**文档：**

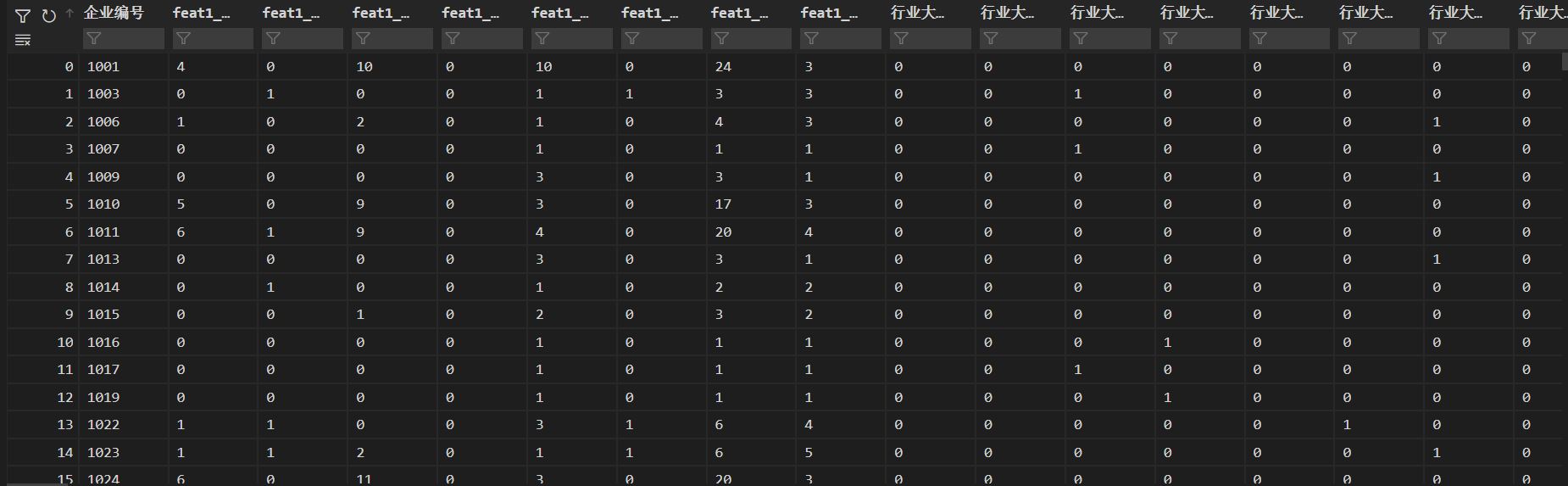
Model.ipynb

**文档说明：**

清洗后的数据整合统一，训练模型，测试集的结果输出。

**数据整合：**

将清洗后的数据进行整合，得到适合模型训练的数据集，数据集的一部分数据如图所示：

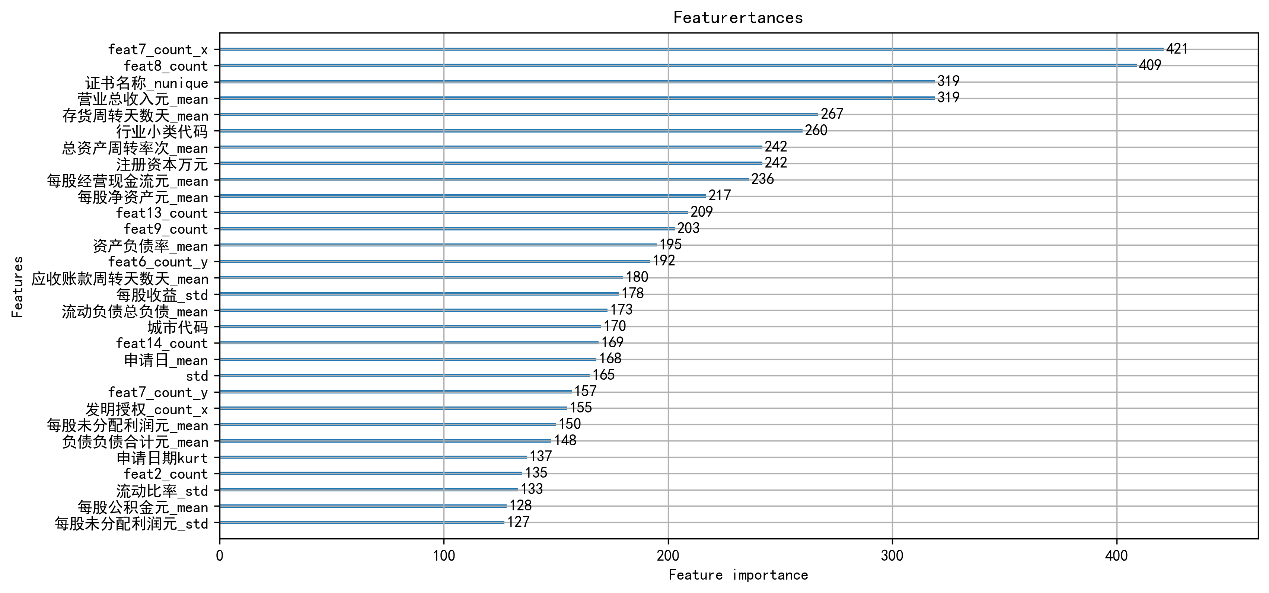


**图1 清洗后的数据合并**

**特征选择：**

使用recursive feature elimination and cross validation(RFECV)进行特征抽取，抽取了439个特征作为最终特征输入。

所有选择特征的重要性排名见附件，下图为重要性排名前30的特征：



**图2 重要性排名前30的特征**

**模型算法思路：**

1.模型选择：

集成学习(stacking)：

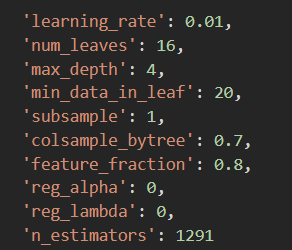
基础模型：Gradient Boosting Decision Tree(GBDT), Xgboost, lightGBM

上层模型：LinearRegression

2.参数调整：

Step1：根据特征和样本数量适当设置每一个备选模型初始参数；

Step2：根据GridSearchCV查找每一个备选模型的最优参数(开始设置较大步长，再逐渐缩短范围)，如下图为lightGBM模型的最优参数，其他模型的最优参数可以在文档中查看；

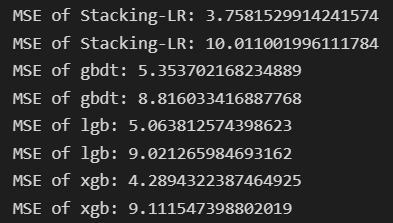


**图3 通过网格搜索得到的lightGBM模型最优参数**

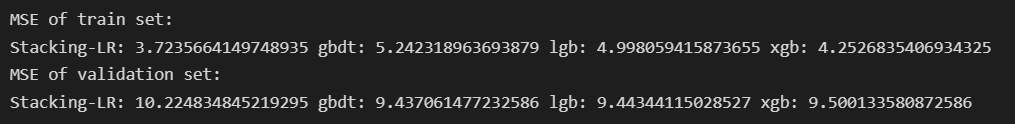
Step3：根据每个参数的作用进行人为调整直到RMSE收敛。

3.模型选择标准：

设置随机数，设置了最优参数备选模型跑十次，综合考虑测试结果稳定性以及RMSE均值。根据稳定性以及RMSE均值相对最优选择模型——lightGBM，各模型表现如图4和图5所示。可能因为基础模型的相对稳定，集成学习在此次结果中的表现并不是最优的，具体原因还需要进一步探索。



**图4 一次交叉验证中各模型的表现(第一行：测试集，第二行：验证集)**



**图5 各模型在交叉验证中的总体表现（取MSE均值）**

**探索历程和心得体会：**

1.探索历程：

模型的选择中，查阅相应文献，根据数据本身的特点(特征多，样本少)选择初级备用模型，初步测试初级备用模型的表现，再选出备用模型，进行调参，整个过程查阅许多资料，了解很多模型的深层含义。

2.心得体会：

本次题目数据量不大，但是非常“脏”，并且特征多，人工进行特征工程很难寻找到所有的特征，因此要想进一步提高模型准确率需要在特征工程中下更多的功夫。以前效果良好的集成学习在此次题目中发挥并不是很理想，具体原因还需要进一步的探索。