用评分的方法。将Logistic模型转换为标准评分的形式。

（8） 建立评分系统，根据信用评分方法，建立自动信用评分系统。

## 数据获取：

面向的客户群体：

（1） 存量客户数据（已开展相关融资类业务的客户，包括个人客户和机构客户，主要为个人客户）。

（2） 潜在客户数据（未来拟开展相关融资类业务的客户）。

数据集描述：

（1） 名称：application\_data.csv

（2） 来源：Kaggle上一个经典的评分卡案例（Loan Defaulter）

（3） 网址：<https://www.kaggle.com/gauravduttakiit/loan-defaulter>

（4） 内容：

基本属性：包括了借款人性别、职业、贷款类型等属性；

家庭情况：包括了借款人拥有的孩子数、家庭成员数等属性；

偿债能力：包括了借款人的收入、是否拥有房子、公寓、车等财产等属性；

联系方式：借款人是否提供电话、邮箱等属性；

信用往来：30天逾期次数、60天逾期次数等属性；

财产状况：包括客户居住建筑的标准化信息等属性；

贷款属性：包括客户是否提供某文件等属性；

其他因素：包括客户申请前一小时、一天等时间内向征信局查询客户的次数。

样本数量：307511条

特征数量：122个

## 数据预处理：

大数据时代，要进行分析的数据量迅速膨胀，同时由于各种原因数据通常是不完整的（缺少某些感兴趣的属性值）、不一致的（包含代码或者名称的差异）、极易受到噪声（错误或异常值）的侵扰的。因为数据库太大，而且数据集经常来自多个异种数据源，低质量的数据将导致低质量的分析结果，所以分析数据所涉及的数据对象必须进行预处理。

