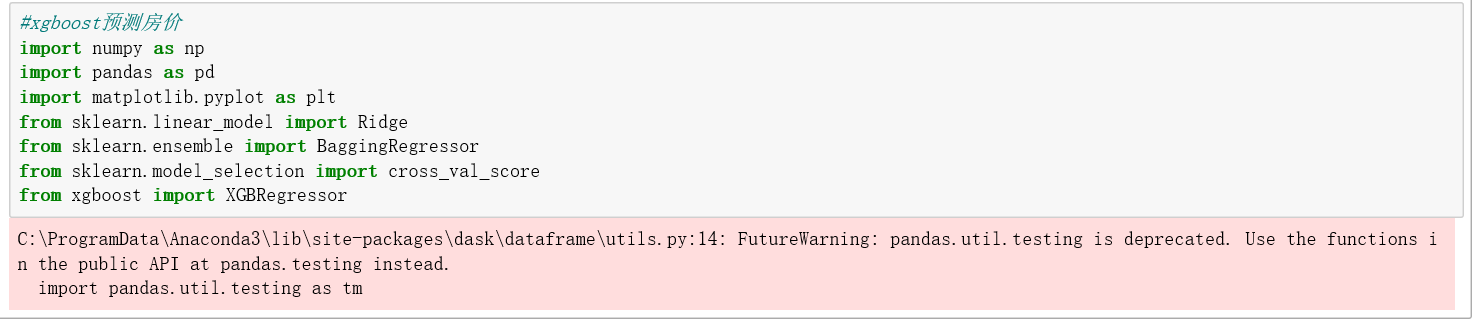
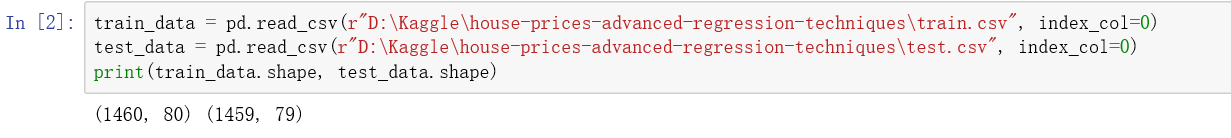
房价预测数据集说明书

这次作业我采用xgboost回归算法预测kaggle house-prices-advanced-regression-techniques比赛数据集

1. 导入所需的依赖包：



1. 导入测试集和训练集



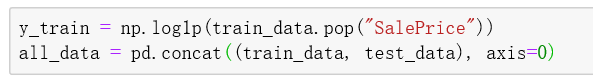
1. log1p=log(x+1),这行代码用原始数据中的房价和取对数后的房价创建了个新的DataFrame

在数据预处理时首先可以对偏度比较大的数据用log1p函数进行转化,使其更加服从高斯分布

(S~QC2D]~B}MI2EJO16DBUT

1. 数据预处理

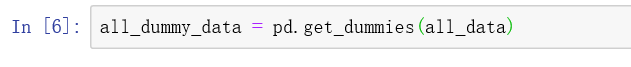
先把训练集中的房价拿出来，将测试集合和训练集合进行拼接，共同作为训练集训练数据



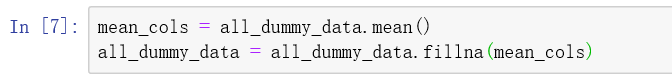
观察数据集中的特征，其中MSSubClass是一个类别特征，他们之间没有大小关系，但是pandas默认将其处理成数字，要将其转化成str

