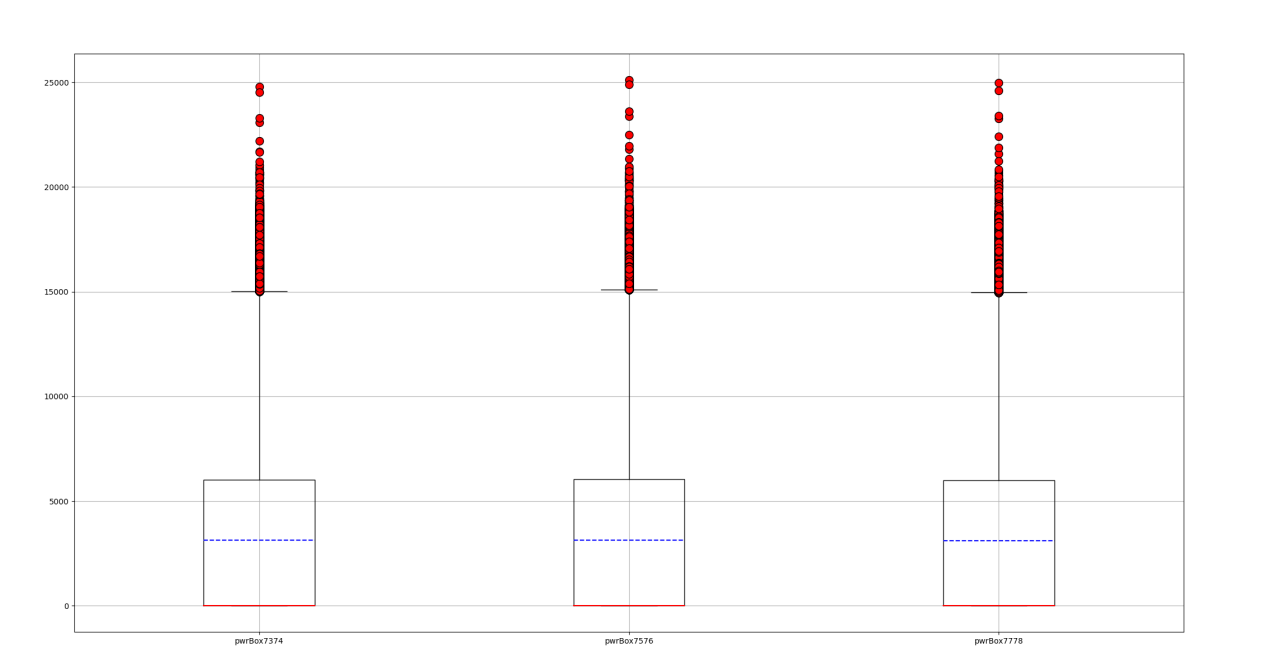
## 电源车项目数据探索

### 查看特征数据分步

通过箱线图查看每个特征数据分布情况，同时对数据中波动较大的数据进行进一步判断是否是脏数据和异常值,箱线图部分代码如下：



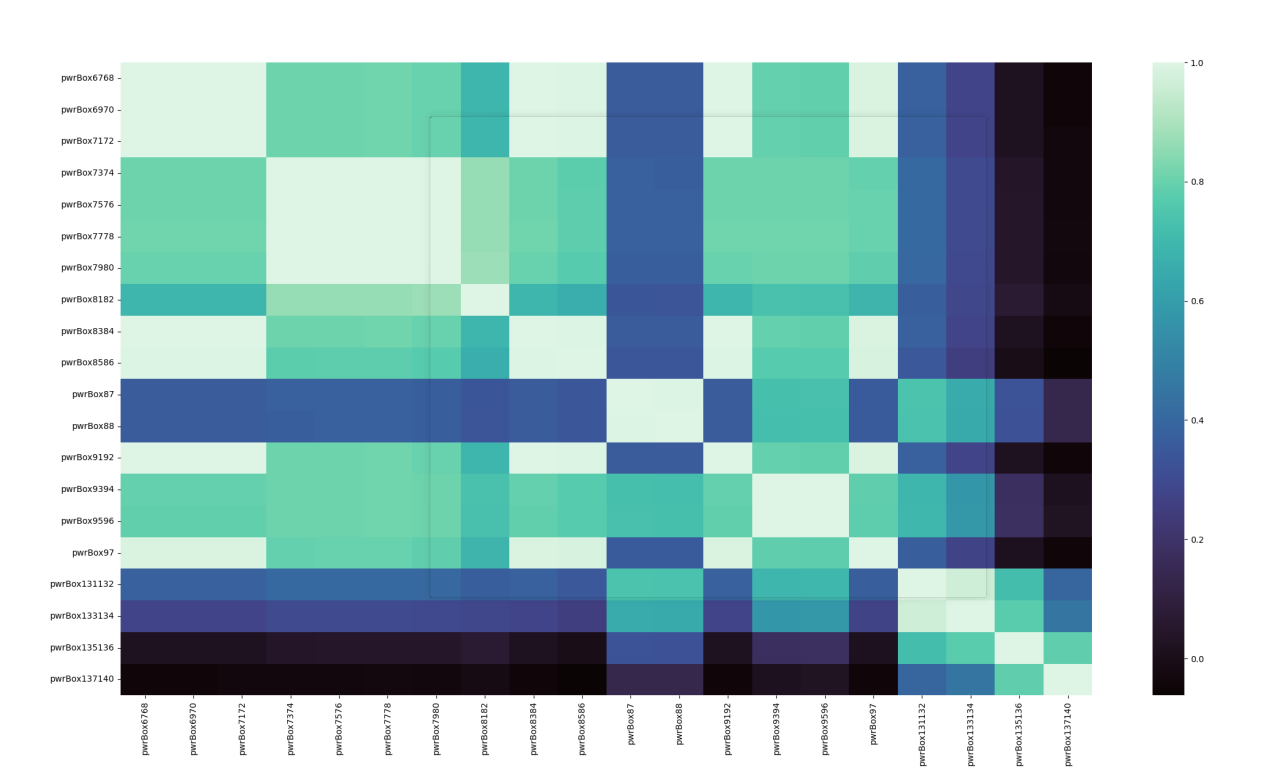


### 数据相关性分析

相关性分析是数据探索中研究变量之间最常用的方式之一。常用的相关性分析有pearson(皮尔逊),kendall（肯德尔）和spearman（斯伯曼/斯皮尔曼）三种，其具体的使用范围如下：

Spearman相关系数又称秩相关系数，是利用两变量的秩次大小作线性相关分析，对原始变量的分布不作要求，属于非参数统计方法，适用范围要广些。对于服从Pearson相关系数的数据亦可计算Spearman相关系数，但统计效能要低一些。Pearson相关系数的计算公式可以完全套用Spearman相关系数计算公式，但公式中的x和y用相应的秩次代替即可。  
          Kendall'stau-b等级相关系数：用于反映分类变量相关性的指标，适用于两个分类变量均为有序分类的情况。对相关的有序变量进行非参数相关检验；取值范围在-1-1之间，此检验适合于正方形表格；  
        计算积距pearson相关系数，连续性变量才可采用;计算Spearman秩相关系数，适合于定序变量或不满足正态分布假设的等间隔数据;计算Kendall秩相关系数，适合于定序变量或不满足正态分布假设的等间隔数据。  
        计算相关系数：当资料不服从双变量正态分布或总体分布未知，或原始数据用等级表示时，宜用 spearman或kendall相关。  
          Pearson相关复选项积差相关计算连续变量或是等间距测度的变量间的相关分析   
Kendall 复选项、等级相关、计算分类变量间的秩相关，适用于合并等级资料。

考虑三种方法的使用范围及运算效率，本项目使用Pearson系数来分析本项目不同特征之间的数据分析。下图是本数据源相关性矩阵图表，从图中可以清晰的观察出不用特征之间的关系。



