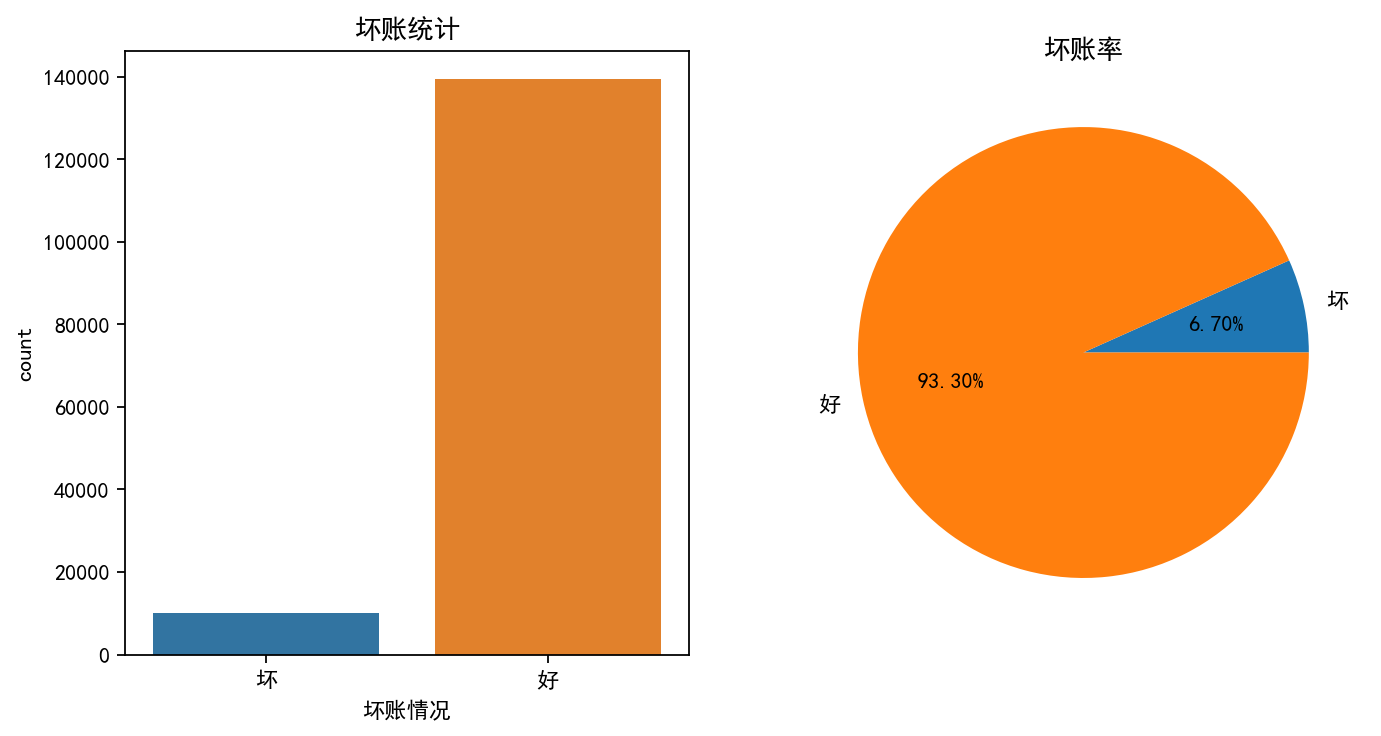
# 四：特征工程

### 4.1 样本平衡化处理

## 4.1.1 不平衡数据产生的原因

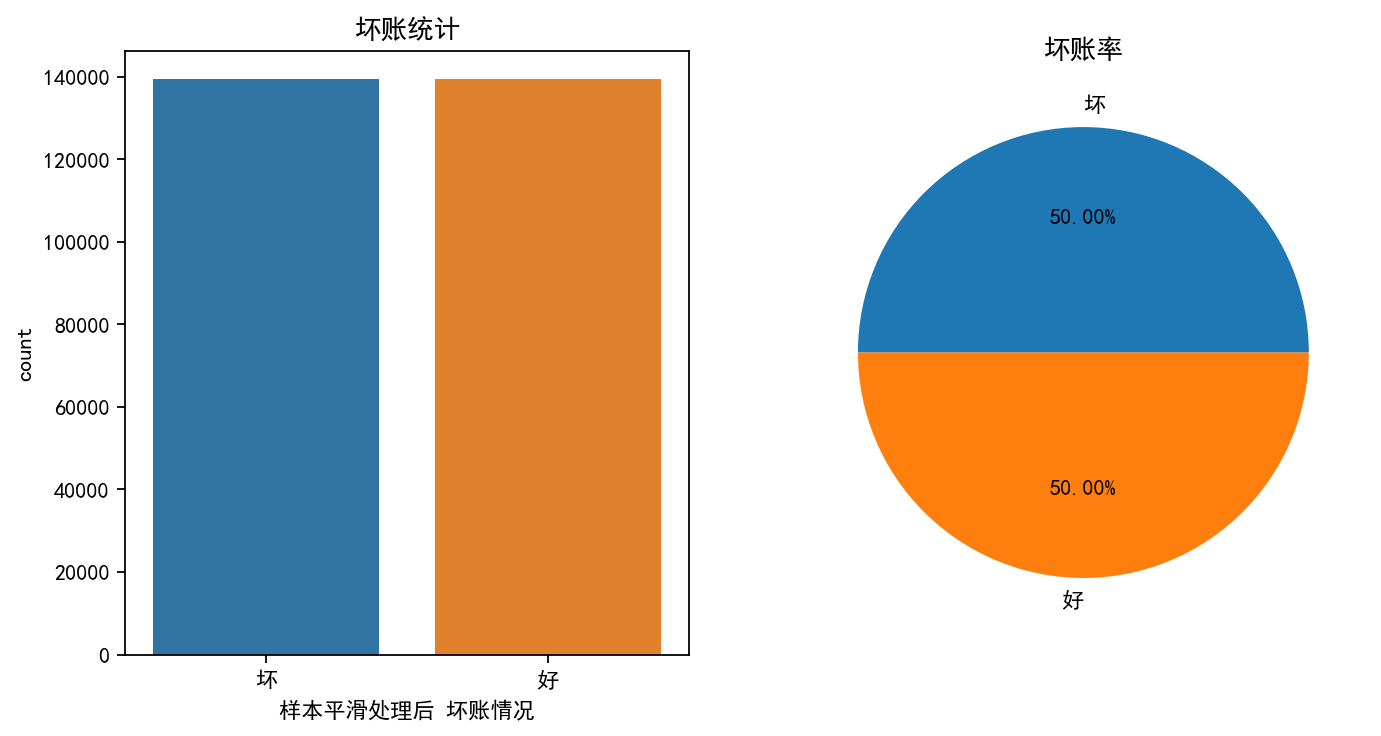
观察样本分布，发现严重不均衡，其原因之一，是人们皆在有意识的避免产生不良信用，故而实际违约者并不多。

从银行的角度出发，贷款给用户，用户遵守协议偿还本金以及利息，银行便可获利，故而银行真正的需求是筛选出 “恶意违约“ 之人，故而原因之二，便是银行并不会死板的将所有逾期的用户记录为坏账，定会尝试与客户沟通、证实，只要能把银行的钱还上，银行不会记此客户为坏账。



## 4.1.2 解决样本不平衡的方法

我们在采样方法上使用SMOTE算法来平衡样本，以保证之后逻辑回归的准确度。平衡之后的数据如下所示。



### 4.2 特征选择

## 4.2.1 数据分箱

#### 4.2.1.1为什么分箱

制作评分卡，目的是给各个特征进行分档，以方便业务人员能够根据新客户的信息为其打分，因此需要分箱，其本质是将连续变量转化为离散变量，让组之间的差异尽可能的小，不同组之间的差异尽可能的大，使不同属性之人划分为不同类别，对应不同的分数。

此外，数据特征的分布大多很极端，存在大量的围绕均值波动的数据，同时有少量远离大部分数据存在的极端高值，但是却不能将其归类为异常值，拿月收入Monthincome来说，存在一部分收入极高，脱离大部分数据分布的值，但是我们却不能判断其为异常，是因为现实中确确实实存在着收入远远高出普通人存在的超高收入人群。

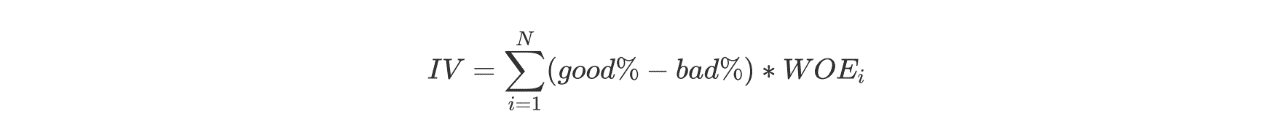
为了避免这些‘异常数据’极端值产生的影响，同时又避免特征中无意义的波动，我们采用分箱的方法，来提高模型的拟合度。

#### 4.2.1.2 关于分箱的个数

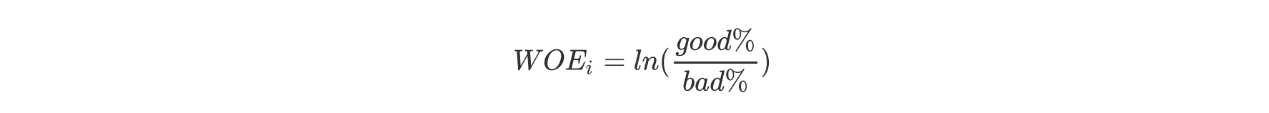
即然是将连续变量离散化，箱数想必不能过多，又考虑到用于制作评分卡，理想箱数应当在4~6左右。

**我们会对特征进行分箱，然后计算每个特征在n个箱子数目下的WOE值，利用IV值的曲线，找出每个特征合适的分箱个数。**

Information value(IV)：银行业用于衡量特征上的信息量以及特征对预测函数的贡献。



N是特征上分箱的个数，i代表每个箱子，good%是这个箱内的非坏账率（标签为0者之比例，银行认为的优质客户），bad%是这个箱子中的坏账率（违约者，标签为1的比例），而WOEi则写作：



银行业中用来衡量违约概率的指标，中文叫做证据权重(weight of Evidence)，本质其实就是优质客户比上坏客户的比例的对数。WOE越大，代表了这个箱子里的优质客户越多。而IV是对整个特征来说的，IV的意义如字面（Information value）意思一般，是特征上的信息量以及此特征对模型的贡献，由下表来控制：



因此，IV并非越大越好，我们需要找到IV值和箱子个数的平衡点。箱子越少，则IV值必然越小，若存在IV值足够大，并且箱子个数合适的点，那便是理想情况。

#### 4.2.1.3 卡方分箱

我们希望不同属性的人有不同的分数，因此我们希望在同一个箱子内的人的属性是尽量相似的，而不同箱子的人的属性是尽量不同的，即”组间差异大，组内差异小“。对于评分卡来说，我们希望一个箱子内的人违约概率是类似的，而不同箱子的人的违约概率差距很大，即WOE差距要大，并且每个箱子中坏客户所占的比重（bad%）也要不同。

