2024秋 大数据技术与应用 课程大作业

电 25 吴晨聪 2022010311

# 思路阐述

在完成这项大作业的过程中，所做的最多的事情就是寻找模型和参数调优。其中模型的选择可以带来最直接的改变，好的模型无论是求解精度还是求解速度都会有较大的提升。而同一个模型内部又有很多的参数可以选择，对于不同的训练集以及测试集，都有其各自最佳的参数，这就需要经过不断的尝试来提高精度。

# 调试过程

# 2.1 模型选择

首先采用的是Randomforest的方法进行分类和回归，再利用网络调参的办法分别对其n\_estimators、max\_depth、min\_samples\_split、min\_samples\_leaf和max\_features进行提调参。采用同一种模型进行分类和回归的好处在于在对分类完成参数调优以后，其将其参数带入到回归模型中也往往会在一个局-部最优点的附近，这样对于寻优的效率可以提升。

重复几次调优后发现已经落入了几个局部最优点附近，再尝试调优改进不大，于是便选择尝试别的模型。此后又使用LogisticRegression、LGBMClassifier、MLPClassifier等模型进行了分类的尝试，对RidgeRegressor、XGBRegressor、MLPRegressor等模型进行了回归的尝试，每个模型通过不断地参数调优可以得到更好的结果，而多次的测试下，最后的模型选择为MLPRegressor。

# 2.2 数据预处理

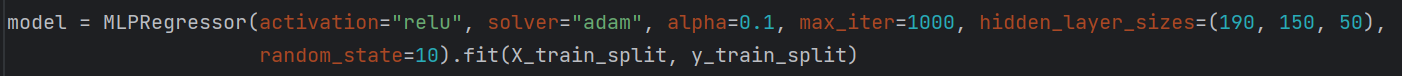
(1) 采用了计算参数和结果间相关系数的方法来观察特征与结果间的重要程度，并按去掉相关系数较低的一些特征。发现计算速度因为特征的减少得到了提升，但精度也有明显的下降。最终为了保证足够的精度，不批量删去特征；

(2) 归一化处理。因为各特征的取值范围大小不一，所以采取了归一化处理的方式将所有特征的取值变为(0,1)之间，但对于预测精度而言并没有太多的改变；

# 2.3 调试方法

第一周: 本周的主要完成工作为初步建立好模型的框架，包括数据的读入和按要求的格式输出、对数据进行初始化、处理时间列、填充缺失值、并且训练模型，以及通过均方误差来简单判断训练结果的准确性，而数据特征值只选择了'demand', 'capacity', 'type'这三项，原因是给定的预测集只有提供了这三项数据，而目前也未能联想到如何利用其他'price\_band'和'volume\_band'的特征值来帮助模型更好地学习。

在测试过上文所提及的众多模型后，最后使用了MLPRegressor模型，此模型的主要参数如下:



由于时间原因，初步调试后把此模型的神经元层设定为(190, 150, 50)，而此时得到的均方误差为4751.132261817674，在第一周的动态排名上为第11名，分数为0.52489108。

第二周: 本周在第一周的模型上利用for循环开始不断对模型进行参数的调整，主要为数据读入的数量大小、MLPRegressor模型中的激活函数、求解器类型以及最为重要的神经元层设定。

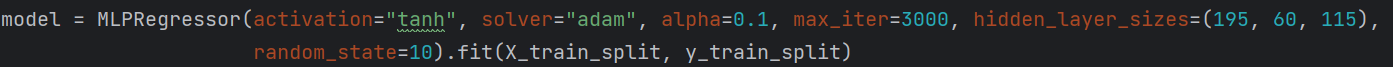
最后结论是给定数据集的数据量实在十分巨大，并且由于数据的离散程度也相对较大，把数据全部读入给模型进行学习并不能得到一个很好的预测结果，经常产生过拟合的情况，同时也因为这个原因，让模型学习第一周测试集的结果也并不能让模型有明显优化，所以这周的训练集数据只取了data\_2021和data\_2022的前1200行，这是能让模型训练出较好结果的数据读入量。

同时，还需要对模型的求解器类型和神经元层进行反复调试，发现参数为以下值时:



其均方误差为953.7367446407645，较第一周有十分明显的进步，此模型训练的结果是本周在动态排名上取得的最优成绩。但同时也发现了在其他参数的组合下其实有更低的均方误差，但把这些数据上传到网站上与答案对比却没有获得更好的分数，这个现象也说明了均方误差只是评定模型准确性的其中一种标准，并不能完成反映出模型的优劣。而最后在第二周的动态排名上为第15名，分数为0.56854152，虽然排名有所降低，但模型自身的精度还是有了十分明显的提高。

第三周: 本周由于对模型的参数调整其实已经到了瓶颈位置，为了能更好的优化模型，我向班里的同学们请教了一下。在讨论过后，最后决定尝试只把前两周的答案集读入训练集来让模型进行学习，结果发现模型的精度并没有得到很好的提升。因此本周便把思路是对模型参数调整的基础上更优化对数据读入的预处理。最后在试过多种预处理方法和模型参数调整的双配合下，模型的均方误差能达到45左右，这个数字理论上能反映这个模型已经有十分优越的精度，但事实上也是和第二周一样的问题，均方误差并不能完成代表模型的精度，本周取得最优成绩的反而是均方误差较大的参数:



此模型的均方误差为731.1410078931026，并且能发现可能是因为给定模型的特征值只有三个，数量太少，以致于即使对数据进行了不同方法的预处理，精度也没有明显的提高，甚至有些预处理方法会让精度大幅下降。最后在第三周的动态排名上为第18名，分数为0.56854152，虽然排名比上周又降低了一些，但本周对于模型还是用了不少方法去提高精度。

# 心得体会

经过三个星期的模型训练，得到的结果是在一周周变好并且一次次在认为要达到极限的时候超出了自己的预期。总体而言，对于预测而言最重要的是前期思考好大的方向，包括模型的选择以及对数据的分析，因为无论是哪种模型最后都需要花大量的时间来训练得到最优的结果，如果一开始没选择对方向那后期大量的花费在模型训练上的时间都是白花的。

此外数据本身的理解也是很重要的，任何一个模型要达到最好的效果肯定得先对数据的含义有深刻的理解。很可惜在完成这一次大作业的时候所用到的一些数据预处理的方法都没有对最终的预测精度有很大的提升，但这也仅仅是从精度的角度来看，如果考虑实际应用，若牺牲一点精度能节省大量的训练时间那也是成功的处理，所以数据的理解和预处理还是十分重要的。

# 动态排名

第一周动态排名:



第二周动态排名:



第三周动态排名:

