# 房价预测模型实验报告

1. **最终版本简述:**

**（一）数据读取**（参照Homework11）

**（二）数据处理**

1.删除 ‘ADDRESS’, ’APARTMENT NUMBER’

2.['BOROUGH', 'BUILDING CLASS CATEGORY','TAX CLASS AT PRESENT','TAX CLASS AT TIME OF SALE','BUILDING CLASS AT PRESENT','BUILDING CLASS AT TIME OF SALE','SALE DATE','NEIGHBORHOOD']

全部转换为one\_hot\_feature编码

1. ’GROSS SQUARE FEET’,’LAND SQUARE FEET’先将’- ‘替换成0，然后转换为int类型
2. 构造‘oTAX CLASS AT PRESENT’
3. 构造‘oSALE DATE’，为以2016年一月为1计数的相对月份(即，若2016年3月出售，记为3）
4. 仿照11处理，做数据变形
5. **模型构建**

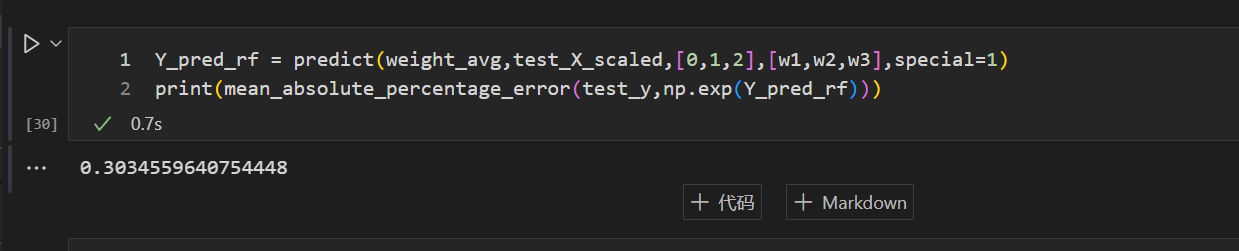
1.

取RF，Extra，Xgb模型(参数如下)  
RF = RandomForestRegressor(n\_estimators=600,n\_jobs=200,verbose=False)

Extra =  ExtraTreesRegressor(max\_depth=160,n\_estimators=160,n\_jobs=200)

Xgb = XGBRegressor(max\_depth=6,n\_estimators=1600,learning\_rate=0.16)

1. 分别fit后predict，最后取最小值，即可。（因为(二)中进行了log，所以预测值要exp）
2. 最终结果

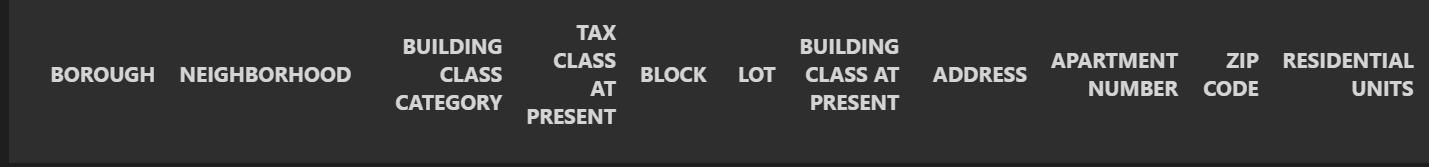


1. **探索过程及详细介绍**
2. **背景介绍**

**该数据集由大约45000个数据集和10000个测试集构成。**

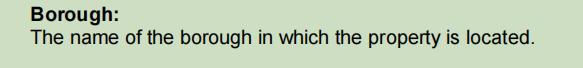
**提供了诸如“地址”，“街区”，“出售年份”等十九个（不包括售价）的信息。**

1. **数据探索性分析**





这是提供的各种信息



我们知道，地点很显然应该是对房价有相当大的影响。因此,BOROUGH’,’NEIGHBORHOOD’,

’BLOCK’，‘ZIP CODE’,’LOT’等都很有必要;但address太过特殊，没有学习的意义，二apartment number理论上不应该与房价有联系。

