# 房价预测(AI基础大作业)实验报告

Zhuiy presented

1. 背景介绍

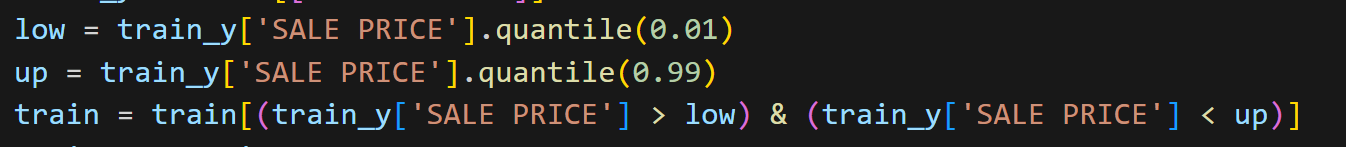
通过详细的房产交易信息对纽约不同地带, 不同规格的房产价格进行预测, 属于比较庞大的的回归任务.

作为一次大作业, 目前最好的loss可以在0.32以下.

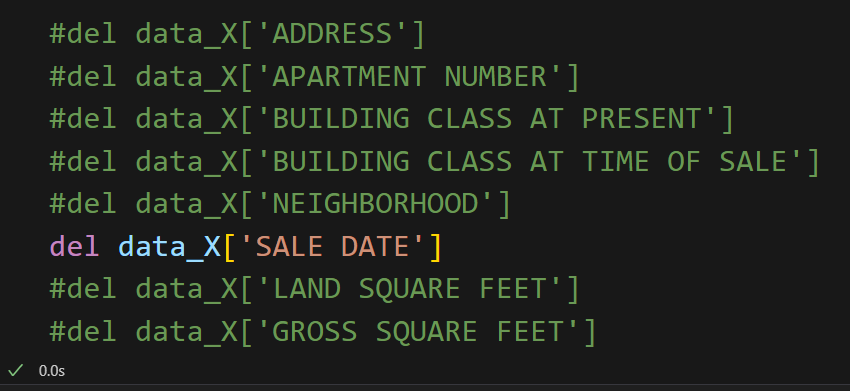
1. 数据处理

按行考虑: 空缺值过多的数据参考价值低, 可以考虑去除, 但也要加入随机性, 因为由于地区数据收集的差异, 空缺值多的数据可能集中于某些特定区块.然而最后发现, 似乎不去除的效果最好……

当然, 以上分析的时参考价值的一个角度, 还需考虑是否删除极端值, 结果表明删除更好:



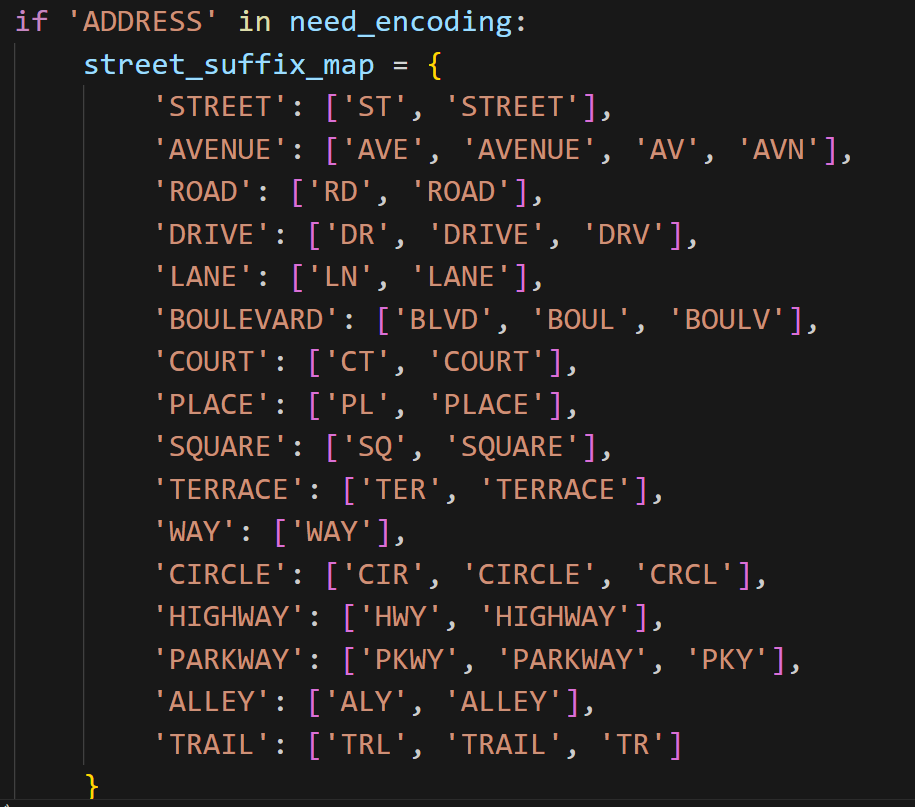
按列考虑: 没想到最后的结果表明, sell date去除有益处:



有基本序与和关系的数字数据(占地面积, 总面积, 交易时间, 总单元数等)无需编码, 使用平均值补全空缺数据(不能用test\_X来计算均值).

种类维度较少的特征使用1-hot编码, 较高的特征使用频率编码.

值得注意的是, 地区属性作为重要特征维度过高, 而且特异性明显, 故频率编码效果差, 我选择先按照街、道、路等性质分类后, 在选用频率编码(这一部分比较复杂, 应用了re包的字符匹配, 详见代码).



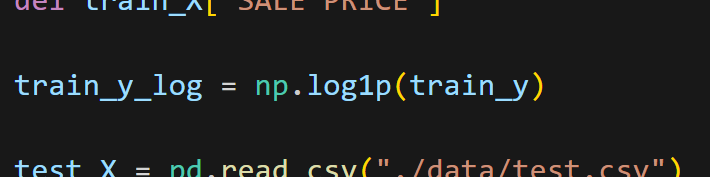
最后, 对所有训练特征进行均值归零, 方差归一化(不能用test\_X来计算均值).

具体选择那些数据进行训练, 实验表明, 虽然看起来很多数据没有用处, 但最好的训练结果仍然来自所有数据同时参与训练的模式.

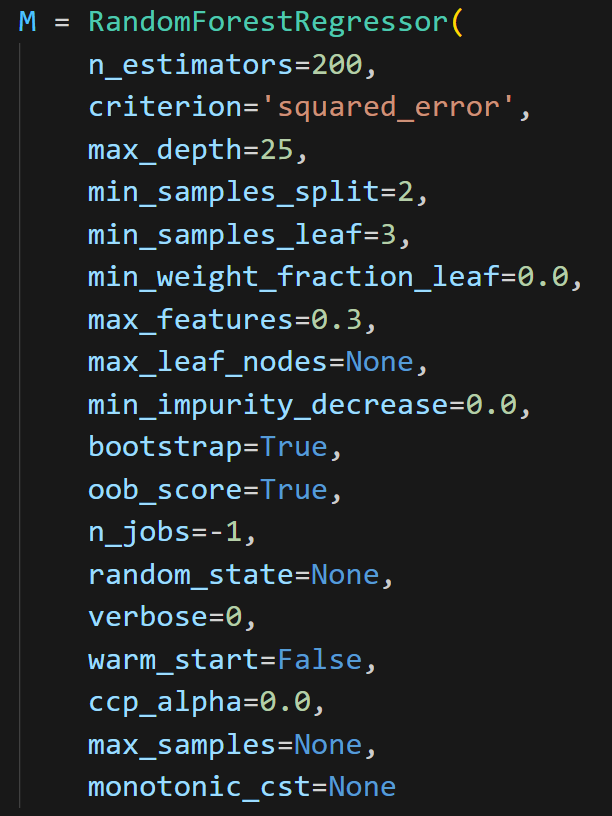
1. 基础模型选择

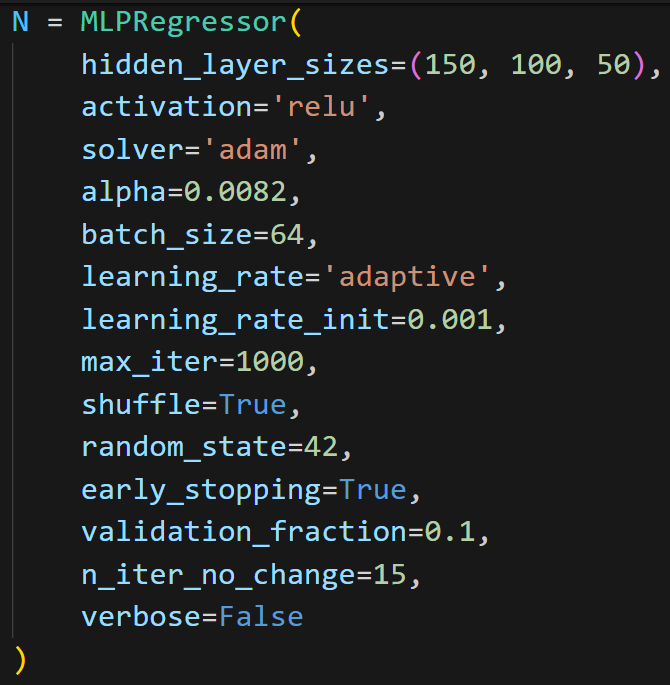
首先选用random forest进行尝试(lightlgb不如初始的模型好用), 引入GridsearchCV调参后, 最好结果能达到mape loss = 0.34, 很难再下降, 可能是由于数据处理或者模型的限制.

1. 不断改进整体流程
2. 首先发现, 房价的分布符合长尾定律, 而这种分布对于随机森林算法没有益处, 所以我尝试将房价取对数, 结果明显变好, loss下降了约0.02, 但仍未进入0.32.



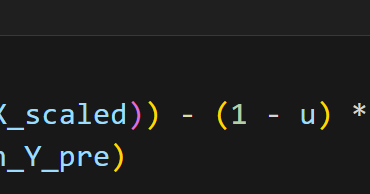


1. 尝试集成学习思路, 另加一个并行的MLP模型(通过RandomsearchCV调参), 最后进行加权平均得到结果, 然而似乎loss并没有明显改善.

