

课程设计报告

（ 2021秋季学期 ）

|  |  |
| --- | --- |
| 课程名称 | 机器学习课程设计 |
| **项目名称** | **房价-高级回归技术** |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 傅凯 | 学号 | 20190440908 |

|  |  |  |  |
| --- | --- | --- | --- |
| 专业 | 软件工程 | 班级 | 三班 |

|  |  |  |  |
| --- | --- | --- | --- |
| 地点 | 机房 | 教师 | 刘永彬 |

1. **项目概述及计划**

**项目概述：**Kaggle之入门预测竞赛题：房价-高级回归技术 预测销售价格并实践特征工程、RF和梯度提升。数据集由Dean De Cock编译，用于数据科学教育。

**计划：**根据所学知识，通过机器学习流程来制定计划：

数据收集—>数据探索—>数据预处理—>模型选择—>模型训练—>模型评估—>性能改进—>处理结果

**1.数据收集、数据探索：**本项目数据平台已提供，不需要收集数据，将网站提供的训练集、预测集读取，进行数据质量分析、数据特征分析或者图表分析。

**2.数据预处理：**查看各变量与预测目标的相关性，查看各变量缺失率、填补缺失值、进行类别特征编码，并检查是否仍存在缺失值

**3.模型选择、模型训练：**在上面进行了数据预处理部分后、需要对处理后的数据构建模型。本项目可以使用线性回归、梯度提升回归、决策树回归、随机森林回归等模型训练并评估数据

**4.性能改进：**通过不同模型选择、训练，找到预测最接近准确的模型预测，加以结果处理，完成预测。

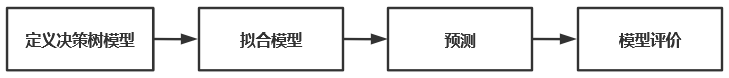
1. **问题描述**

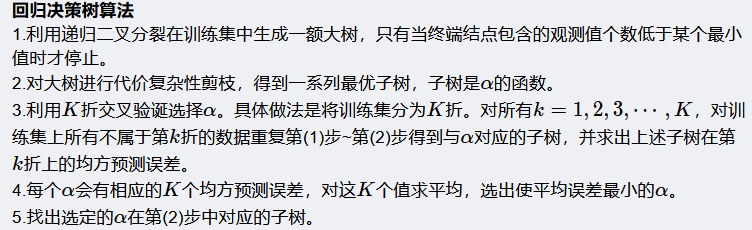
让购房者描述他们梦想中的房子，他们不可能不会从地下室天花板的高度或离东-西向铁路的距离开始。但这场比赛的数据集证明，影响价格谈判的因素远远大于卧室数量或白色栅栏。

通过79个解释变量来描述爱荷华州埃姆斯住宅的各个方面（几乎），本次比赛挑战您预测每套住宅的最终价格。

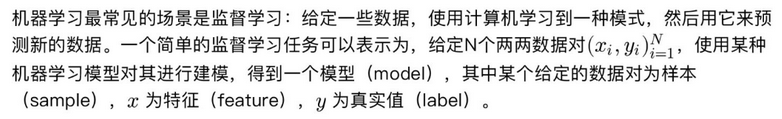
1. **模型建模**

**1.决策树回归：**





**2.线性回归：**



当变量之间保存等比例的关系，同时预测的变量输出集合是无限且连续，这时候就可以用到线性回归模型。

线性回归模型有RidgeCV()和LassoCV()两种，通过分别训练两种模型L1,L2。通过两种回归模型训练出结果，比较真实目标变量与真实预测的目标变量比较，也可以进行进行模型融合，最终得到两种模型的均值。

**3.LightGBM模型：**

GBDT (Gradient Boosting Decision Tree) 是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。LightGBM模型是一个梯度提升决策树（GBDT）的实现，其本质原理就是利用基分类器（决策树）训练集成，得到最优的模型。相比于其他梯度提升决策树模型来说，LightGBM模型可以更直观的提升训练速度，使用算法构建决策树，减少了不必要的计算，降低了内存消耗。