]X~GYIOJ}VEWPJ4VW9P8N`K

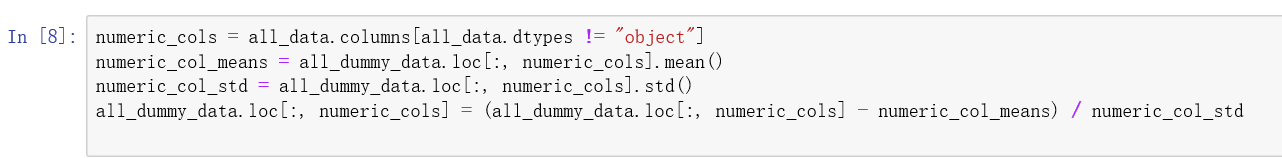
将字符串特征转化成One-Hot编码形式



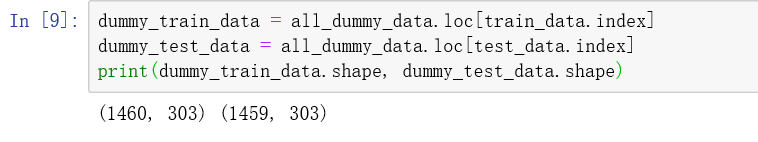
用平均值填补缺失值



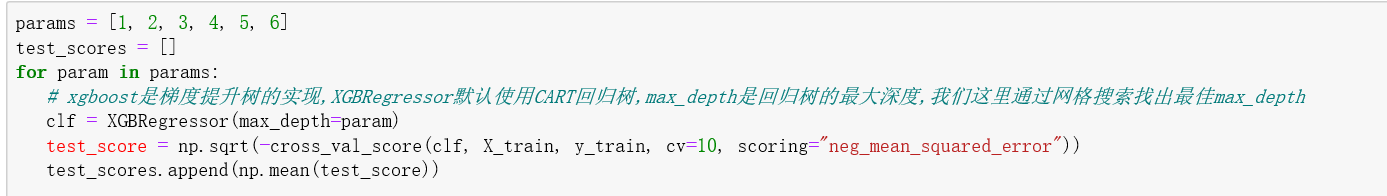
数据归一化

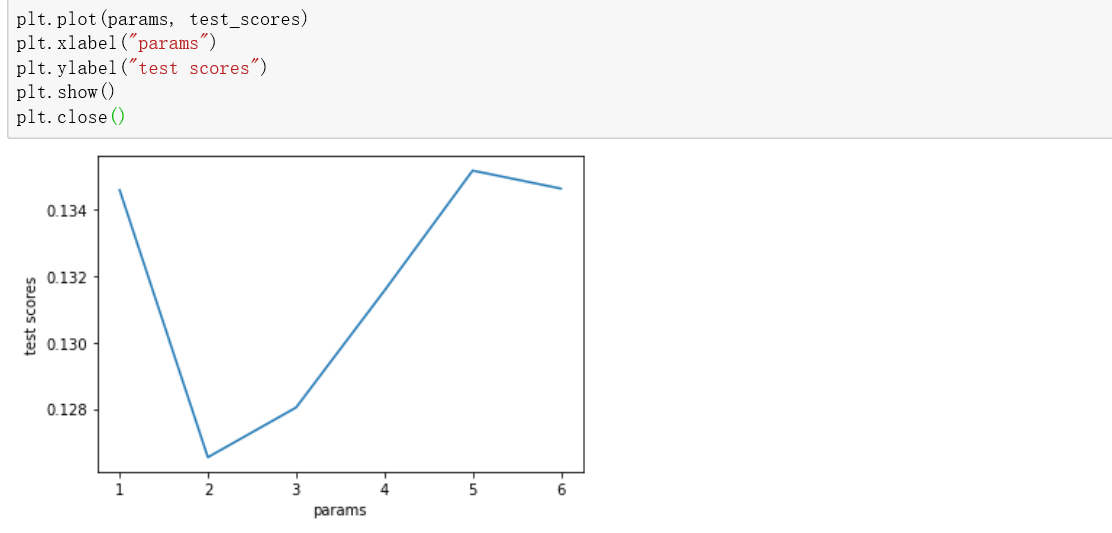


1. 重新划分训练集和测试集

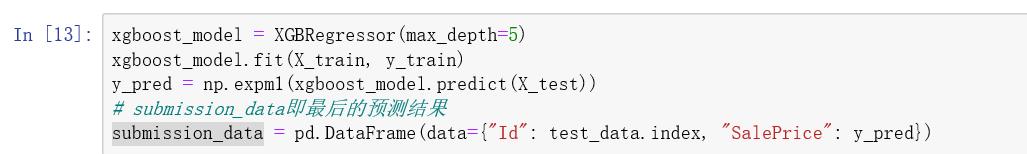


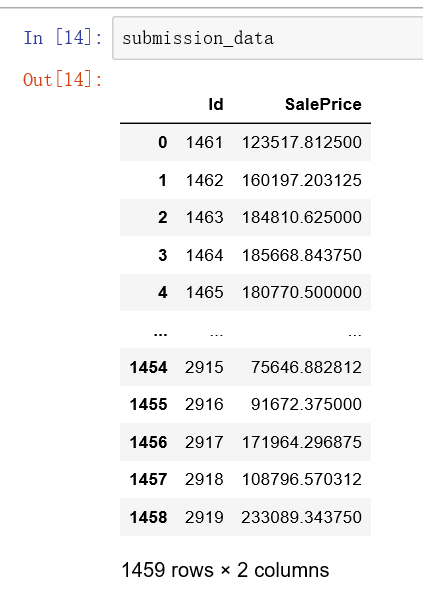
1. 网格搜索找出回归数的深度最佳深度（这里之前有范围较大的网格搜索但时间较长，最后取较小范围中选择最好的max\_depth）





1. 用最佳参数进行训练





1. 保存到本地

8~~@1DFY50LQ{){F3_O816G