可以使用卡方检验来对比两个箱子之间的相似性，如果两个箱子之间卡方检验的P值很大，则说明他们非常相似，那就可以将这两个箱子合并为一个箱子。

基于这样的思想，我们总结出对一个特征进行分箱的步骤：

·1.我们首先把连续型变量分成一组数量较多的分类型变量，比如，将几万个样本分成100组，或50组

·2.确保每一组中都要包含两种类别的样本，否则IV值会无法计算

·3.我们对相邻的组进行卡方检验，卡方检验的P值很大的组进行合并，直到数据中的组数小于设定的N箱为止

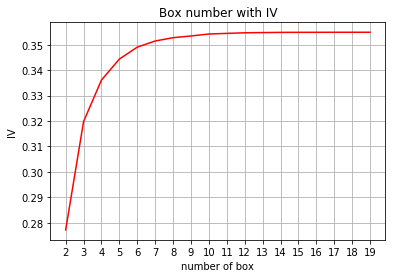
·4.我们让一个特征分别分成[2,3,4.....20]箱，观察每个分箱个数下的IV值如何变，找出最适合的分箱个数

·5.分箱完毕后，我们计算每个箱的WOE值，观察分箱效果

这些步骤都完成后，我们可以对各个特征都进行分箱，然后观察每个特征的IV值，以此来挑选特征。

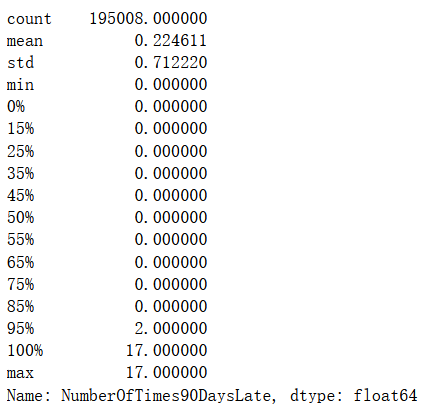
接下来，在构造完算法和定义完WOE和IV函数后（具体算法请看附录），我们以特征"age"和“NumberOfDependents”为例，展示分箱如何完成。

首先，利用卡方分布合并箱体，并画出IV曲线，选择转折点处，使得IV足够大而箱数足够合适，所以这里对于age来说选择箱数为6。而后为其余特征用相同的方法。



并非所有特征都可根据图像选取最佳箱数，对于分布不适合此方法，无法绘画出图像的特征，如特征“NumberOfTimes90DaysLate”，我们采用手动分组的方法分箱得到的结果为[0，2，7]，这是一种观察分布而后主观分组的方法。

#下面二图摆放为同一行



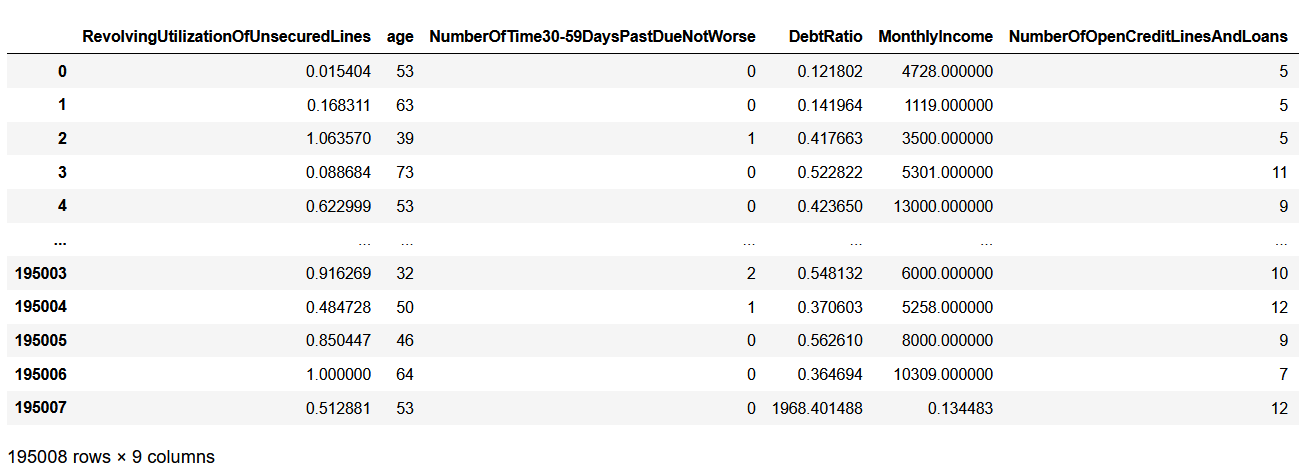


## 4.2.2 单变量分析

由于特征的IV值，可以反应其特征对模型的贡献，故将三组IV小于0.1的特征：NumberRealEstateLoansOrLines、NumberOfOpenCreditLinesAndLoans、NumberOfTime60-89DaysPastDueNotWorse。

分别单独的删去三者其一，再绘IV曲线图，发现只有当删去IV0.04（最小）的特征NumberOfTime60-89DaysPastDueNotWorse时，才能观察到原本Month-Income和DebtRatio此种主观认为对模型显著的IV值有上升的趋势，**故而决定删去此特征**，并调整两特征的分箱数。