（1） 目的：为了将获取的原始数据转化为可用作模型开发的格式化数据。

（2） 过程：特征清洗、缺失值处理、异常值处理

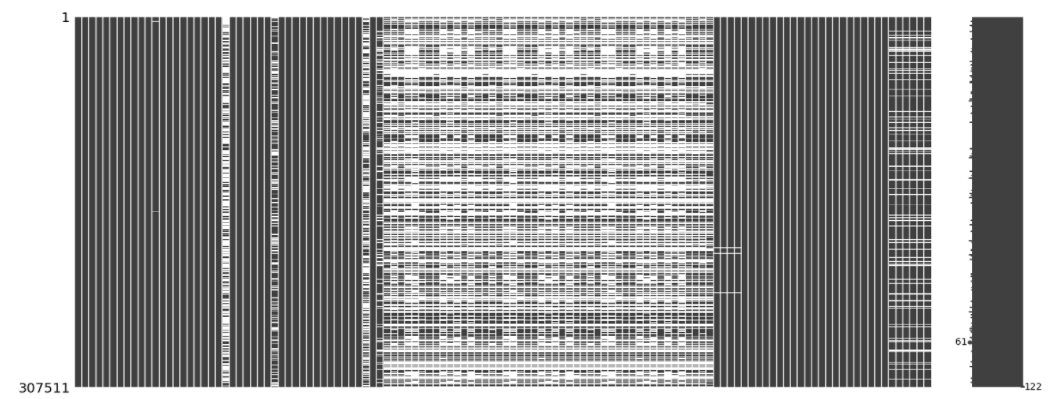
（A） 特征清洗：特征工程Dropping

（a） 统计特征的样本缺失率，若缺失值样本占总数比例极高（阈值），直接舍弃该特征，因为作为特征加入数据反而会引入噪声值。

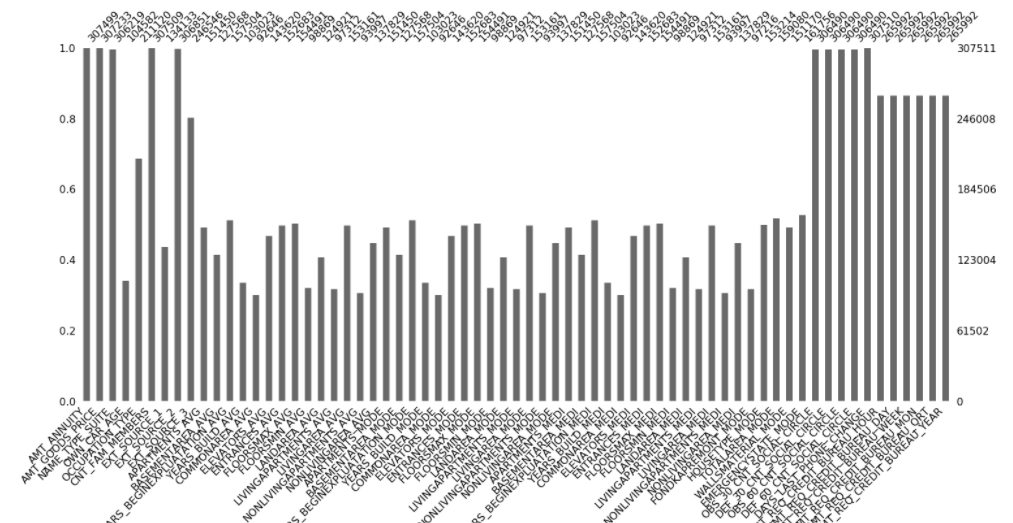
技术：缺失值可视化（missingno）、无效矩阵的数据密集显示（matrix）、无效数据简单可视化的条形图（bar）、缺失性相关性热图（heatmap）、树状图（dendrogram）。