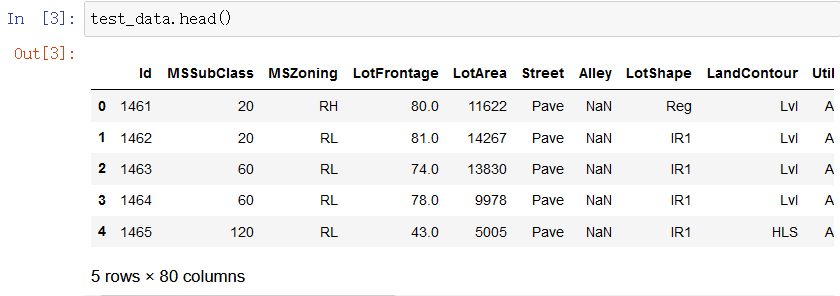
1. **实验分析**

**4.1 数据分析**

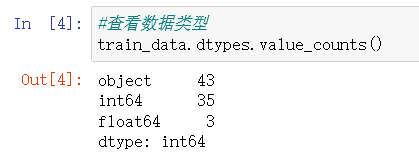
数据来源于kaggle数据集，train.csv为训练集 、test.csv为预测集

首先查看训练集合和预测集的结构

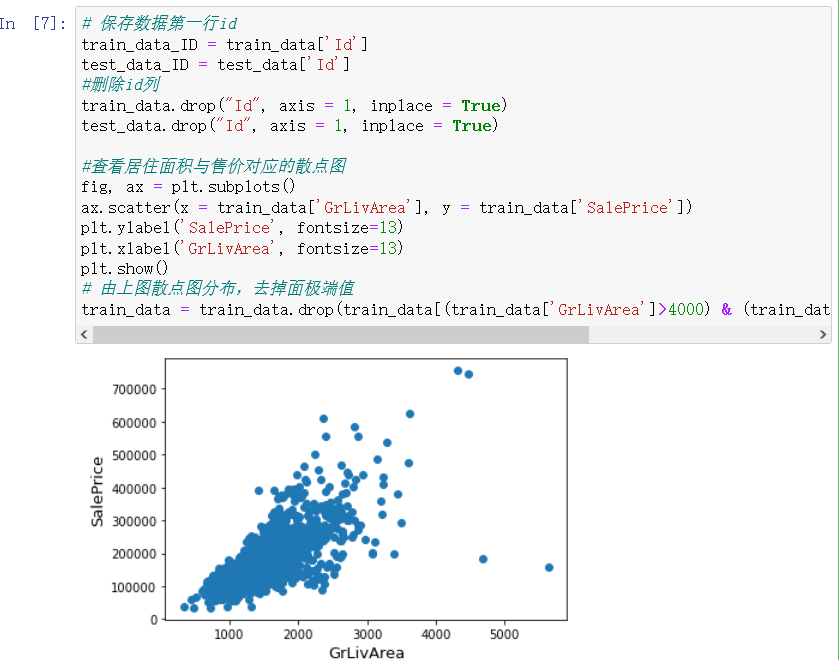




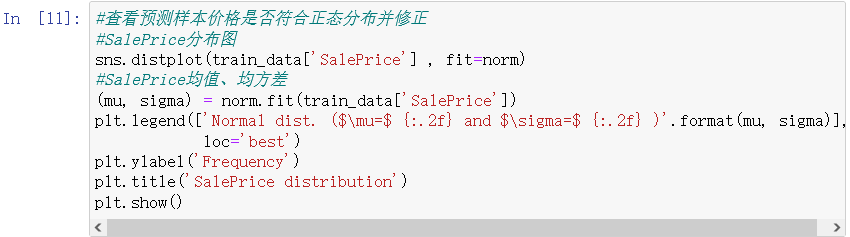
查看数据源中的数据类型：

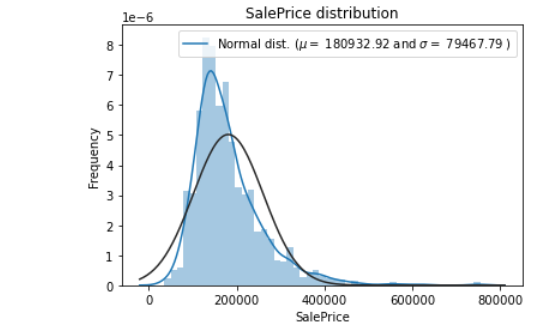


知道最基本的信息之后，就要开始进行数据预处理了。先画出一个代表性的散点图、查看数据大概情况、然后去掉极端值

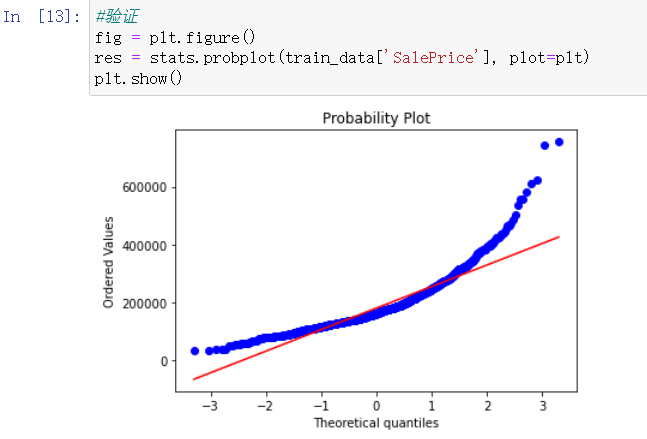


查看预测样本价格是否符合正态分布并修正（线性回归模型需要用到）

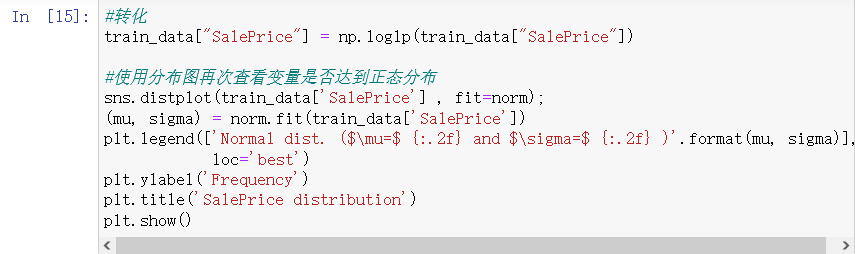


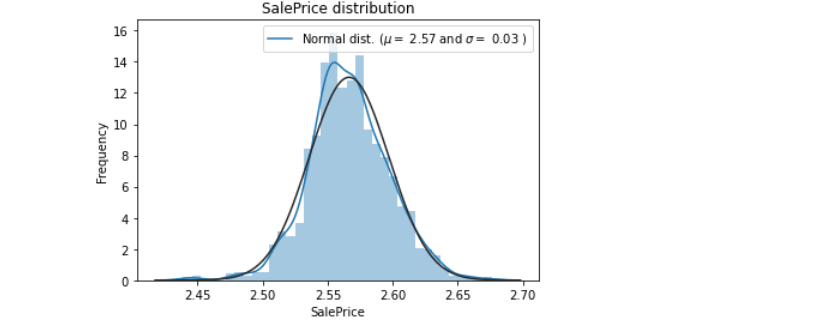


验证某两组数据是否来自同一分布

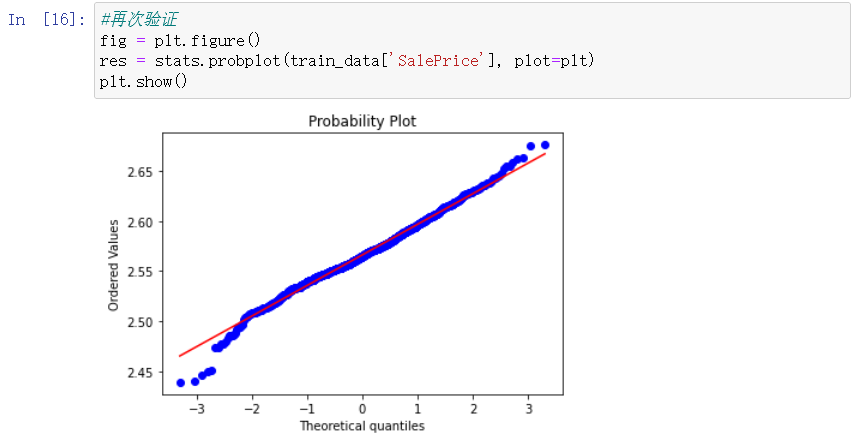


由上面的直观图可知，目标是右倾斜的，模型需要转化为正态分布的数据，使用log(1+x)对售价进行转化

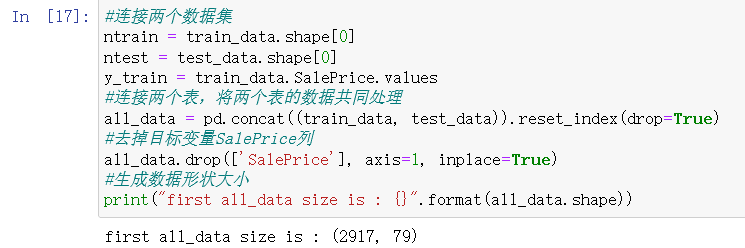




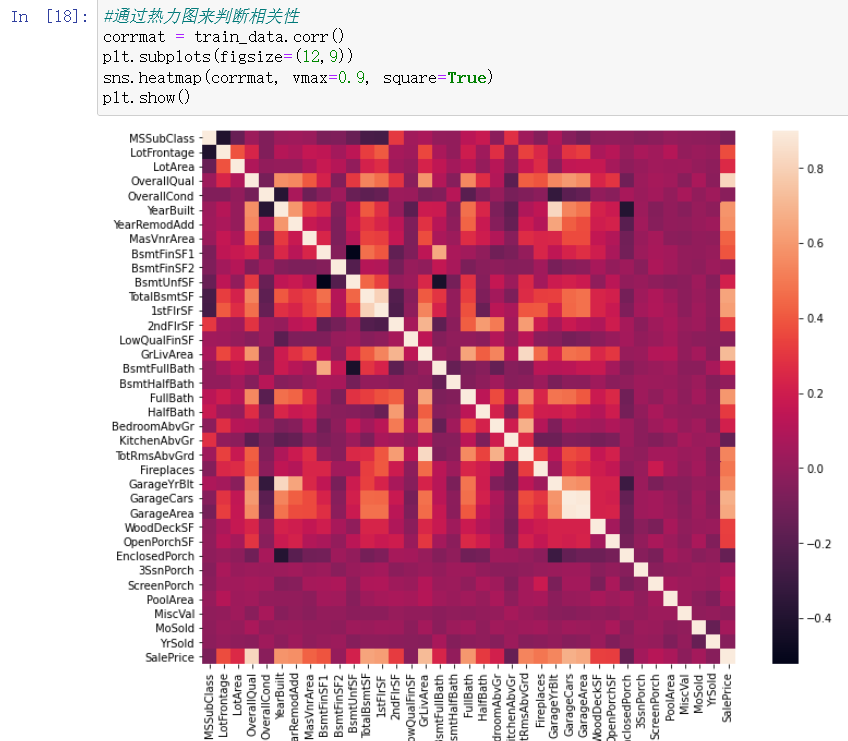
再次验证



完成上述工作后、需要开始构造特征工程了（其他模型需要用到）

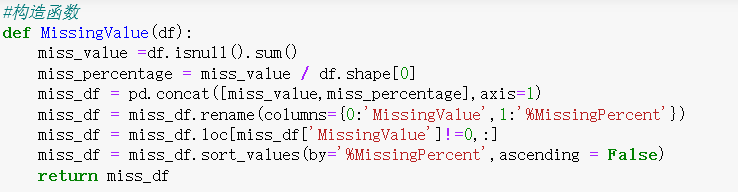


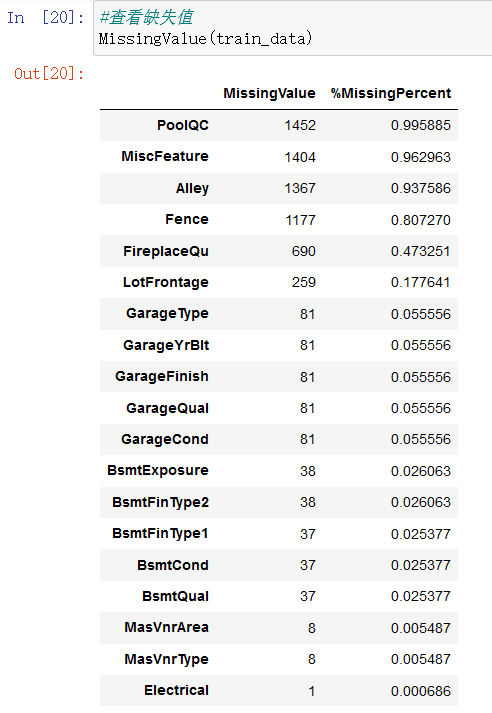
通过热力图查看各变量与预测目标相关性，根据颜色条判断

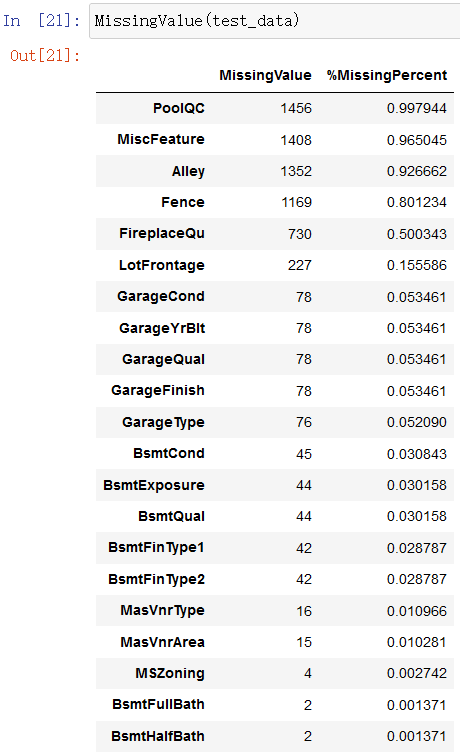


之后需要进行缺失值处理

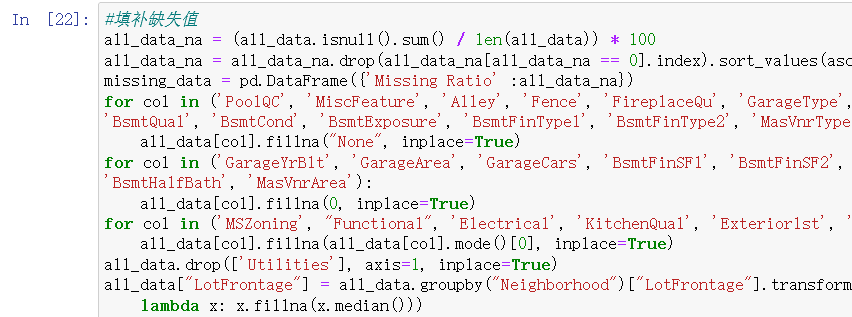