至此，10列特征变为9列。



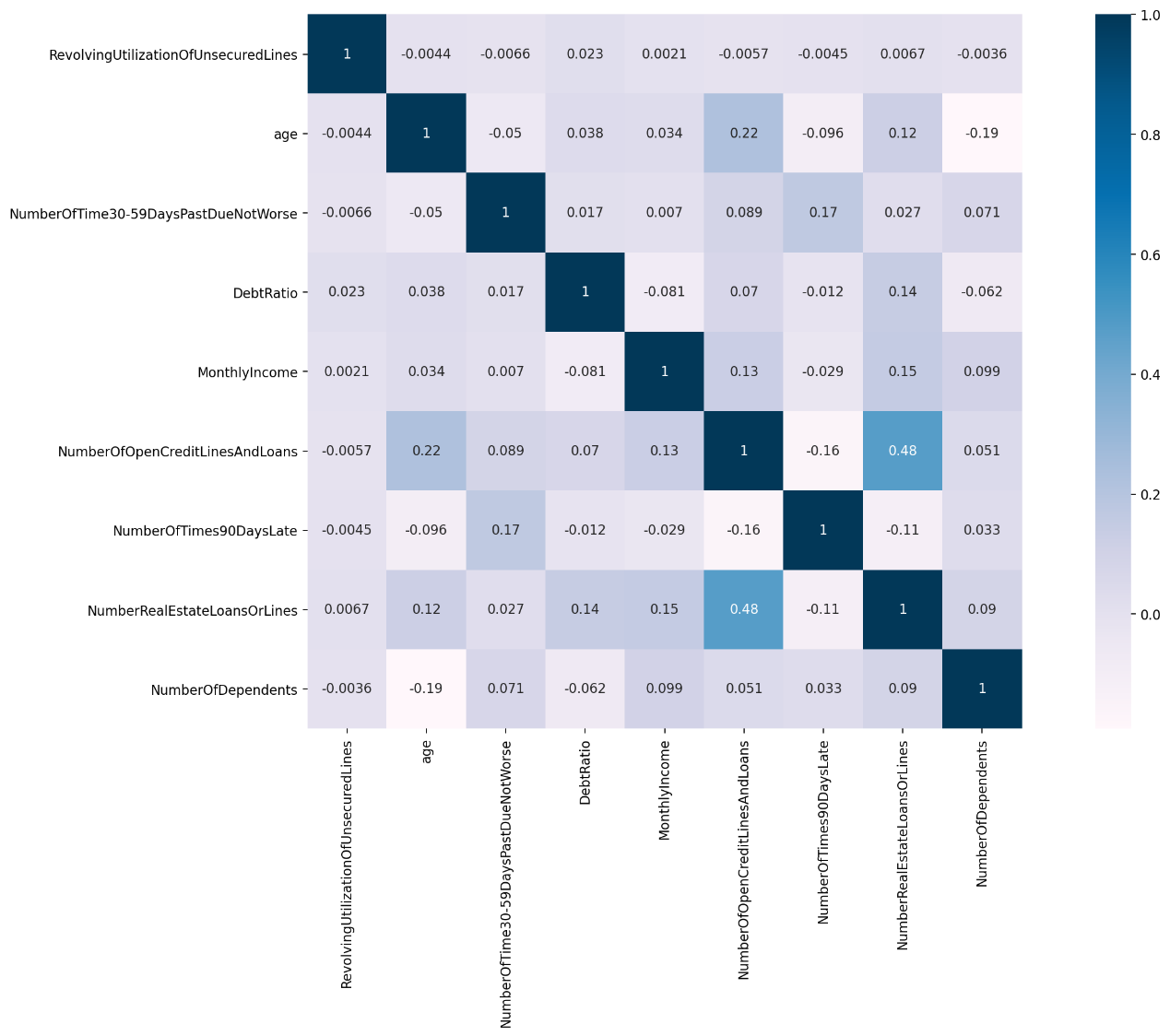
## 4.2.3 多变量分析

#### 4.2.3.1 皮尔逊检验（热力图）

可见特征NumberRealEstateLoansOrLines和NumberOfOpenCreditLinesAndLoans之间的相关系数接近0.5。

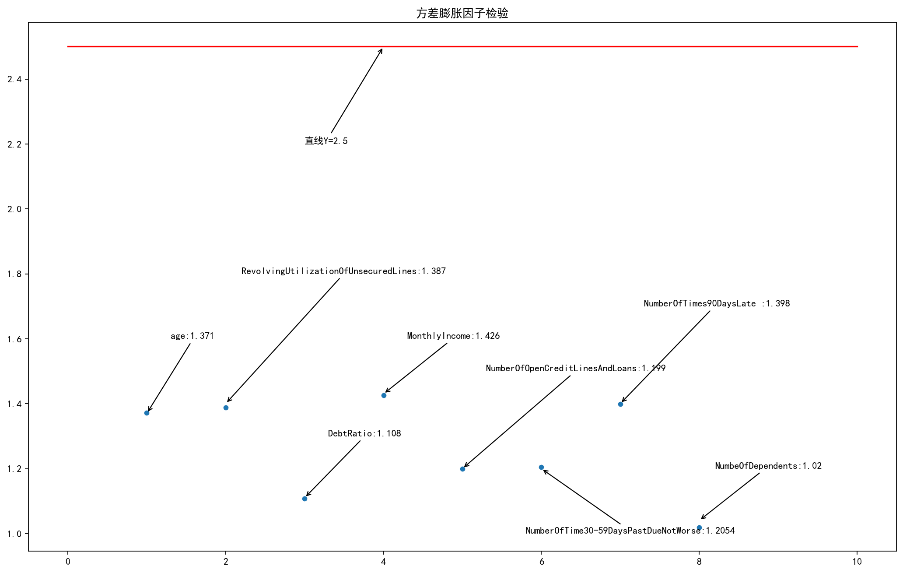
对于广义线性模型而言，若特征间存在近似的共线性，会导致拟合参数β估计精度很低，参数估计量经济含义不合理，变量的显著性检验失去意义，可能将重要的解释变量排除在模型之外，模型的预测功能失效。变大的方差容易使区间预测的“区间”变大，使预测失去意义。