举例：

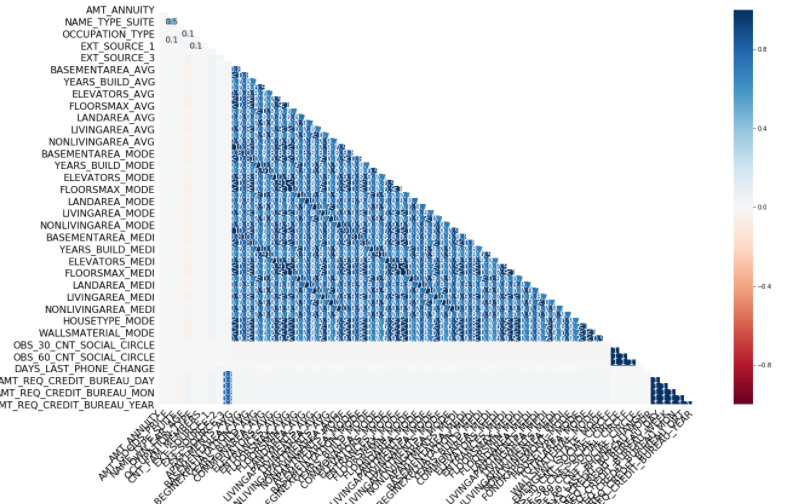
整体数据缺失值的无效矩阵数据密集显示（白色代表缺失）：



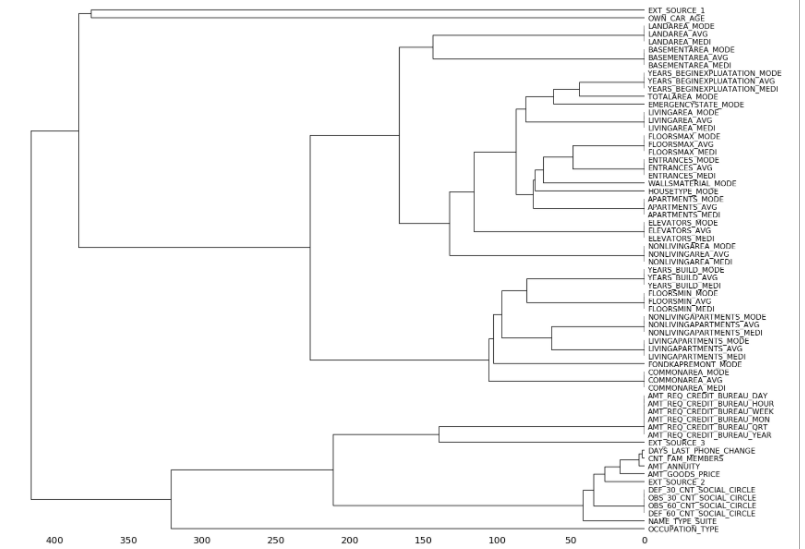
特征无效数据的条形图：



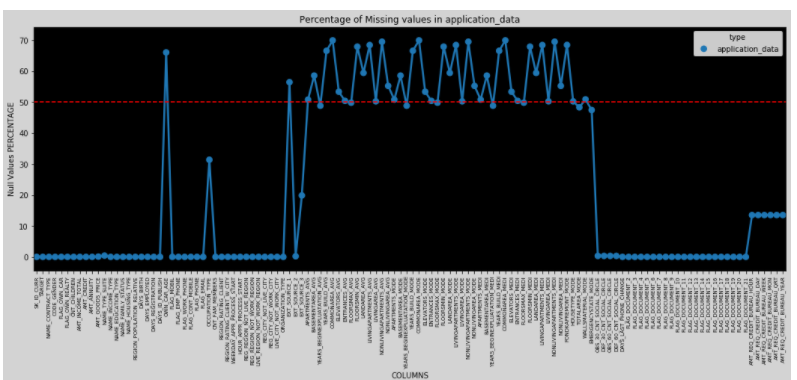
缺失性相关性热图：



树状图：



删除特征变量样本缺失率高于50%的特征：

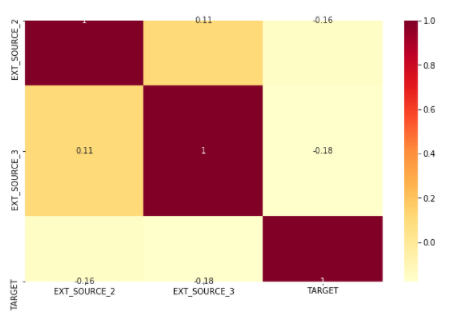


（b） 通过目标变量与自变量特征之间的相关性，结合数据的分箱统计和逻辑分析，对无关特征进行删除。

技术：数据相关性热图（heatmap）、分箱统计条形图（countplot）。

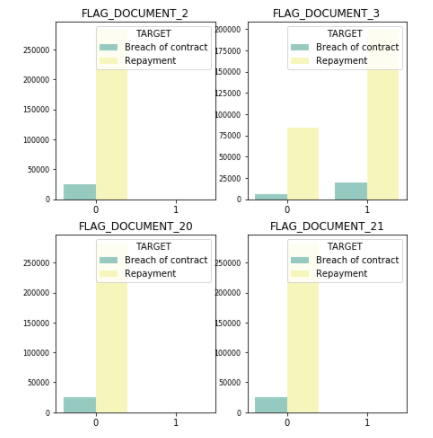
举例：

EXT\_SOURCE\_2、EXT\_SOURCE\_3与TARGET的相关性热图：



EXT\_SOURCE\_2、EXT\_SOURCE\_3与TARGET间不具相关性。

FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_3、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21与TARGET的分箱统计条形图：



FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21变量下，用户的违约与还款均发生在变量取值为0时，分析FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21变量无法得出其取值为1时对目标变量的影响。

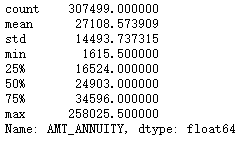
（B） 缺失值处理：特征工程Imputation、An extension to imputation

