房价预测参赛笔记

经验历程：

1.对数据进行清洗:删除缺失值较多的列，对缺失值较少的列，进行填充。填充分为两种方式，根据特征类别进行区分：类别型特征将空值单独归为一类，数值型特征使用均值填充。

2.基于xgboost进行建立回归模型并预测。一开始这里疏忽了，模型用错了，导致第一版的结果出乎预料的差。更正后预测结果得分0.13945。

3.基于随机森林建模，得分0.151，同样的数据，结果不如xgboost。

4.由于预测结果分布明显畸形，使用log变换符合正态分布，预测目标使用log变换比不使用log变换成绩能提示0.001。

5.从数据方面考虑提升预测的准确率，深入探索一下数据的联系，如某些特征与预测值是否有某种相关的关系。

发现某些特征有正相关关系，分为类别型和数值型，类别型可以转化为数值直接参与建模计算。数值型可以做多项式特征构建新的特征。

6.从之前跑的模型输出的最重要的特征中，分析其分布。找到离群值去掉。

7.随机森林本地测试集RMSE比xgboost提升不少，但是提交后得分却下降0.02，怀疑是不是过拟合了。增加树的数量试试。

8.几次调整模型后分数提升遇到瓶颈，想起老师说的，大幅度的提升不是模型改进带来的，而是数据预处理，遂从数据处理方面下手。之前将缺失率较大的特征直接砍掉了，这里是不是可以做填充而不是drop。将一系列的特征进行填充，分数提升了0.006，名次提升了近300名，密度好大。

9.xgboost树最大深度一开始设置为5，但是网格搜索后发现3更合适。

10.TotalBsmtSF这个特征表示地下室面积，但是从其分布来看，不仅分布偏左，而且有大量0值，假如强行进行log变换，而0不会参与，会导致数据之间距离变小，破坏原有数据间的关系。因此，可以通过新增一维特征，来表示面积大于0。

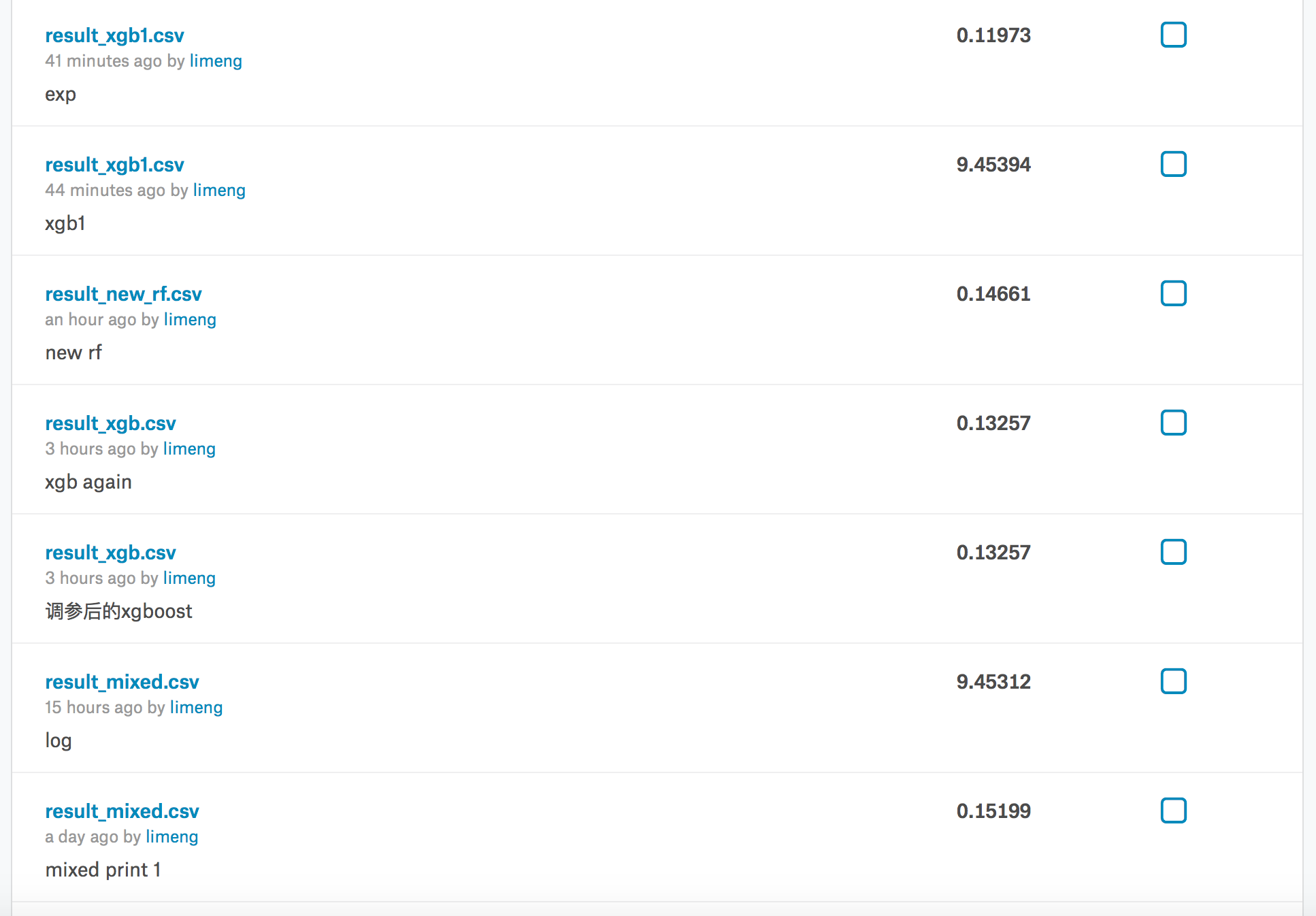
Your submission scored 0.13322, wh

11.推掉之前的特征工程，重新处理。外加参考了一篇kernal后，排名提升到700+。主要改进还是特征方面，进一步细化了重要度比较大的特征处理。

12.第11步挣扎了好久。

得分记录：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model     /  submit | xgb | rf | gbdt | ridge | lasso |
| 1 | **0.13945** | **0.15100** | \ | \ | \ |
| 2特征重新整理后 | 0.13322（1685名） | \ | \ | \ | \ |
| 3 | 0.13322 | 0.15146 | 0.14103 | 0.17275 | 0.2447 |
| 4融合 | mixed 0.15199 | \ | \ | \ | \ |
| 5 | 0.0.119733 | \ | \ | \ | \ |



参考kernel：

<https://www.kaggle.com/hamzaben/tuned-random-forest-lasso-and-xgboost-regressors>