第六赛道：大数据建模-交通赛题

说明文档

1. **队伍介绍：**

学校：北部湾大学

队伍名称：hello\_world

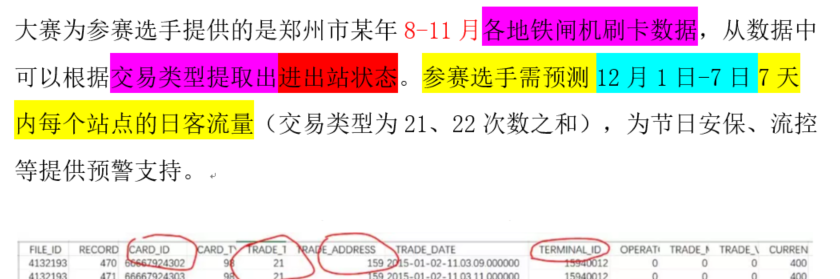
队长：\*\*\*

队员：\*\*\* , \*\*

**二、建模说明：**

随着我国轨道交通的建设发展，地铁已经成为公共出行的重要方式。在节假日、比赛、演出等重要事件发生时，地铁客流会面临巨大压力。我们获得大赛主办方提供的是郑州市某年8-11月各地铁闸机刷卡数据，根据数据建立模型预测12月1日-7日7天内郑州每个地铁站点的日客流量，为节日安保、流控等提供预警支持。

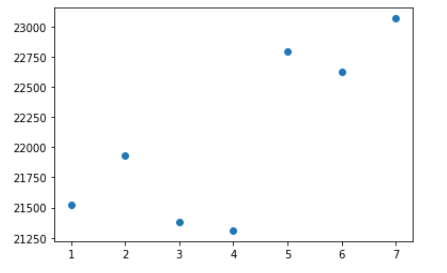
1. **解析题意：**



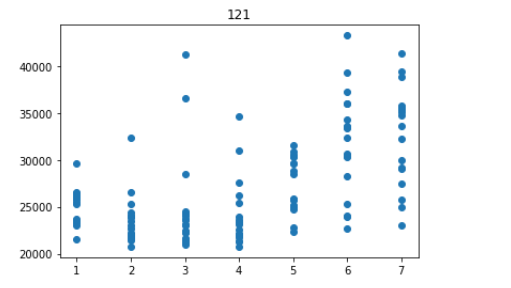
1. 目的是 预测12月1日~7日这7天内每个站点的日客流量,为车站提供预警的参考以及准备的支持。
2. 客流量可能会受天气、所处地域、时间、其它等影响，但题给出的是闸机刷卡数据，也就是说有些数据字段是对于结果没有什么影响以及特殊影响,只是达到了记录的作用。比如交通卡旧类型、持卡人、读卡器文件名等等。
3. 此外，题目还出了如何计算日客流量的方法。
4. 由于数据的字段内容有限，且有诸多字段名称是对于客流量来说是相关性比较小的。
5. **数据清洗与探索过程:**

原始数据是8-11月共14.9G的csv文件。

1. 根据建模需求，经过**MapReduce**转换汇总得到每天每个站点的人流量汇总，共2440行数据。包括时间（年月日），星期，车站编号，当日人流量汇总，共四个字段。
2. 把MapReduce处理得到的数据转换成dataframe并拼接数据,首先通过map阶段洗去没有用的字段,封装成(日期,车站编号,1),再通过reduce汇总同一天同一个车站的总人流量(日期,车站编号,日流量)。
3. 我们对于模型的选择思路:(1)根据现有线性模型进行构建(因为这并不是一个分类问题，而且所给特征与目标也不符合其它算法类型的任务要求、目的),并在有必要的情况下采取模型融合的手段完成任务； (2)思考各变量之间的联系和关系，在现有原理和知识上进行公式联系以得出结果；
4. 我们先采取现有模型建模，如果现有模型无法优良的解决问题，再采取第二种方式。为此，有必要采取试探性的深入探索。
5. 对数据进行线性模型的探索和尝试。
6. 对现有数据进行绘制单个车站单周的散点图以进行观察数据，之所以绘制单个车站单周的图，一是因为联系现实之后，可以发现各个车站之间有极大的可能性会因为车站所处地域不同而导致客流量的级别也就大不相同；二是因为：假如把所有的点绘制在同一张图上将不利于查找一周的联系，至于为什么是一周，根据一周观察单个图(如下)，有可能性是存在某种线性关系的:

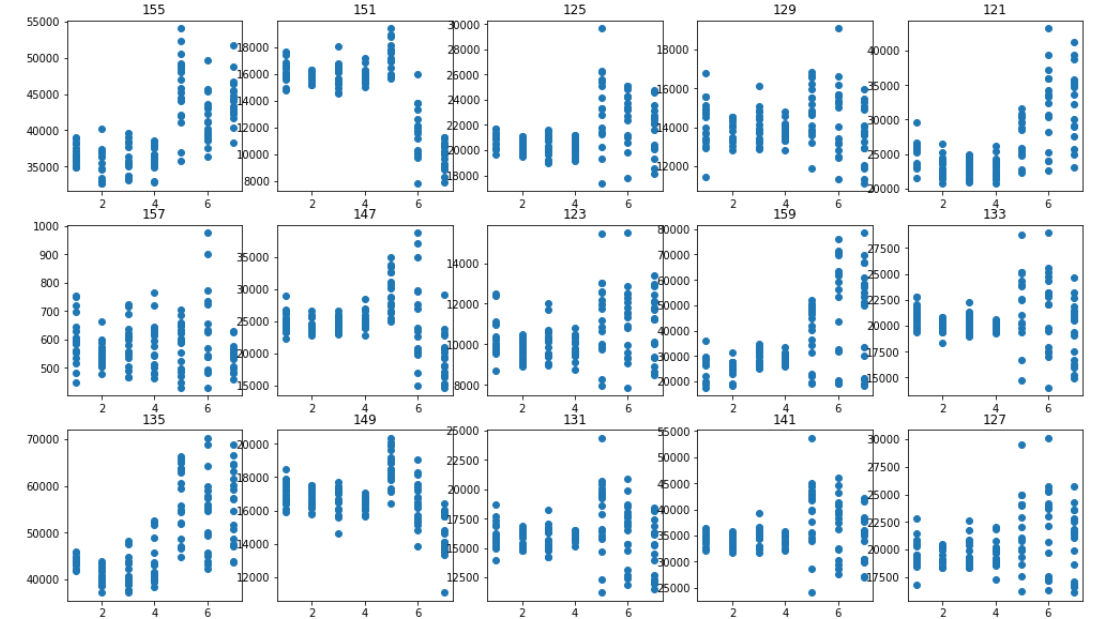


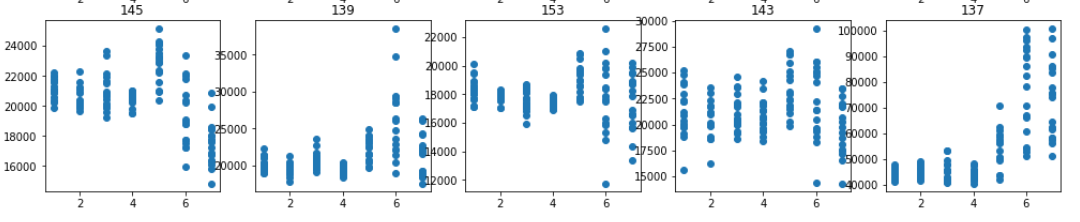
把单个车站的散点图画出来:



可以看到该车站在每一个天(这里的天 指星期几)里面都有某块比较密集的点的范围，当然不是100%的每个天都有,这主要是受到众多不确定性的影响。

绘制所有的车站来观察其点的分布:

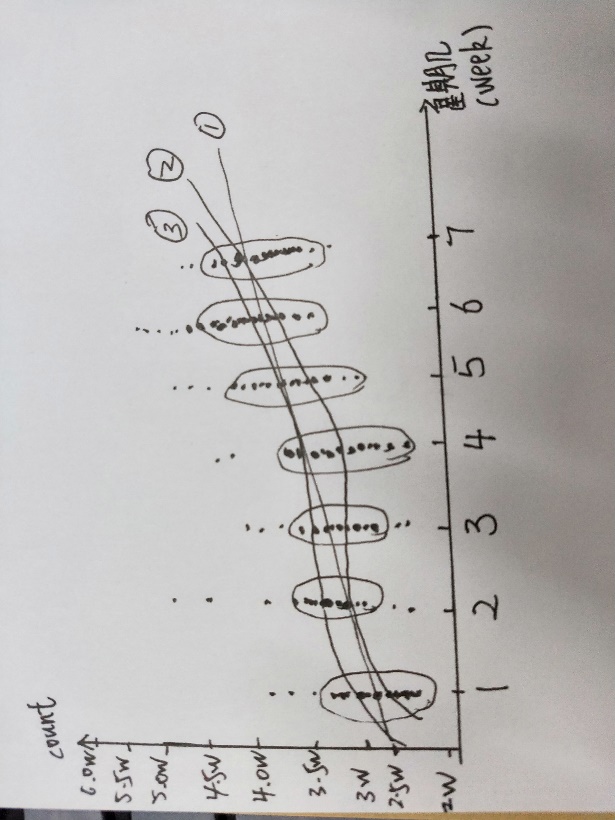




可以知道的是每个车站在每个天里面几乎都有自己的点的集中分布，而在有些天里面比较离散的点则是由现实生活中众多因素导致和影响的。

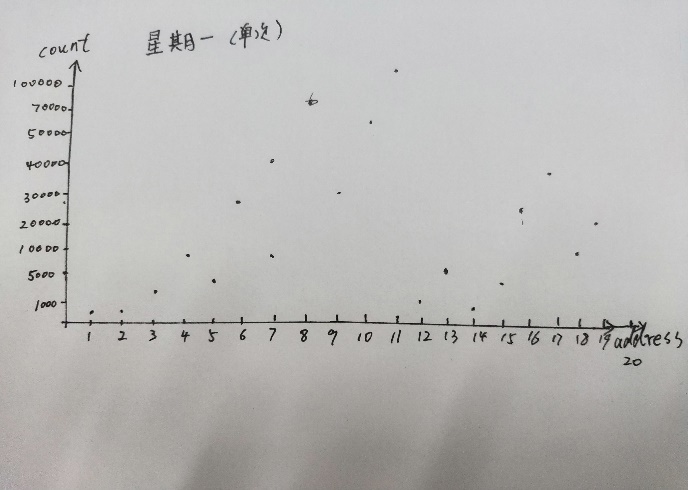
但每一个天都由其各自集中分布趋势的时候，可以将这些点集其视为一个线性关系，因为当多个点都在一定范围内的点集时，可以明白其点集里面点与点之间具有共性，也就是: 假如有真实值点a和点b在一个点集里面，其预测值A和B，那么(a-A)和(b-B)之间的差值将不会很大。