（a） 连续性特征、二值特征根据同特征样本之间的相似性，利用平均数、众数、中位数填充填补缺失值。

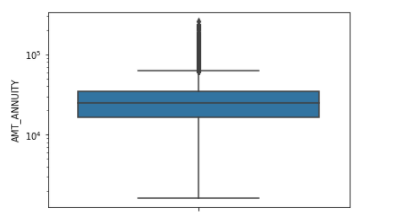
技术：描述性统计（describe）、箱线图（boxplot）

举例：

AMT\_ANNUITY的描述性统计：



AMT\_ANNUITY的箱线图：



AMT\_ANNUITY存在较多异常值超出上限，因此使用中位数填充缺失值。

（b） 非连续性特征（分类值）将属性值进行分类，作为新的特征加入到数据集。

技术：特征工程独热编码（one-hot encoding）

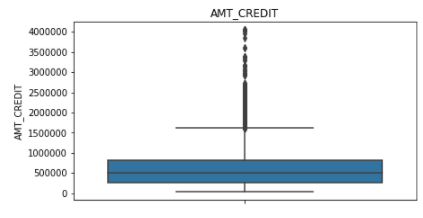
（C） 异常值处理

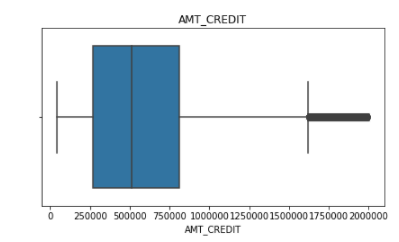
对连续性特征由于人为错误、数据处理错误等原因产生的异常值进行删除处理。

技术：箱线图（boxplot）、直方图（distplot）、核密度估计（kde）

举例：

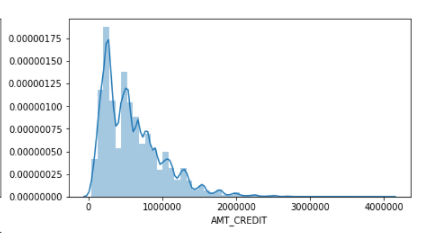
AMT\_CREDIT的箱线图：





对AMT\_CREDIT超过2000000的数据进行删除。

AMT\_CREDIT的直方图和核密度估计：



## 探索性分析：

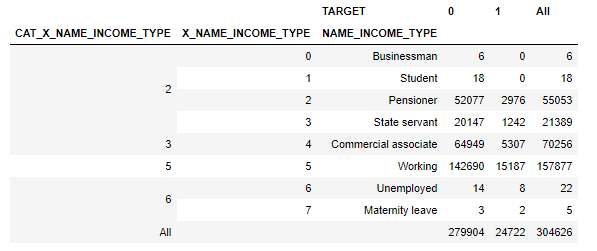
（1） 数据分箱：局部平滑方法，通过考察“邻居”（周围的值）来平滑存储数据的值，即连续变量离散化，达到去噪，增加粒度，提高变量的可解释性和鲁棒性的目的。

方法：决策树最优分箱（使用sklearn提供的决策树，将内部节点的阈值作为分箱的切点，利用变量值x和目标变量y之间产生一个映射关系，从而将特征用树模型拟合目标变量）

意义：数据分箱要求变量分箱后每箱的坏样本占比具有一定的区分度。决策树分箱通过信息增益或者gini系数进行分裂特征，信息增益或者gini系数也可以说成分裂点最大可能的提升分裂前后变量的区分度，所以分箱的每个分裂点都是根据目标变量选择的最优分裂方式。

举例：

列联表显示部分分箱结果：



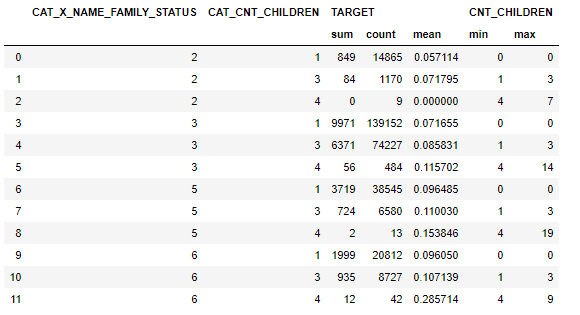
（2） 特征分析和特征衍生

（A） 特征分析

对一类特征变量（如工作/收入相关的特征变量）及其分箱结果与目标变量TARGET之间进行描述性统计，从而初步分析变量与目标变量之间是否存在相关性影响。

举例：

CAT\_X\_NAME\_FAMILY\_STATUS、CAT\_CNT\_CHILDREN分箱结果与特征变量CNT\_CHILDREN和目标变量TARGET的描述性统计：



