1. 分析，理顺，项目要求并落实在文档中。20分

1.1房价预测相关需求，这个一个典型的回归类问题。通过给予的79维的数据来预测未知的房价。

2进行载入数据，清洗，建模，训练，预测等流程，并对相应步骤（及思想）进行说明并落实在文档中 50分

2.1 载入数据，info了一下数据,看到有int float object的类型，以及发现有很多是有缺省值的，有的缺的还比较多，除去saleprice有79维的数据（细看每个属性的话，工作量不小），所以从这些信息可以知道，要做的工作有1缺省值的处理，类别和数字型的，2独热向量编码是要做的，同时有的类别是int型的数据其实是类别，所以要转成str，如果是类别型的数据然后空值填充均值的话就坑爹了

2.2 数据清洗

2.2.1空值的处理

1像PoolQC这个字段有近95%缺少的情况，可以把没有的值当作一种情况，也可以完全丢弃，这里我都是将nan统一当做一种情况处理的2,另外像Electrical因为某一类的值非常多，对缺失值的填充就填充最多的那一种值，3还有一些连续的值4像GarageCars这种的没有的直接填0。5另外一些连续值的话我填充的均值

2.2.2 异常值处理

在GrLivArea 与房价的散点图中可以发现在GrLivArea大于4000的时候有两个点明显低于一般情况的，可以去除

2.2.3 数据正态化

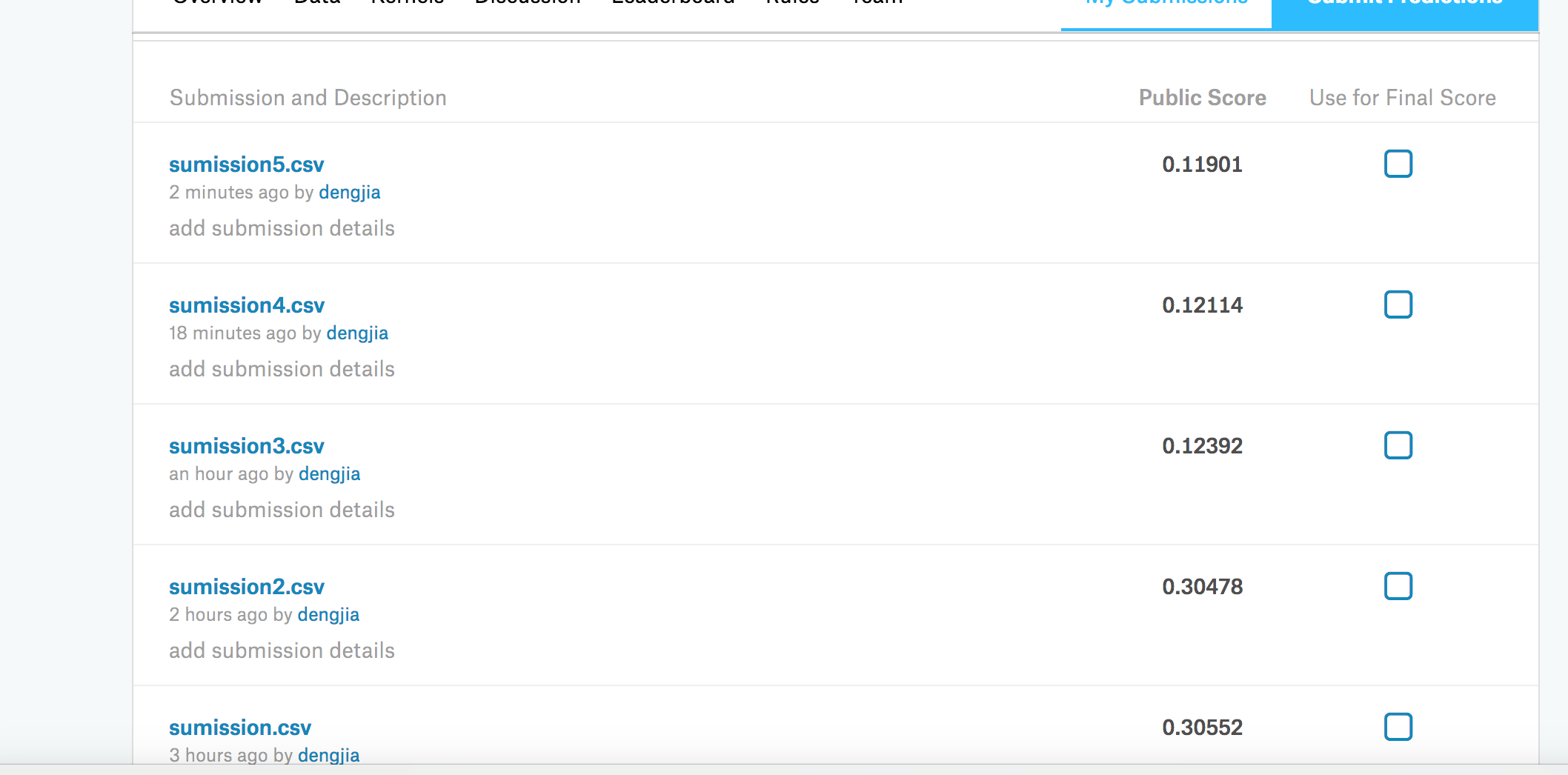
对房价做了正态化的处理,以及对一些参数做一些正态化处理

2.2.3建模

用了三个模型分别是RandomForestRegressor Ridge XGBRegressor 三个基模型来进行预测

预测结果用的是三者的平均

3在kaggle完成至少两次或以上提交。记录成绩，分析成绩差异的原因。并落实在文档中。20



第一次是做的非常的粗糙，特征工程直接str的转成类别进行独热编码，数值进行中值填充，所以出来的结果很不理想，第三次对特征做了较为细化的分析修改，其他的没改，成绩由4000多跳到1000多，第四次对数据进行了正态化的处理，成绩进一步提升，第五次，多式了几个模型取均值，成绩仅一步提升，

3项目总结。对该项目做最后总结，可以是心得，也可以是教训。 并落实在文档中。10分

本次还是有很多两点还没时间去做，一个是将重要的特征进行维度扩增，2维吧可以先试下，如果过拟合的话，可以用特征筛选的方式筛出部分特征。第二个就是模型的融合，这块做的还不够，像blending boosting bagging等，应该成绩还能够进一步提升，

4如在项目文档中参考引用了其它kernel。额外10分（但总分不超过100分）