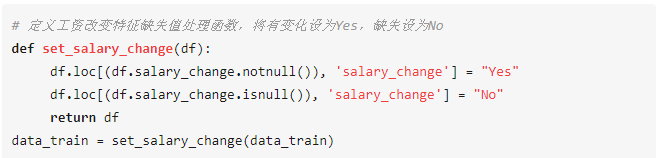
如何处理缺失特征

1. 缺失值较多的特征处理

可以采取将该特征舍弃，但会对结果产生不良影响。

还可以不采取删除特征的操作，而是换一种方式，将该特征分为两类，一类是非缺失的，我们将其设为 “工资收入有变化”，另一类是缺失的，将其设为“工资收入无变化”。

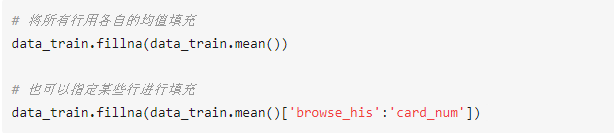


2. 缺失值较少的特征处理

方式一：把NaN直接作为一个特征，假设用0表示，实现如下:

data\_train.fillna(0)

方式二：用均值填充：

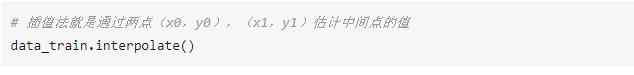


有时候遇到一种情况，那就是训练集train中有缺失值，而test中无缺失值。这时候最适合的处理方式应该是对缺失值取条件均值或条件中值，即根据该用户的label值类别，取所有该label下用户该属性的均值或中值。

方式三：用上下数据进行填充：



方式四：用插值法填充：



方式五：用算法拟合进行填充：



当然，针对我们这里的数据，我们对那些缺失值不是很大的特征都采用方式五来填补其缺失值。即使用随机森林算法，利用数据表中某些没有缺失的特征属性来预测某特征属性的缺失值。