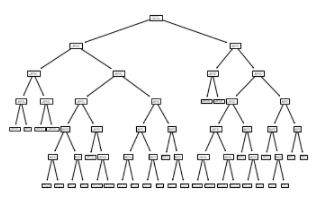
（B） 特征衍生

根据特征分析结果，采用技术衍生的手段利用现有的特征进行某种组合衍生成新的特征，再结合特征分析判断衍生的特征对目标变量是否存在较强的影响。

本项目主要采用多个特征同时对目标变量拟合的决策树分箱方式衍生出新特征。

举例：

CNT\_FAM\_MEMBERS、CAT\_X\_NAME\_FAMILY\_STATUS和CNT\_CHILDREN对目标变量TARGET拟合的决策树分箱衍生CAT\_NAME\_FAMILY\_STATUS\_MEMBERS：

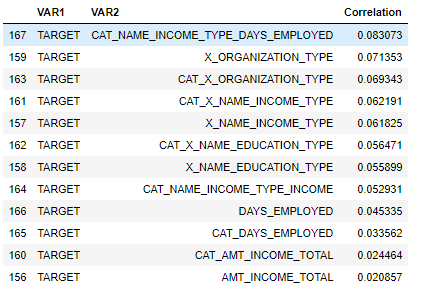


（3） 相关性分析

使用相关性矩阵对特征变量与目标变量进行分析，从而衡量该特征变量与目标变量的相关密切程度。

举例：

特征变量、衍生的特征变量与目标变量的相关性（部分）：



## 变量选择：

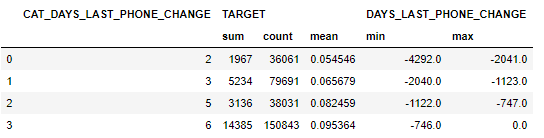
通过统计学的方法，筛选出对目标变量即违约与否影响最显著的指标。

（1） 特征分析

通过特征变量与目标变量之间的描述性统计，人为分析特征变量与目标变量之间的关联性，判断该特征是否对目标变量有影响，作为变量的初步筛选。

举例：

CAT\_DAYS\_LAST\_PHONE\_CHANGE与TARGET、DAYS\_LAST\_PHONE\_CHANGE之间的描述性统计：



最佳更换电话号码越频繁，最终违约概率越大，因此特征与目标变量之间相关。

（2） 双样本比例Z检验

对特征分析中描述统计为两种情况的特征变量进行对目标变量的比例检验，分析该特征变量对目标变量是否存在显著影响。

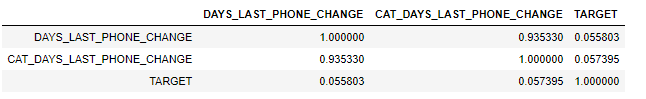
技术：statsmodels.stats.proportion.proportions\_ztest

（3） 相关性分析

使用相关性矩阵得出特征变量之间、特征变量与目标变量之间的相关性，分析特征变量与目标变量之间的相关性，选择与目标变量相关性高的特征变量作为已选择特征；分析已选择特征变量之间的相关性，对其中彼此相关性高的两个或多个变量进行选择剔除，尽量确保已选择的变量之间具有低相关性，减少变量间的关联，避免变量重复、多重共线性问题和过拟合问题。

举例：

DAYS\_LAST\_PHONE\_CHANGE、CAT\_DAYS\_LAST\_PHONE\_CHANGE与TARGET之间的相关性：



相关性分析中，TARGET与DAYS\_LAST\_PHONE\_CHANGE、CAT\_DAYS\_LAST\_PHONE\_CHANGE相关性不高，但从描述统计中可以发现，彼此间存在联系，由于DAYS\_LAST\_PHONE\_CHANGE和CAT\_DAYS\_LAST\_PHONE\_CHANGE高度相关，保留和TARGET变量更相关的CAT\_DAYS\_LAST\_PHONE\_CHANGE。

（4） 变量的评价

（A） WOE（Weight of Evidence）：证据权重类似于信息熵反映了自变量取值对目标变量的影响。WOE越大表示该特征正向作用越大，它对数据进行了归一化处理，也就是将所有不同特征划在了统一的尺度上。公式：WOE=ln(good attribute/bad attribute)。

（B） IV（Information Value）：WOE没有考虑分组中样本占整体样本的比例，如果一个分组的WOE值很高，但是样本数占整体样本数很低，则对变量整体预测的能力会下降。IV信息量考虑了分组中样本占整体样本的比例，相当于WOE的加权求和。计算特征内部的信息含量，如果信息含量足够大，则表示其为有价值的特征。