所以可以考虑按照单个车站按7天一周次来进行线性拟合:如下图

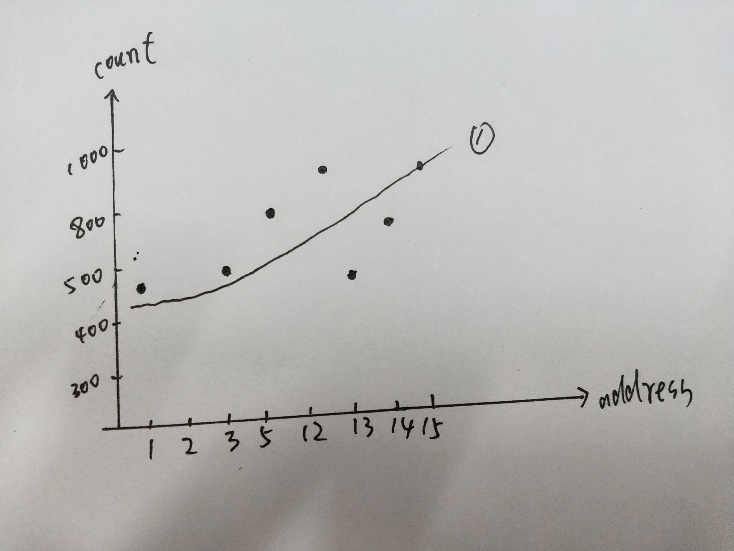


如上图,、、这三条拟合线(当然会有更多的),中会寻找出一条进行最佳拟合；但是此做法(以下统称此做法为**线性1**)会导致忽略了车站与车站之间的联系和会受到不在点集里面的点的严重影响，从而导致拟合的不是很好。

当然，上面是以车站为对象；如果以天(星期几)为对象那么将会解决掉第一个问题，因为这时候车站与车 c站之间的关联已经在里面了(以权重的形式)，如下图: 可以看到点的线性分布的概率很小，因为受到个别车站的影响，由于一条公式里面参数过多而导致容易出现过拟合！



因此，将具有同一数量级的客流量的车站划分到同一组里面进行建模。故将此做法记为**线性2**.



可以看到其线性关系明显提高了许多，没有绝对的事情，实践是检验真理的唯一标准，后面将采取构建多模型,选取最优模型进行预测。

**线性1和线性2后面都会用到，后面就直接用这两个名词替代这两种思想了。**

1. 观察散点图，可以发现有类似线性关系的痕迹，却又不像是，因为存在着一些离 集群 很远的点，而这些点成为了当前的点的分布的异常点，但不幸的是，这些异常点确是真实合理的存在以及影响模型拟合效果的因素。
2. 根据上一点，按照常理来说，我们需要 和谐 这个异常点，但又不幸的是异常点也属于我们的预测目标之一，题目的目的不仅仅进行预测以及提高准确率就ok，而是要尽可能的降低泛化误差(这里可以将泛化误差指定为题目要求的评价方法即可、即平均绝对误差)、同时异常点也不可忽略(因为联系现实容易发现:不可能每个车站每一天的客流量都是处于平稳状态的)，很显然，这两者是矛盾的！因为异常点的存在会导致模型拟合效果不好，出现欠拟合现象,也就是会产生训练误差很大而且泛化误差也很大的情况；而异常点被直接和谐的话又可能会出现过拟合现象，即训练集表现良好，但泛化误差离奇的大。
3. 因此，由上一点可以得出解决方法:(1).对异常点进行恰到好处的处理，即对异常点进行既不会导致过拟合又不出现欠拟合现象的处理(当然，这个方法是不存在的或者说是难以发现的，因为我们只有一点数据，而且现实生活的随机性和偶发性始终有一天会打破这个 恰到好处的处理)。(2).对异常点单独建模拟合得到a模型，对处理过异常点的数据集进行线性拟合得到b模型，对于这两个模型，现在都是优秀的，之后进行模型融合。假如a模型在测试集上预测得到某个结果A与b模型在测试集上预测得到的某个结果B的距离过近，直接取A和B的均值作为预测结果；假如AB之间距离过远，那么就是说出现了极大异常或者极小异常，将A作为预测结果。当然采用模型融合且提高模型融合效果的最佳方式是数据量足够大。
4. 当然，以上全是基于数据集的分析。
5. 对于**异常值的处理**:采取单独对车站进行按星期几处理的方式来处理。也就是处理方式是:在a车站中对星期一处理掉星期一的异常的数据，再对星期二处理掉星期二的异常的数据、依次类推。处理规则是: 使用0.25分位数(记为q)和0.75分位数(记为Q)对数据进行判断，如果大于(Q+1.5\*(Q-q))记为极大异常，如果小于(q-1.5(Q-q))记为极小异常；全部使用平均数代替，这其中其实如果加上一个个数K作为选取点的参考，并求出目标点对K个点的马氏距离,再做平均距离,根据这个平均距离与阈值的对比的话,就加入了KNN的思想，当然我们并没有那么做，因为这样做的迭代运算估计会达到n\*n-1次极大的消耗电脑的IO和硬件寿命。可以知道的是这类处理方法非常简单，但效果不是特别的好，假如采用SVM进行检测估计效果会非常好，但是也就变得极其复杂。
6. **总结模型的选择**：