至于两个buildclass，两个taxclass，很显然与房子质量有关，自然和房价有很明显联系。

Saledata表述时间

三个unit和两个面积展示房子架构，都和房价有明显联系。

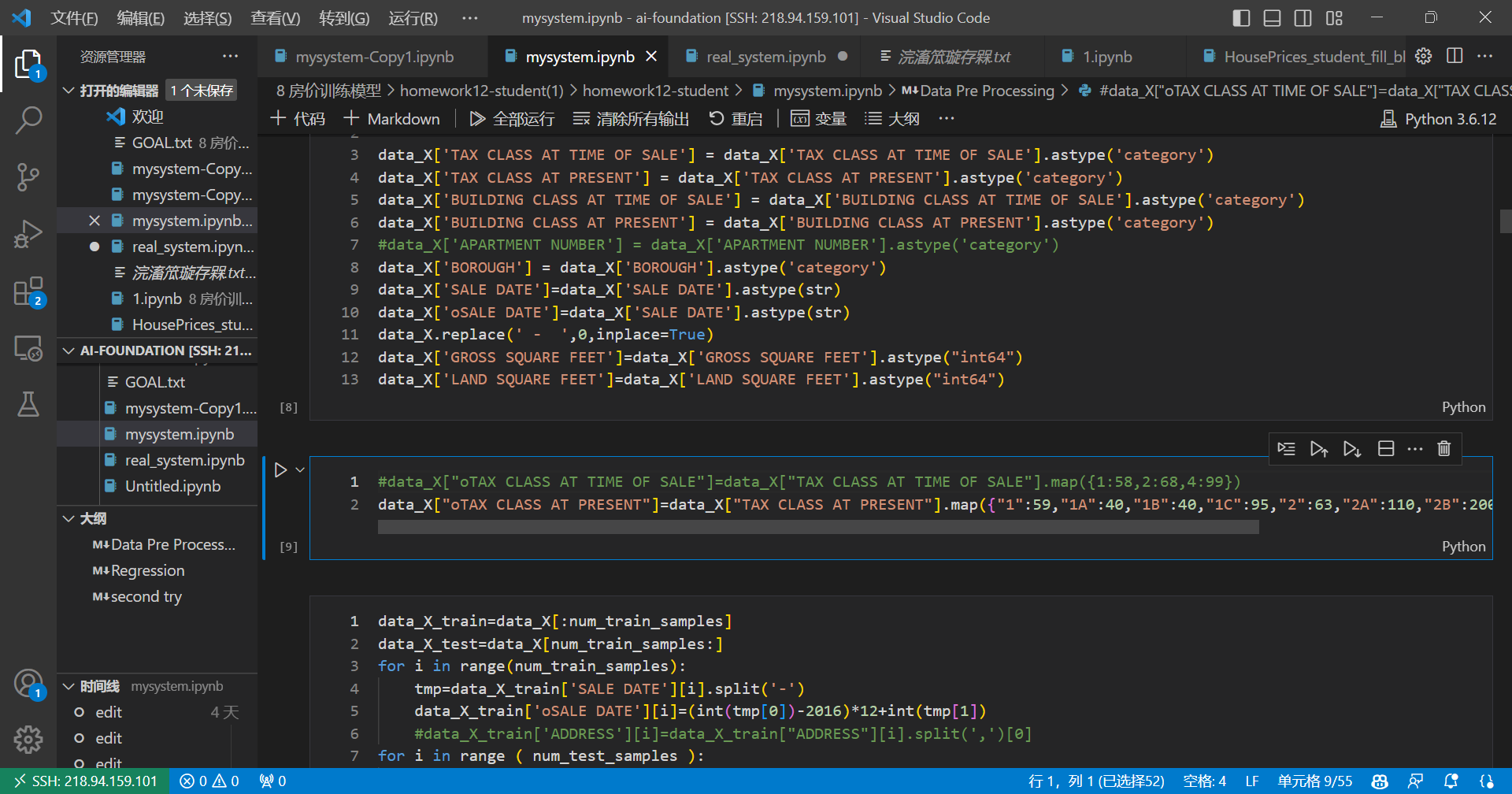
而仿照11利用可视化手段也可以看出这一点。

1. **特征工程**

绝大多数特征都应该使用category然后用one\_hot编码，但saledate我们认为随时间是有比较明显的连续性的，所以我们尝试将其转换为数字，我们尝试了年份后，发现效果不佳，原因可能是都是2016/2017两年售出的，类别太少，于是我们使用了月份，为了简明，我们以最小月份(2016年1月)为1，便于处理。

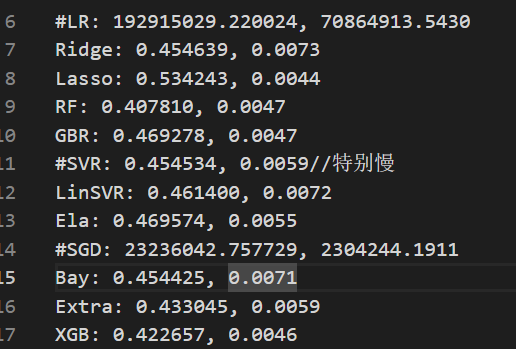
我们利用了11中按某一特征分类查看某一特征不同的值与售价的关系，试图构造一些映射函数。显然我们需要找类别比较少的那些，我们选择了taxclass;

构造了两个映射



但令人遗憾的是，第一个映射效果并不佳，因此我们将其删除。

1. **预测模型的建立**

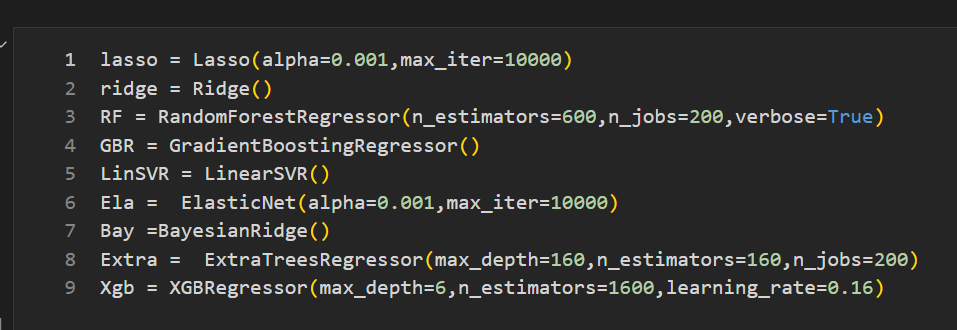


我们先利用五折交叉验证探索了模型的平均水平。

我们选用Ridge,RF,GBR,BAY,EXTRA,XGB作为我们的基础模型。

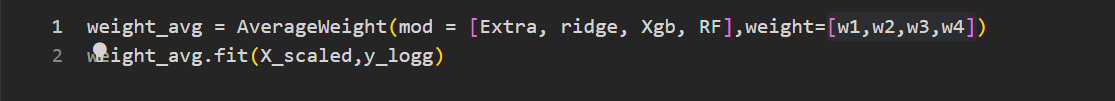
同时，我们注意到RF明显优于其他的模型，于是我们尝试对模型进行改进。

查找资料后，

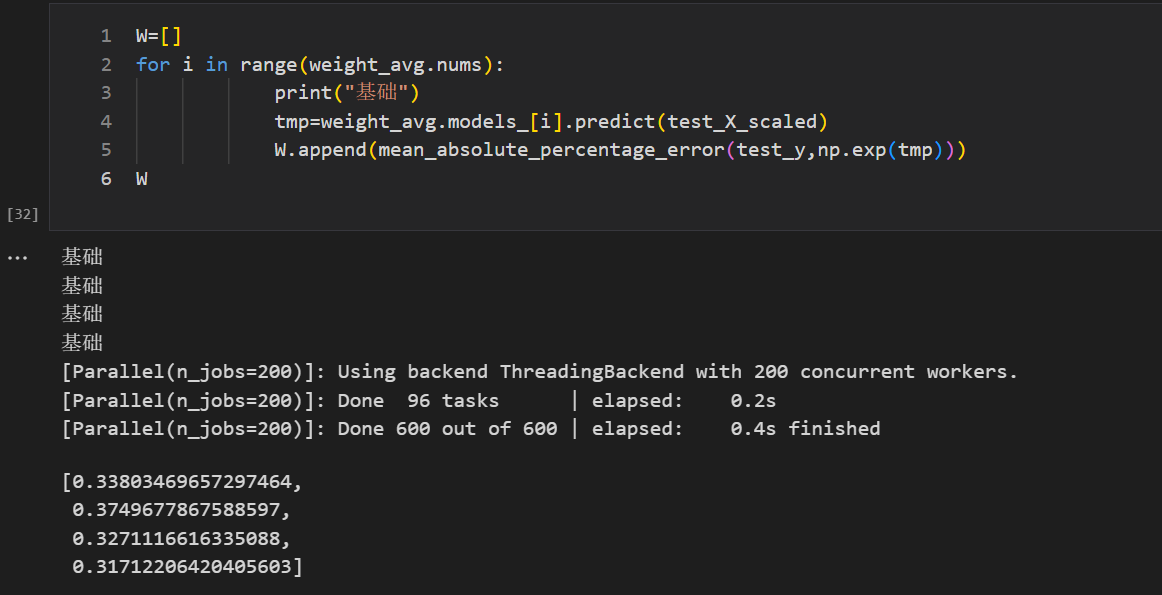


我们使用了一些参数，其中n\_jobs可大幅提高运行效率。

再进一步筛选后，我们选定了四种模型



同时我们采用了11中的transformer手法。



这是基础预测结果。很明显二完全不合格。

我们压缩为三种模型。

利用加权法，我们大约可以将结果修正至0.315左右。但我们注意到，对y进行log比不进行log要好上很多，分析可知，进行log后的结果，会更靠近较低的预测值。于是我们进行猜想，是否能激进一点，直接取最小值?实验后便得到0.3035左右的结果

1. **结果的分析与总结**

很明显，综合之前的结果，我们看出，不同模型的基准值差距巨大，而部分含参模型，则有更高的潜力。同时，要善于观察思考猜想模型的结合手段，尝试取长补短，多进行尝试，只有这样，才能取得成功。