（一般）标准：

小于0.02：预测能力极低（Unpredictive）

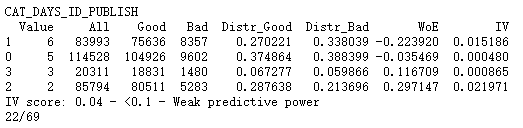
0.02到0.1之间：预测能力较低（Weak）

0.1到0.3之间：预测能力中等（Medium

0.3到0.5之间：预测能力较强（Strong）

举例：

CAT\_DAYS\_ID\_PUBLISH的WOE和IV值：

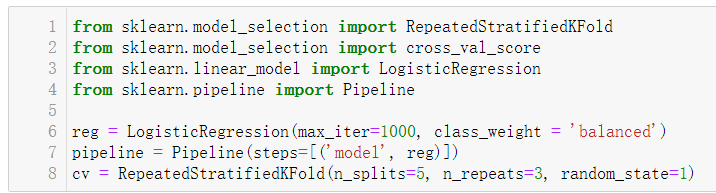


**模型开发：**

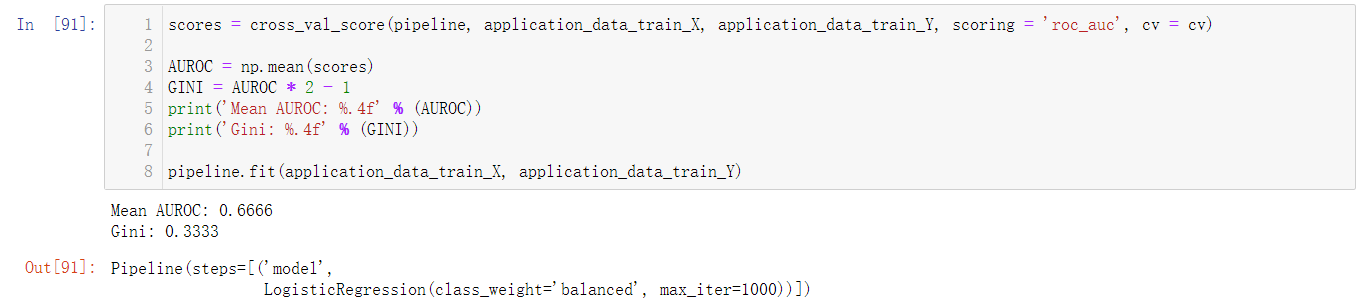
在信用评分卡建模中，可以使用很多方法，但基于信用评分卡的实际应用背景，很多方法建立的信用评分卡模型很难简化成一种评分表。为了简单性和可解释性。我们还是主要选用了应用广泛的Logistic回归模型。

Logistic回归模型：

逻辑回归是一种用于解决二分类问题的机器学习方法，用于估计某种事物的可能性。这里的“可能性”不等同与数学上的“概率”，它的结果往往用与和其他特征值加权求和。所以运用逻辑回归模型，可以给我们选择的特征赋予权重，再对特征加权求和得到评分。



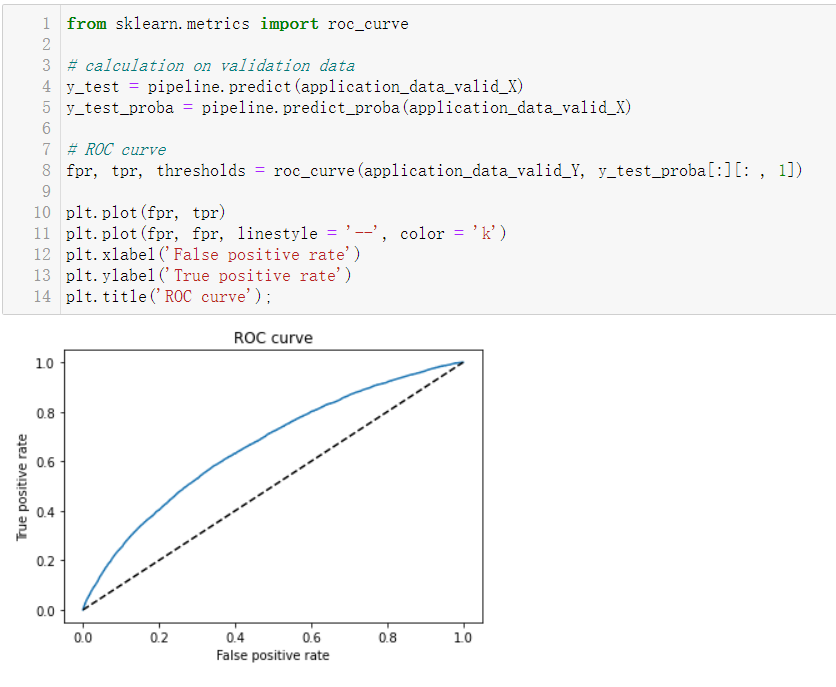
除了使用逻辑回归模型，我们还使用了Pipeline封装了训练模型的步骤，可以使参数和步骤可以重复地在新数据集上使用，也可以结合grid search对参数进行选择。使用了分层的K折交叉校验，将原始数据D随机分成K份，每次选择（K-1）份作为训练集，剩余的1份（红色部分）作为测试集。交叉验证重复K次，取K次准确率的平均值作为最终模型的评价指标，这样可以有效避免过拟合和欠拟合状态的发生。



