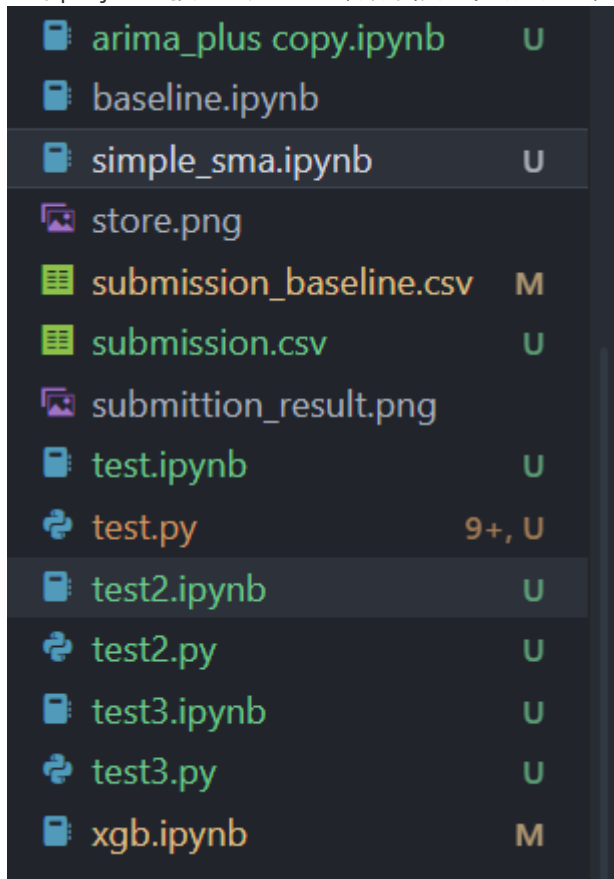


Data Mining Project 1 时序数据预测

贡献总结

1. 发现了助教所给出的baseline中, 学长使用了数据集中出现的cluster来进行先集中后平均的操作, 察觉到了这种操作方法可能会掩盖掉各个商店的独特性, 因为数据最终要求预测的是以每个商店每个种类作为单位, 因此舍弃掉了学长提供的baseline, 选择自己重新书写一个baseline
2. 尝试了不下三种深度学习的方法并在本地跑了模型, 但是之前确实没有接触过很多AI相关的内容, 在这个project上投入了超过25h, 效果依然不尽如人意, 以下是各种尝试的文件目录...



3. 尝试使用了油价和节假日数据进行模型构建, 一开始在学长提供的baseline基础上进行尝试, 发现收效甚微(<0.05), 后来换到自己写的baseline上进行尝试之后发现反而使模型的表现大大下降了, 于是决定不使用这些数据, 最朴素的方法效果反而不是很差

技术方案

以每个商店为单位进行SMA预测

```
#存储最终预测结果
forecast = pd.DataFrame()



for store in set(train_df['store_nbr'].unique()):
    for family in train_df['family'].unique():
        #每个商店中的每个类别进行循环
        series = train_df[(train_df['store_nbr'] == store) & (train_df['family']
        == family)]
        #提取销售数据
        sales = list(series['sales'])
        #对未来16个时间点进行预测
```

```

for i in range(16):
    sales.append(mean(sales[-7:]))
#预测的销售值赋值给sales列
    target = test_df[(test_df['store_nbr'] == store) & (test_df['family'] ==
family)]
    target['sales'] = sales[-16:]
#连接，存储结果
    forecast = pd.concat([forecast, target])

```

实验分析

274			0.46865	3	38m
 Your Best Entry! Your submission scored 3.64169, which is not an improvement of your previous score. Keep trying!					
1	vngtono		0.37770	13	2d
478	meilulu		0.59890	2	7d
479	ziyang xiao		0.60028	1	1s
 Your First Entry! Welcome to the leaderboard!					
480	何勤廷 HE,JIN-TING		0.60889	8	9d
481	Stephen		0.61587	6	1mo

我的结果相比于助教的结果提升了25%的表现, 与前排差距不到0.1, 表现还算中规中矩, 主要是算法很简洁, 努力了这么久最终却用了最简单的方法, 心里还是有点小难过...