(1)L2正则的线性回归模型；

(2) xgboost.XGBRegressor线性模型。

(3)时间序列分析。

分别建立这三种模型，训练预测以及进行测试集上的MAE评估，然后对比结果，选取MAE最小的作为最终模型选择。

**四、建模过程：**

1. 模型建立：

(1)、线性回归模型的建立:

将week属性和address属性分别进行one-hot编码处理，即构造0-1向量(暂时不考虑稀疏的问题)，以count为label。

也就是得到：



这里的X是**one-hot编码**组成的0-1矩阵。

即



目标函数:(L是L2正则项的系数)



即

接下来就是展开平方项，化简求导，令导数为零，即可求出回归系数O。O是一个回归系数矩阵。

也就是:



求导并令导函数为0:



所以,,也就求出了回归系数矩阵，思路与多元线性回归类似，只不过多了一个正则化项的限制。

X是训练数据集，L是给定的，I是单位矩阵，y是训练集标签。所以，O可求。

其间那些复杂的矩阵运算过程将由计算机来进行计算。

最后输出预测结果和在测试集上的MAE值!

(2)boosting---XGBoost(XGBRegressor)的引入:

对于X、y的处理与之间相同。

XGBoost核心思想是多个基础模型的线性拟合，基础模型使用CART树，因为CART树普遍来讲要比线性基础模型的效果要好。

首先，xgboost采用的是加法训练，也就是要确定第t颗树最优，先确定第t-1颗树最优，依次类推。

所以，其目标函数: =

表示经过第t轮迭代后的模型预测值，表示已知t-1个基础模型得到预测值;表示第t个基础模型(,w是叶子的向量、q是树的结构), **寻找一个使目标函数尽可能最大化降低的**,那问题就解决了！因为前面的t-1颗树这时候已经固定了,找到****之后，模型主要求解部分基本结束，再把正则项化简，即可求解。为第j个正则化项，正则化惩罚项的目的是防止模型的复杂度与防止过拟合。

根据泰勒展开式:

假设为泰勒公式中的,,L为泰勒公式中的f，损失函数的为泰勒公式中的x，那么对于目标函数的求解一切将清晰明了:



也就是说只要定义的损失函数可求一阶导和二阶导，**那么只要再求解正则化项出来**，那么函数求解即可结束，即可开始考虑树结构的分割规则。

当然，**如果**使用最小二乘法作为损失，求导过程也就可以跳过，因为使用最小二乘法化简之后，一切将变得更加简单**:**



当然上面只是一个举例，其它的类似，虽然方法可能不再是最小二乘，但想法都是要把其进行化简,方便后面消去。

其中，对于来说，是已知的，因为这个原理就是利用牛顿迭代法迭代出来的，只要第一个知道，后面迭代运算即可。由于本质上是残差的平方，所以该项是一个常数。令其为C即可，后面再有常数直接与C合并即可，对模型并无影响。

Xgboost的基础模型采取CART树，其实不管是预测还是回归，CART只有一个原理，只是在方式上有所不同，分类的时候采取了投票原则，回归的时候将叶子节点中的样本均值作为了该节点的预测值！对于回归，可以这样认为：假如自变量中X有n个观测，则计算相邻两个观测之间的均值,也就是说有(n-1)个均值,以均值为判断值将数据分为两部分,为n1部分，反之为n2部分，这时候n1和n2中都分别含有两个分类的样本量，基尼指数下降速率为:

基尼指数下降速率表示了自变量对因变量的影响程度，下降越快影响越强。

CART树的字段选择指标是基尼指数,由基尼指数出发可以得到条件基尼指数，条件基尼指数采取二分法原理,以上用的概率都是用经验概率替代：





:某事件第k个可能值的发生概率，使用经验概率代替;

表示A变量在某个二元划分下第i组的概率，经验概率为A中第i组的样本量与总样本量的商。表示内变量D取第k种值的经验概率。

|D|:表示事件中的所有样本点,表示事件的第k个可能值出现的次数。

该树由结构部分q和叶子节点对应的输出值w构成()，T表示叶子节点的个数,则树的复杂度:

也就是说,T越多，树生长的越复杂。γ和λ是xgboost自己定义的，可以设定它们的值，γ越大，表示越希望获得结构简单的树，此时对较多叶子节点的树的惩罚越大。λ越大也是越希望获得结构简单的树。

将目标函数改写:



其中，表示每个叶子节点j中所包含的样本集合，表示第i个样本点的输入值对应的输出值。这里一阶导使用代替，二阶导使用代替，因为mathtype编辑器方便但也不太好写。

一样的道理，求偏导！对求偏导，并令导函数为0。即:



所以,

将代回，得:



使用贪心算法来决定是否剪枝。即在某个已有的可划分节点中加入一个分割，通过计算分割前后的增益值决定是否剪枝。这里不必过多介绍，因为数据集太少，根本无需剪枝，剪枝是一门学问。

(3)时间序列分析建模:

如果画图之后，时间序列不是平稳的，那么用拆分法将其处理成平稳的时间序列。

差分法:时间序列数据在t与t-1时刻的差值。（）,依次类推。

**自回归移动平均模型(ARMA):**

公式定义:

i是将时间序列处理成平稳时所做地拆分次数。

对于参数的选择我们使用自相关函数ACF和偏自相关函数PACF来确定参数:



PACF则是将ACF的去除前t-k-1个对t和t-k的二元独立相关性的影响。

基于ACF和PACF来确定p和q:

|  |  |  |
| --- | --- | --- |
| 模型 | ACF | PACF |
| ARMA（p,q） | q阶后衰减趋于零(几何或震荡) | P阶后衰减于零(几何或震荡) |

那么将会有20个模型被构建，因为有20个车站。

1. 模型训练：
2. RidgeCV：

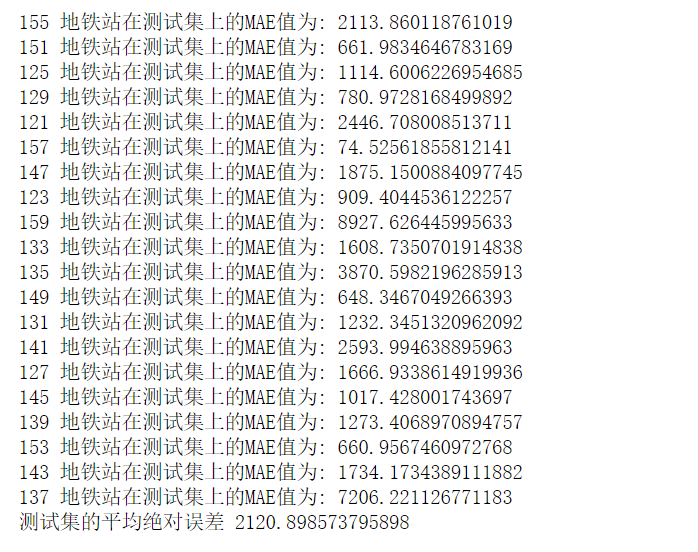
做法1和做法2的区别主要是建模对象的不同，做法1关注的是单个车站的时间联系，做法2关注的是车站之间以及时间联系，也就是X数据集的不一样。采用自带的交叉验证确定正则化系数。

1. XGBRegressor:

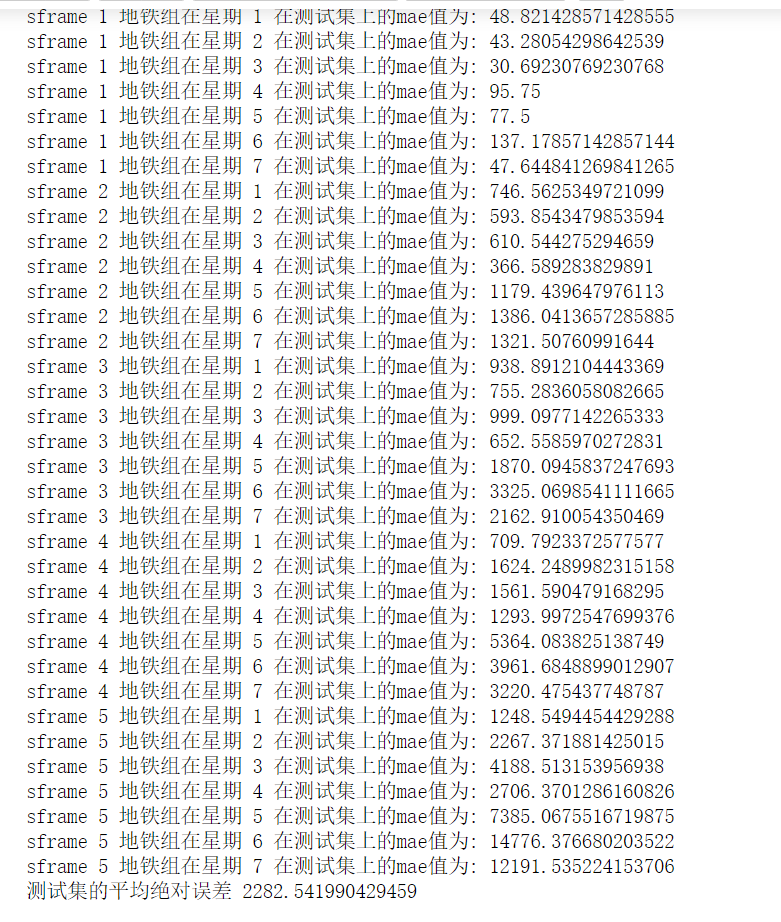
做法1和做法2的区别主要是建模对象的不同，做法1关注的是单个车站的时间联系，做法2关注的是车站之间以及时间联系，也就是X数据集的不一样。然后采取交叉验证确定当前模型最优参数,包括最佳迭代次数、CART树的最大深度、子节点生长的最小权重、损失阈值、子采样参数、整棵树的特征采样比例、L1正则和L2正则系数、学习率，最后采取最优参数字典进行测试集上的验证和预测。

(3)时间序列ARMA:

1. 模型验证：
2. RidgeCV:
   1. 做法1:



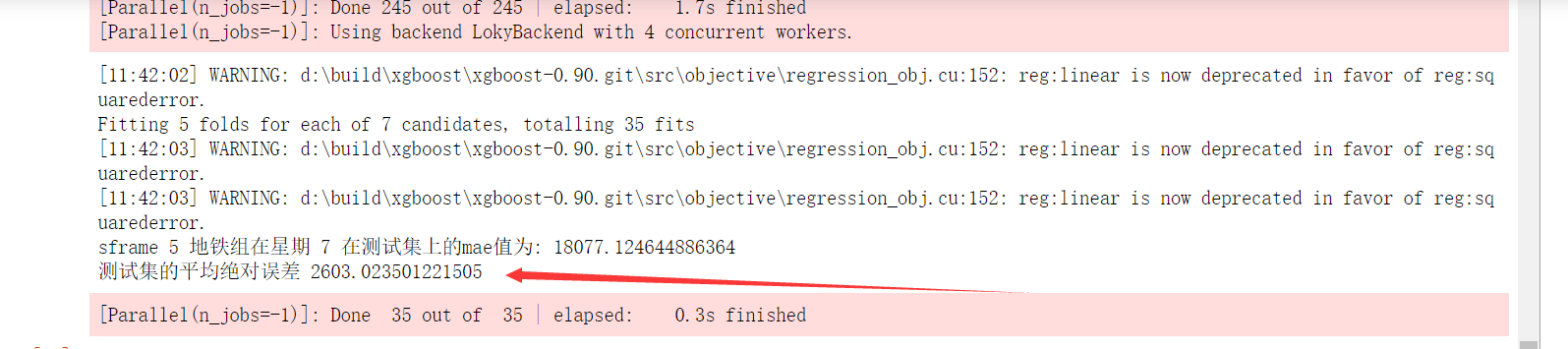
* 1. 做法2:



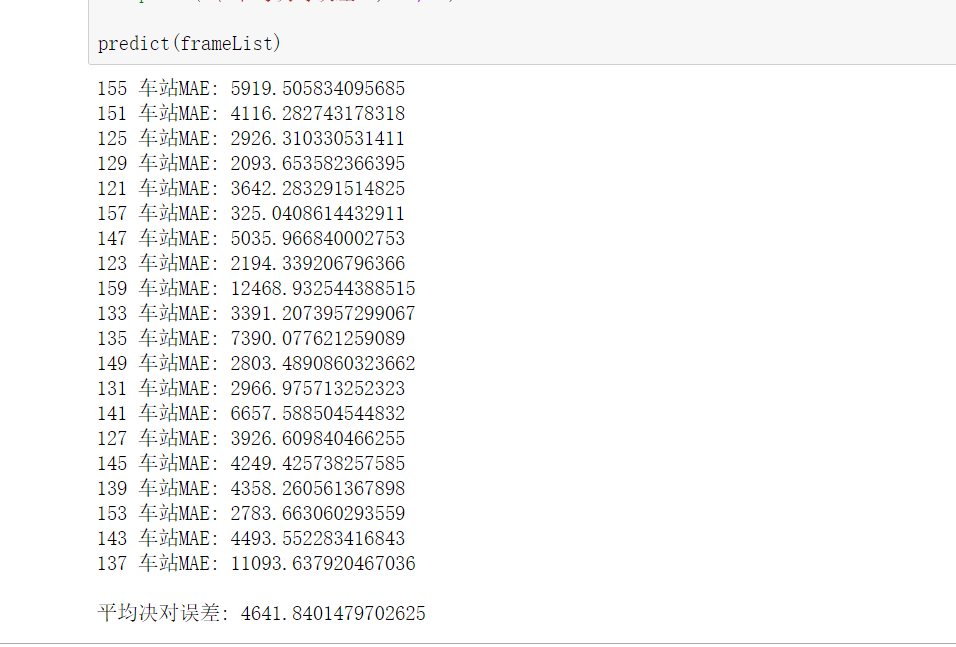
1. XGBRegressor
   1. 做法1：



* 1. 做法2：



1. ARMA：



1. 模型预测：
2. **RidgeCV：只存放做法1的，因为做法1的MAE值小,在 北部湾大学-hello\_world-交通-地铁客流量预测\说明文档\相关代码和数据\三个模型的预测数据对比\七天预测对比\** **RidgeCV做法1\_predict.xlsx里面。**
3. **XGBRegressor：只存放做法1的，因为做法1的MAE值小,在 北部湾大学-hello\_world-交通-地铁客流量预测\说明文档\相关代码和数据\三个模型的预测数据对比\七天预测对比\** **XGB做法1\_predict.xlsx 里面。**

**五、结果解释：**

**通过对比，最终以测试集上的MAE值最小的模型预测出来的数据为提交结果。即北部湾大学-hello\_world-交通-地铁客流量预测\预测结果.xlsx。**

**最终结果在“预测结果.xlsx”里面。**