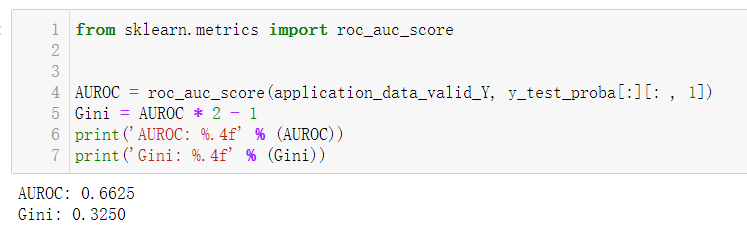
对模型进行交叉检验后，输出评估分数。模型的Mean AUROC=0.6666，而Gini系数为0.3333，总体来看，模型的表现还是不错的。

**模型评估：**

建立完模型后，使用验证集对模型的预测能力进行评估。在评估过程中主要使用ROC曲线和AUC指标。



ROC曲线的横坐标为假阳率，纵坐标为真阳率，当曲线下面的面积越大时，模型的准确率越高，从图形上看，我们模型的ROC曲线虽然超过了中间的分割线，但并没有特别贴近左上角，可能模型的效果不是特别的好。



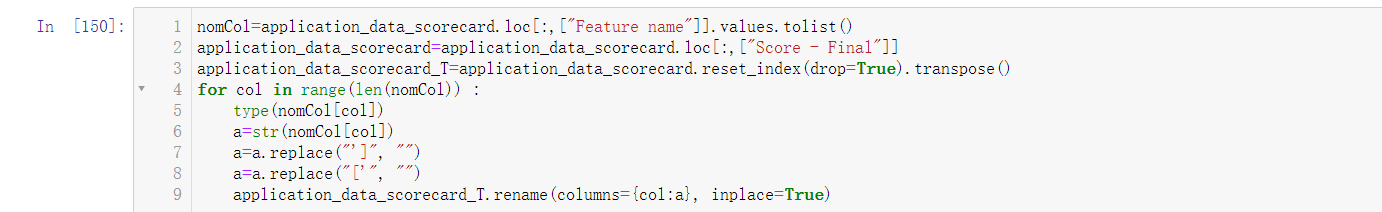
AUROC表示接受者操作特征曲线下面积，该值越大，表示模型的预测效果越好。GINI系数定义为ROC曲线和对角线AC之间的面积占对角线AC曲线围成面积比，该值越大，模型的区分能力越强。通过计算，验证集的AUROC=0.6625，而GINI=0.325，这说明该模型的预测效果有一定的应用价值。

**信用评分：**

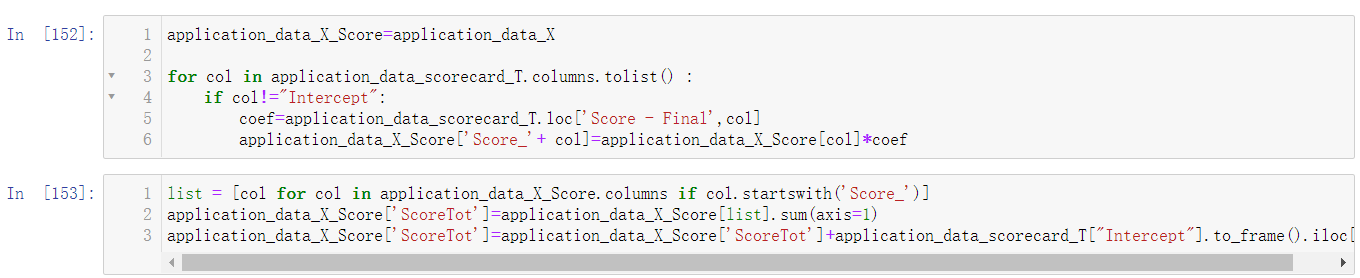
在完成建模相关的工作后，接下的任务就是把Logistic模型转换成标准评分卡的形式。