**因此删除两者中对模型贡献最小的特征NumberOfOpenCreditLinesAndLoans，模型由9个特征变为8个。**



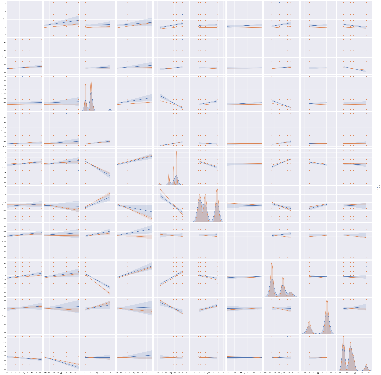
4.2.3.2 多重共线性VIF检验及特征交互性检验

对数据的8个特征的进行VIF检验，发现其VIF皆小于2.5 ，排除这些特征存在多重共线性的可能。



观察交互图，发现产生交互的特征大多为（NumberOfTime60-89DaysPastDueNotWorse与 NumberOfTime30-59DaysPastDueNotWorse 与 NumberOfTimes90DaysLate）又或（age与numberofindepende）认为这一类的特征交互对模型没有价值。

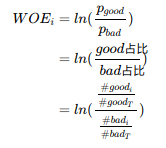
亦增加过特征如：‘月收入的平方除以年龄’ 意在分辨出年龄高而收入低者，但模型拟合度不尽人意。（交互图尺寸过于庞大，请于附录中查看）



## 4.3 woe编码

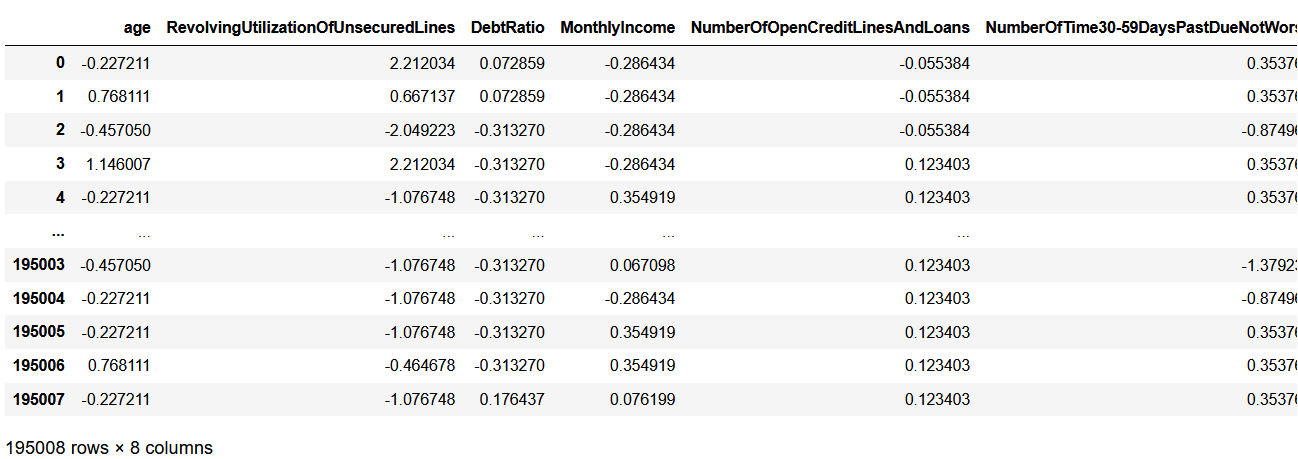
前者，我们基于对业务又或对模型的需求进行了分箱，但是模型是没法理解我们分箱之后划分的区间，我们要把他们编码，使得机器可以理解。

WOE编码是一种有监督的编码方式，将预测类别的集中度的属性作为编码的数值，所以将特征的值规范到相近的尺度上便是它的优势，但对于我们的模型而言，需要分箱后每箱都同时有好坏样本，我们当然也在算法考虑并实现了这一点。



计算各箱的WOE后，我们以此为依据对数据进行WOE映射，并且覆盖原数据，也就是说，我们接下来将用WOE覆盖后的数据进行建模，这样一来我们便获取到各个箱的分类结果，即评分卡上各个评分项目的分类结果，同样的操作我们也会在测试集上进行。

图1为WOE映射后的训练集 图2为测试集。



#上为图1 下为图2 还请标号

