设计说明书

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

代码文件：House\_Prices.ipynb

运行环境：python3.8

编译环境：Jupyter notebook

比赛说明



让购房者描述他们梦想中的房子，他们可能不会从地下室天花板的高度或与东西铁路的接近程度开始。但是这个游乐场比赛的数据集证明，与卧室数量或白色栅栏相比，对价格谈判的影响要大得多。

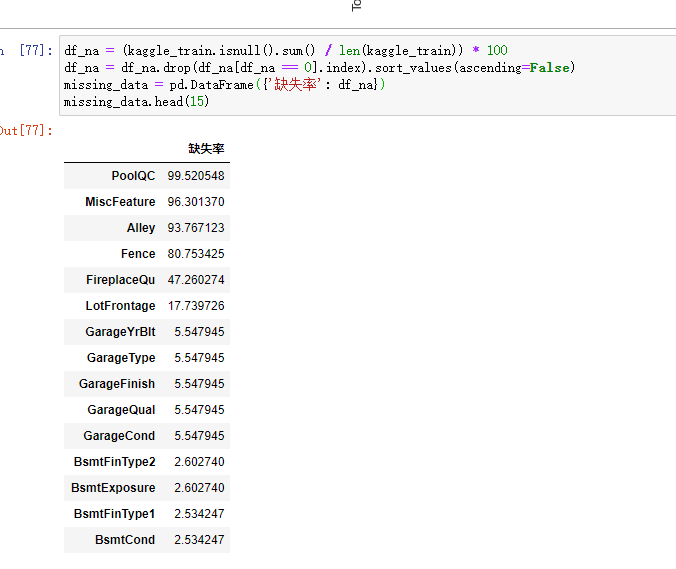
凭借 79 个解释变量（几乎）描述了爱荷华州艾姆斯住宅的各个方面，这项竞赛挑战您预测每栋房屋的最终价格。

目标

影响房价的因素有很多，在本题的数据集中有 79 个变量几乎描述了爱荷华州艾姆斯 (Ames, Iowa) 住宅的方方面面，要求预测最终的房价。我们的任务是预测出每间房屋的价格，对于测试集中的每一个Id，给出变量SalePrice相应的值。

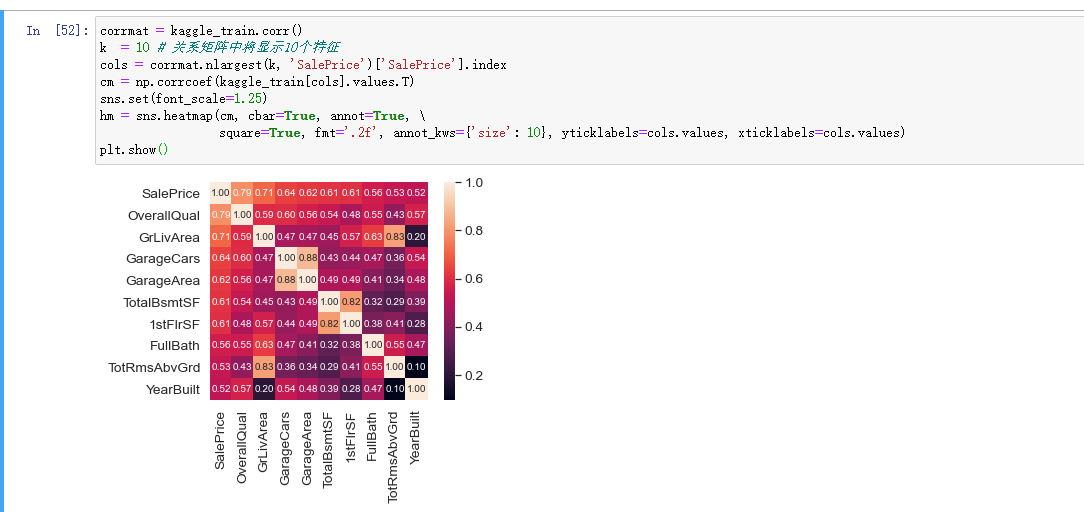
数据分析：

分析发现某些列的数据特征存在缺失值

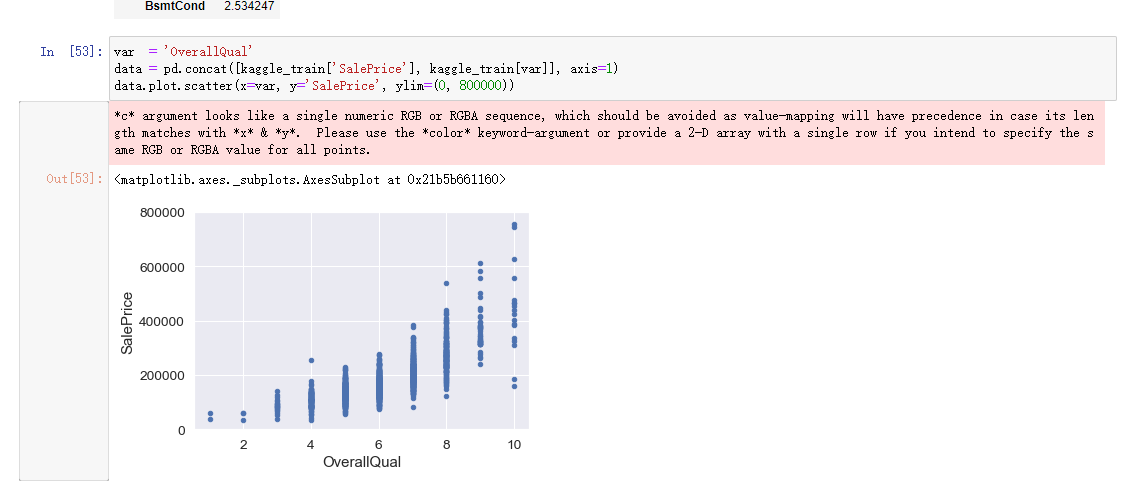


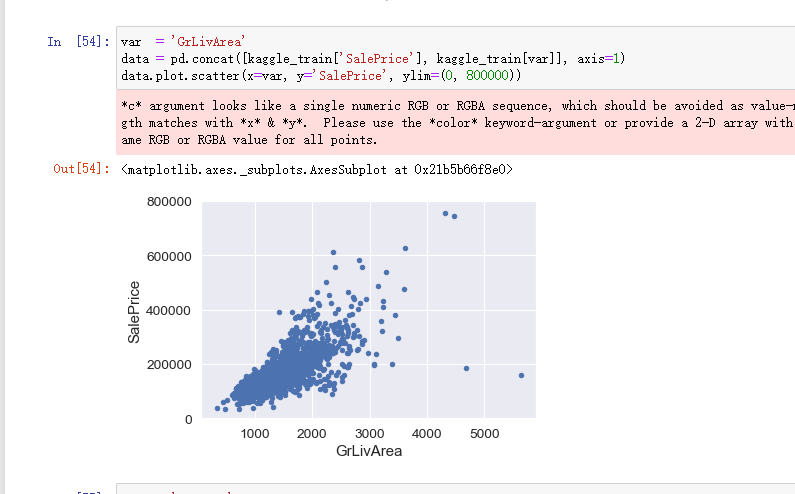
我们可以看到PoolQC、MiscFeature、Alley、Fence、FireplaceQu 等特征存在大量缺失。

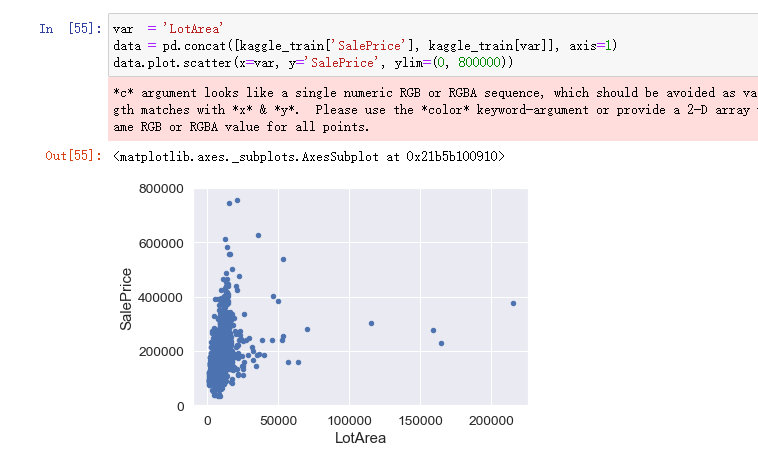
接着分析热力度前十的特征



对于各个特征和SalePrice画出散点图







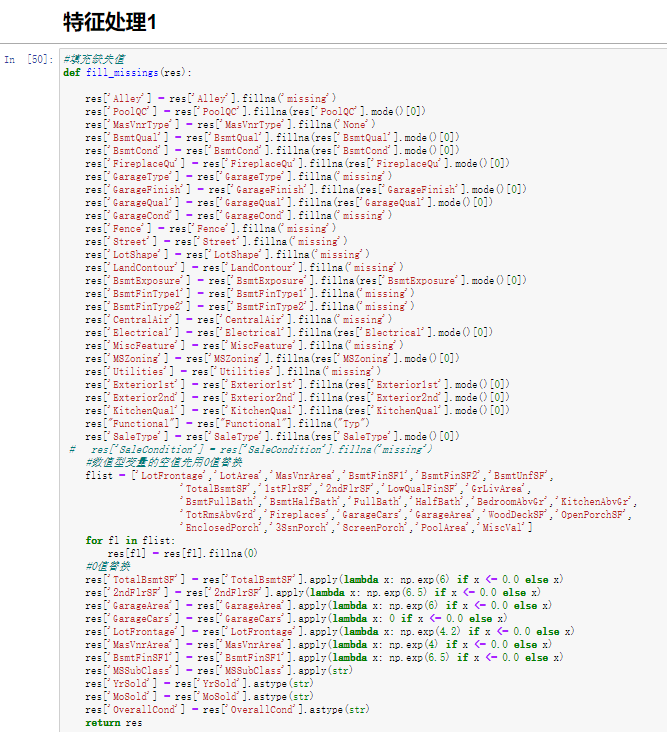
发现热力值靠前则对SalePrice确实有线性关系。

特征处理：

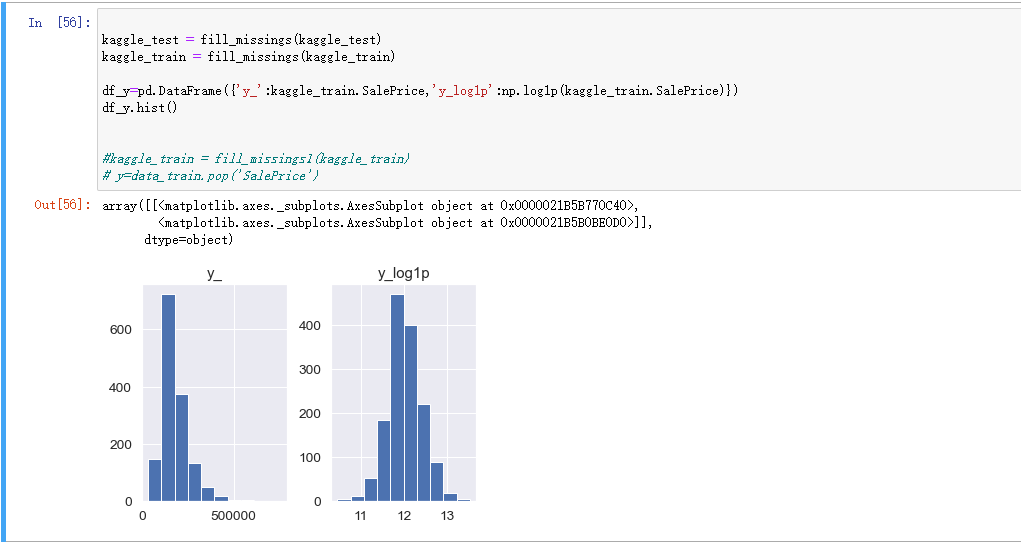
好的特征能提高预测准确度的上限，而好的模型只是能让我们靠近这个上限。

首先我们发现在数据中存在许多的缺失值，我们对一部分缺失列采用众数填充，对另外一部分缺失列将缺失值表示为一种新的类别。

还有一些特征其被表示成数值特征缺乏意义，例如年份还有类别，这里将其转换为字符串，即类别型变量。



对于回归模型中，预测值尽量要处理成正态分布，由于SalePrice特征偏度较大，所以我们对其进行平滑处理。使用log1p函数对其进行转化，使其更加服从高斯分布



对于不同的特殊数据进行不同处理：

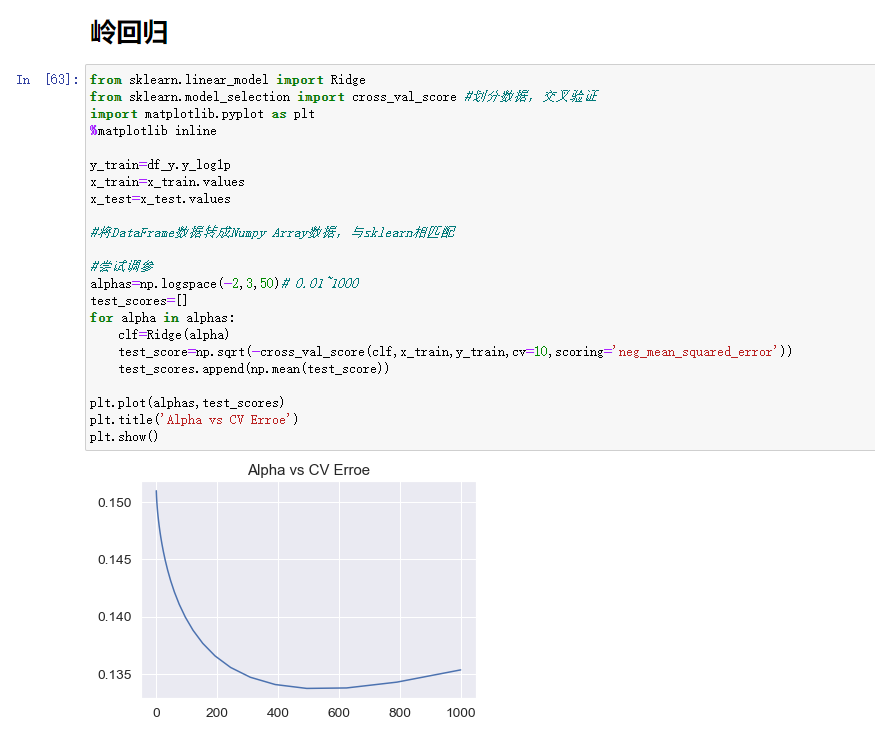
1、category数据：变成one-hot编码，pd.get\_dummies()

2、numerical数据：处理缺失和偏差等等

建立模型:

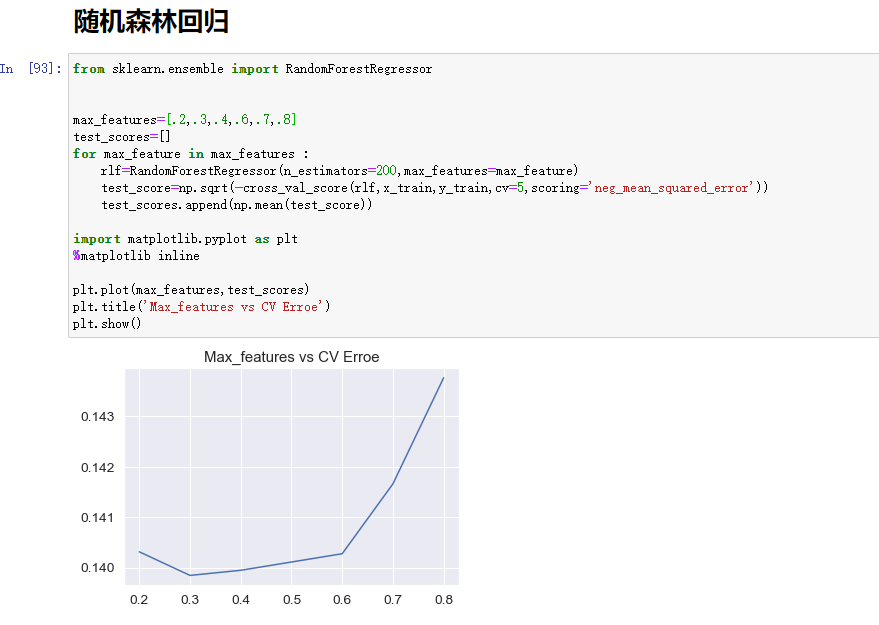
通过sklearn里的正则化+岭回归和随机森林回归分别建立两个模型

在使用岭回归模型进行建模时，需要考虑Ridge的alpha参数，我们对建立不同的alpha参数的模型，分别求出损失值



在alpha=400左右，取到最低值。

在随机森林回归模型中，max\_features参数是随机森林允许单个决策树使用特征的最大数量。我们可以使用“0.X”的格式，来控制使用特征x%来使模型准确度达到最优值。

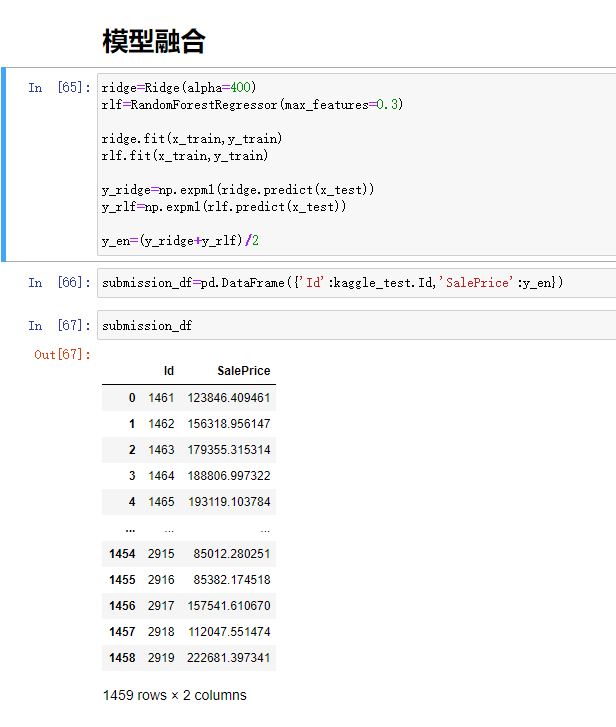


分别选取20%、30%、40%、60%、70%、80%进行尝试，发现当max\_features参数取到0.3的时候效果最好。

模型融合：

可以通过数学证明模型，随着集成中个体分类器数目T 的增大，集成的错误率将指数级下降，最终趋向于零。所以我们可以训练多个模型，然后按照一定的方法集成各个模型，使最终的预测准确率提高。

对于上述两个模型分别对test数据集进行预测，然后将预测结果取均值进行融合。得到最终房价预测结果。

.

上传S3：

最后通过boto3连接S3将预测结果写入S3中