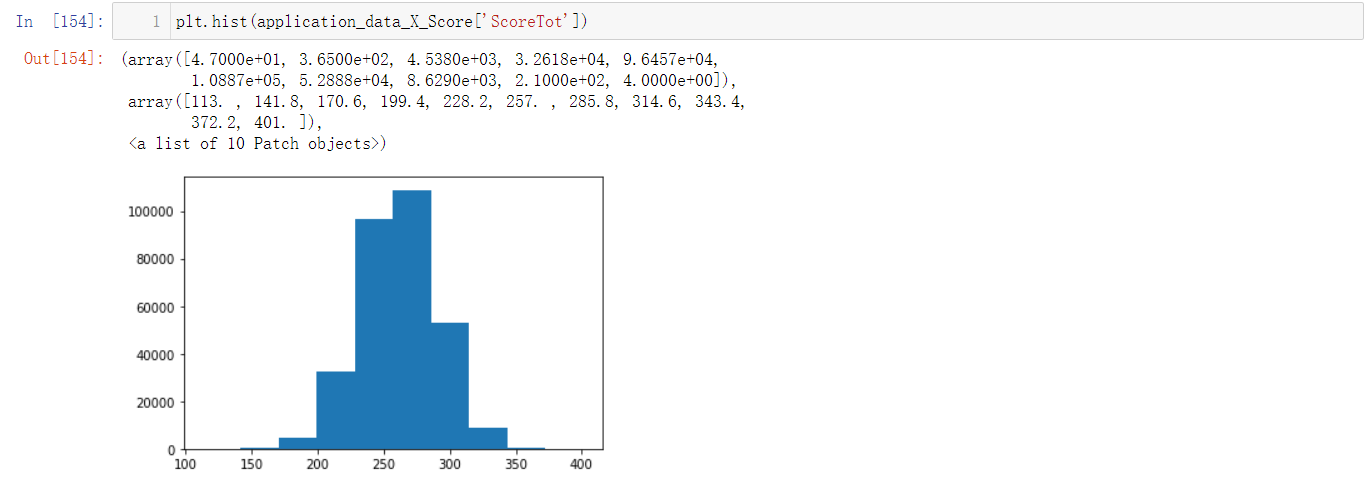
设定信用评分的范围为1到500，同时对逻辑模型得到的系数乘上一定的权重（公式为特征系数乘以设定的信用评分范围之差除以原样本集中各原特征系数和最大值与最小值之差）得到Score - Final（代表一个特征的分数），使计算得到的信用评分在范围之内。



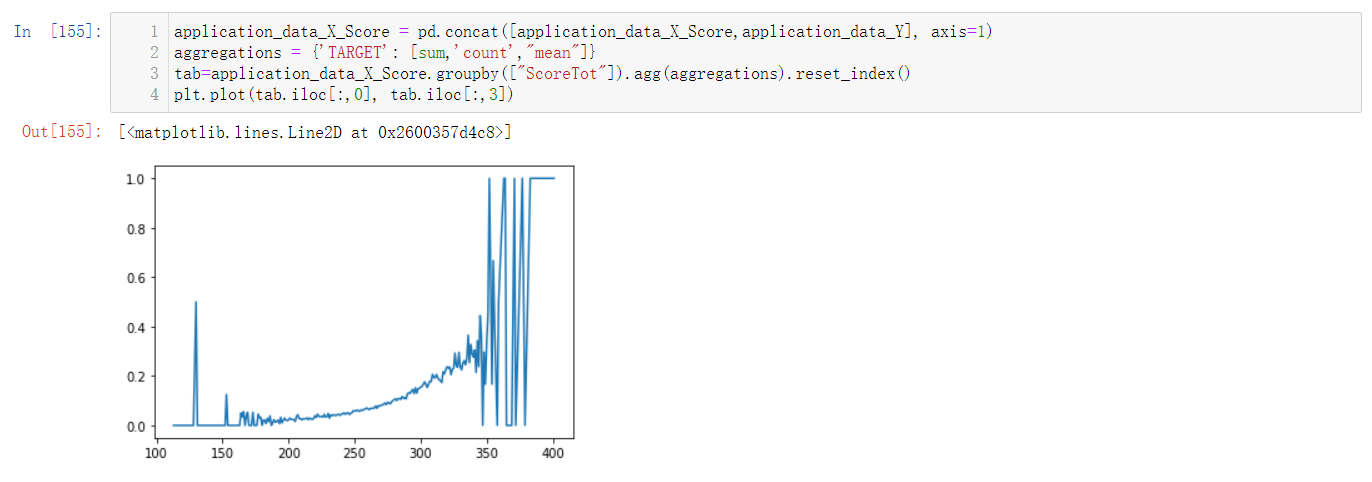
将Score - Final转置，用于后面的计算。



利用每个特征（每个特征都是哑变量，当为1时表示拥有某种特征，当为0时表示不具有某种特征）乘以对应的Score - Final（分数权重）得到总信用评分。



对样本的信用评分分布做图，由图可知，样本的信用评分集中在250到300分。



对信用分数和违约概率做一个统计，发现在我们的模型中，分数处于200-350的违约概率较小，而当分数超过350时，违约概率波动较大，需要堤防违约风险。