先使用函数功能查看缺失值



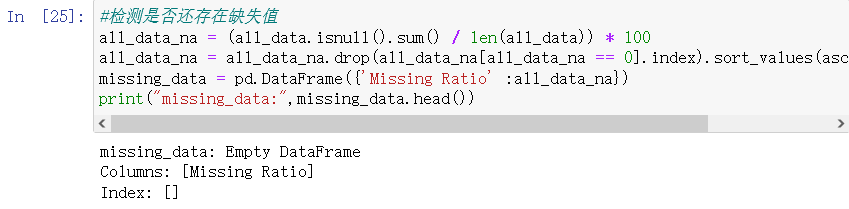




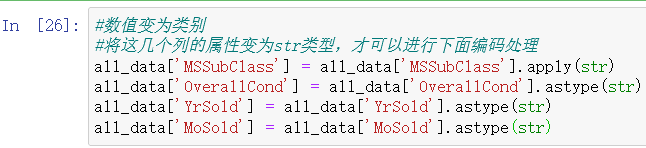
填补缺失值



检测是否还存在缺失值



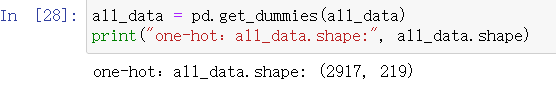
处理完缺失值后、需要开始构造类别特征，将数值变为类别



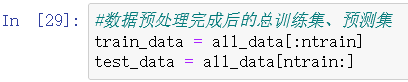
将某些特征进行标签编码LabelEncoder0,1,2···，对于回归问题，更多地使用labelEncoding。对于分类问题，更多使用OneHotEncoding。



将属性特征转化为指示变量，one-hot编码（object类的特征会根据类别使用不同的编码方式处理）



至此，数据预处理环节全部完成



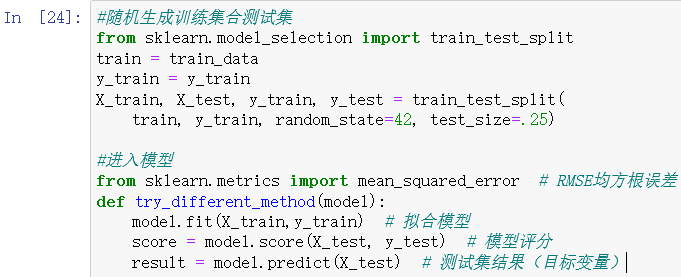
**4.2 实验环境**

**Anaconda3（提供数据包）**

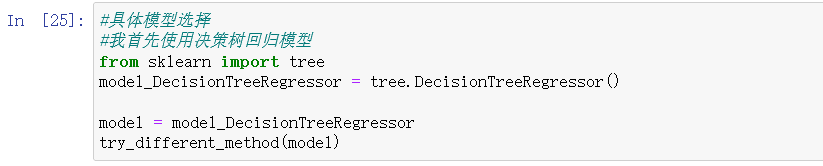
**Jupter Notebook python 3**

**4.3 实验结果**

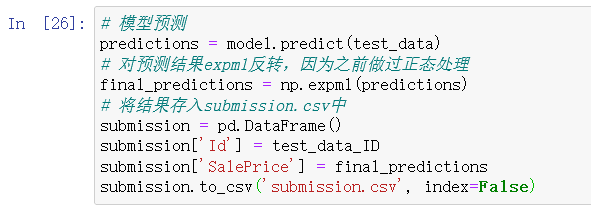
在数据预处理后、需要开始对数据进行模型训练



我首先选择的是决策树回归模型



训练完后开始进行预测



提交结果（误差值0.21909）



再次优化数据（数据预处理部分）

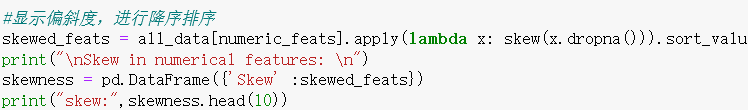
在构造特征工程部分进一步优化：首先将几个特征组合起来，增加一个重要特征



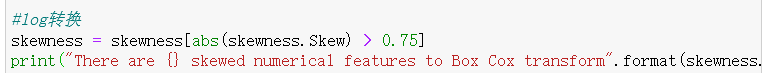
将倾斜数值数值特征提取



列出数据的偏斜度，降序排序



将偏斜度大于0.75的数值做log转换，使之尽量符合正态分布

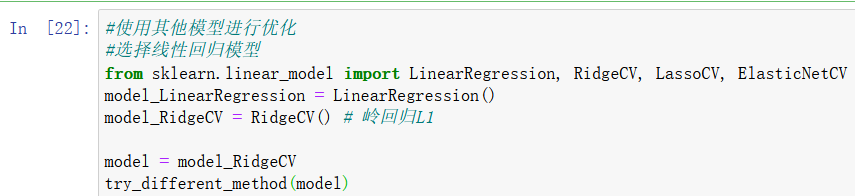


重新进行预测提交结果（误差值降低了，但是不明显，为0.20924）：



使用其他模型进行训练

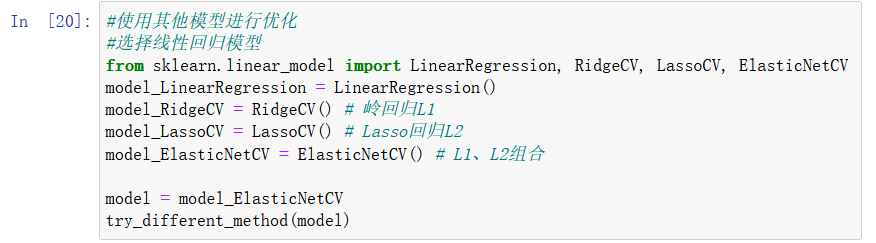
选择线性回归模型来训练数据



提交结果（误差值减少了很多，为0.12999，排名提升了将近3000名）

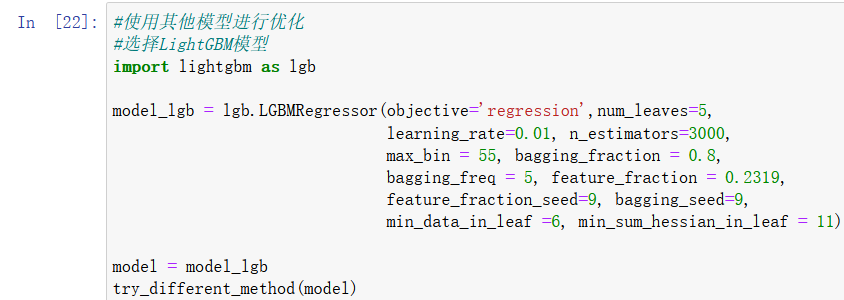


将两个模型进行组合



结果显示当两种模型组合后，误差反而更大了

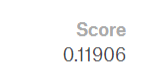
最后再使用LightGBM模型进行优化



结果显示（误差为0.12595）：



使用Kaggle排行榜大佬所写的数据运行：



**4.4 实验对比（竞赛排名）**

首次数据预处理后使用决策树回归模型进行训练：



优化数据预处理后：



用线性回归模型进行训练：

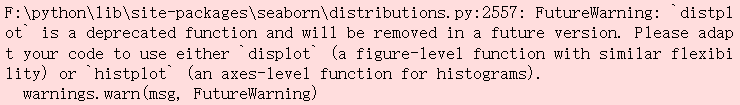


用LightGBM模型进行训练：



1. **项目总结**

1.数据预处理时，使用matplotlib绘制直观图时报错

****

错误原因：distplot是已经弃用的函数、会在以后的版本中删除，需将代码改为“ displot”（具有类似灵活性的图形级函数）或“ histplot”（直方图的轴级函数），但是使用displot后，图形会有所变化，故可以不进行处理。

2.数据预处理的完善对于结果的影响非常大，在我的数据预处理代码部分没有做到更加的细节，很多地方没有考虑到，仍有很大的提升空间。比如缺失值、异常值等地方如果处理到位，可以向Kaggle上面大佬的代码一样，排名能提高很多，这是一个值得反思的地方。

3.使用模型训练之前、需要考虑好使用的模型是否符合自己的项目需求，不能盲目去套用模型、无谓的增加自己的工作量。

4.需要提高自己对于机器学习的深入理解，在刚开始自己根据其他大佬的博文研究花了很大的功夫才明白了大部分代码。然后自己在编写时总是会遇到大大小小的各种问题，边百度边解决问题，导致最后自己也不知道代码是不是到底实现了想要的功能。但是通过这次的课程设计，我也掌握了许许多多上课没有学好的知识，如线性回归模型中的岭回归，如决策树模型等等。