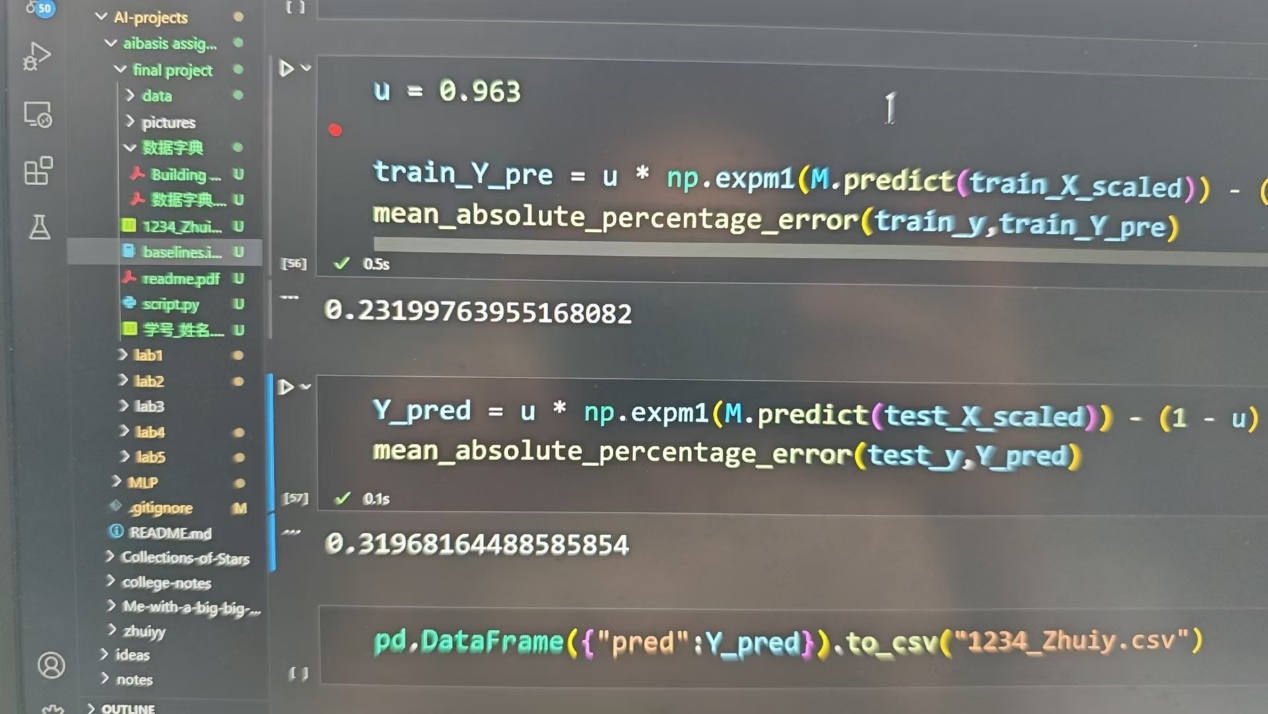




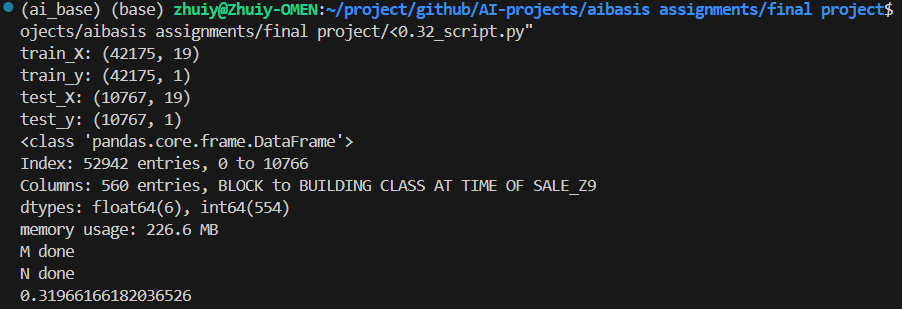
1. 这一步是最令人震惊的, 在以train\_loss为评判标准的微调中, 我发现把线性组合中的加法改成减法居然能让train\_loss得到明显下降!!!!!(实际上test\_loss也在下降, 当然这不能作为训练的标准)



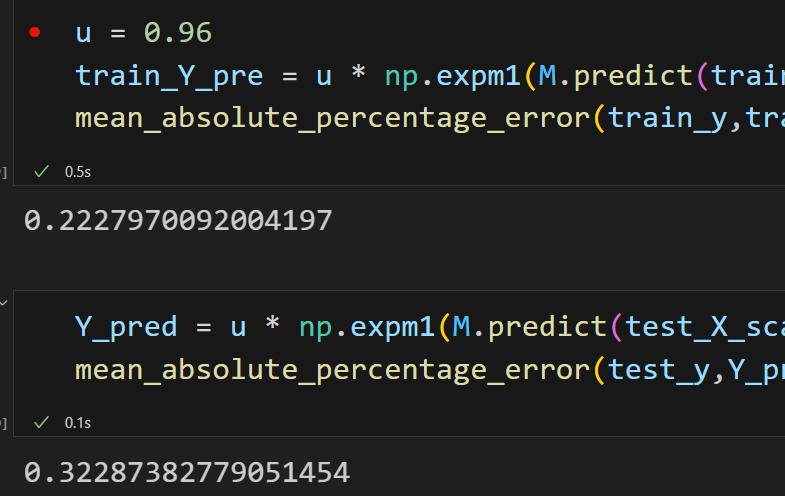
我猜测这种现象的原由是由于数据特征, RF和MLP两个模型都预测出了更高的房价, 所以应用其一作为补偿减去, 反倒能使效果更好: 最终在对u的调整下, train\_loss最低时, test\_loss正好达到0.32以下, 完成这一阶段目标:



具体得到这一结果的代码被我以.py形式保存为’0.32\_script.py’文件, 以下为.py复现, 可以使用此代码进行验证:



1. 如果想更进一步, 似乎只能从数据入手, 我突然发现房价可以取log, 房屋面积等自然数据也很有可能符合对数分布, 也该取log. 然而事实说明, 这种操作并没有什么用:



1. 附:
2. 最主要的python程序写在了baseline的jupyter文件中, 其中包含更多有价值的信息, 请参看.
3. 最好的某次结果存在1234\_Zhuiy.csv中.
4. 由于我使用vscode+wsl工作流, 内存占用大时容易断开连接, 所以加了两次内存清除流程:

