房价预测

Step 1: 检视源数据集

Step 2: 合并数据

使用log x+1 平滑化label

Step 3: 变量转化

类似『特征工程』。就是把不方便处理或者不unify的数据给统一了。

正确化变量属性

One-hot处理string类型，遇到以数字表示的分类，需要先将其改为string类在使用one-hot。

对于数字类特征。遇到缺失的数据，根据具体情况使用均值或0或其它值补全。

使用(X-X')/s或者log是数据平滑化。

sklearn.preprocessing.LabelEncoder()：标准化标签，将标签值统一转换成range(标签值个数-1)范围内

.

检查所有数字特征的偏斜

skewed\_feats=all\_data[numeric\_feats].apply(lambdax:skew(x.dropna())).sort\_values(ascending=False)

print("\nSkew in numerical features: \n")

skewness = pd.DataFrame({'Skew' :skewed\_feats})

skewness.head(10)