#后面文字不要加

优点：在分箱过程中考虑到对于预测变量的影响而使用IV判别法

2.4, 异常值处理

最后加上：

所以我们最后的结论就是，先去除掉这225个样本，之后若是模型的拟合度不好，再选择将其加入。

3.2 特征分箱

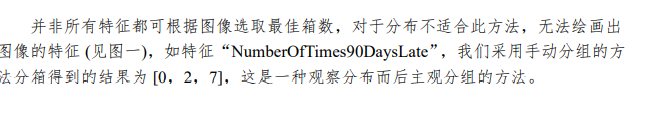
3.2.1 为什么要进行特征分箱

开头加上：在前面，我们已经对数据进行了重复值处理，缺失值处理，异常值处理，平衡化处理，但是

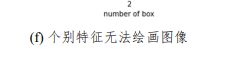
结尾加上：关于特征分箱，我们后面会有更详细的解释。

3.2.2

改成：首先从业务的角度考虑，分箱数不能过多，其次从模型的角度考虑，分箱数过多会导致IV值变大，对模型的贡献过大，会压缩其他特征影响，这就意味着，数据的微小浮动导致样本属于不同的分段的可能性变大，导致模型不稳定，在这里，我们使用Information value(IV)来进行箱数的确定，

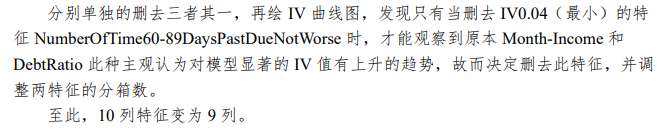


改成：前文提到并非所有的特征都能使用卡方分箱，一些类别变量，诸如“家人”就无法分出20组的箱数，这个时候我们只能采用等频分箱的方法，手动得到分箱区间。如特征“”我们通过观察他的分位点分布手动分出的区间为：[0,2,7]



改成：类别变量无法使用卡方分箱

P14:



改成：依次删除后发现，只有当删除IV值最小的特征(IV=0.04)“”时，MI和DR的IV上升，故决定删去“”，并根据新的IV重新确定分箱，至此，特征变为9列

将3.4 整体放到3.2.3 WOE编码

3.3.2

3.3.4 最后添加：处理完以后，再次映射各箱的WOE值，最终成为我们建模数据