图中颜色越浅带边两个特征之间的正相关性越强，颜色越深代表两个特征之间的正向相关性越强。

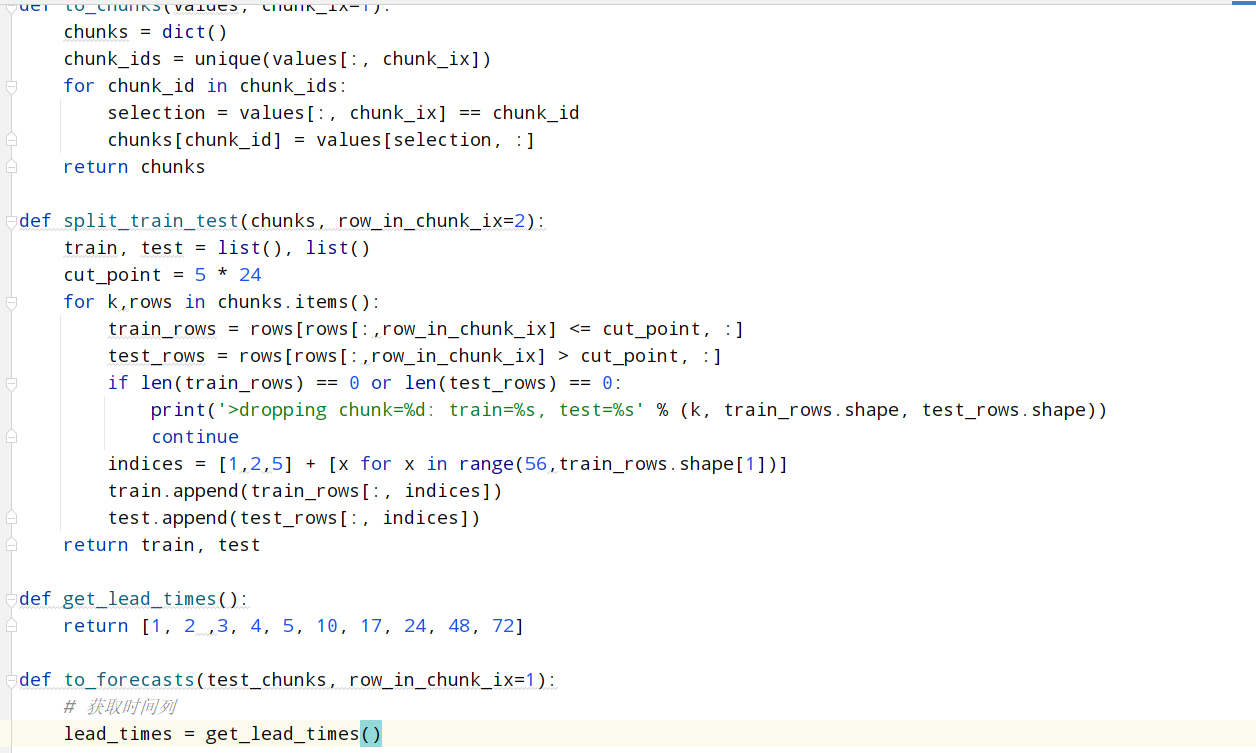
### 数据筛选及空值处理

由于源数据集中含有许多无用数据，本项目通过pandas先对数据中干扰数据进行筛选，其次鉴于数据中可能存在空值已经野值会对后期建模的准确率造成影响，本项目根据数据不同时期的具体状况分别采用上值、下值、中值、均值已经众数等多种方式对数据进行填充。保证处理后的数据具备较强的可用性。最后根据柴油发电机运行原理已经数据实际情况过滤到建模过程中不需要的特征，经过多次探究最终保留一下特征作为模型训练参数：'datetime', 'pwrBox6768', 'pwrBox6970', 'pwrBox7172', 'pwrBox7374','pwrBox7576', ‘pwrBox7778', 'pwrBox7980', 'pwrBox8182', 'pwrBox8384','pwrBox8586', 'pwrBox87', 'pwrBox88', 'pwrBox9192', 'pwrBox9394','pwrBox9596', 'pwrBox97', 'pwrBox117', 'pwrBox118', 'pwrBox119','pwrBox120', 'pwrBox121', 'pwrBox122', 'pwrBox123', 'pwrBox124', 'pwrBox125', 'pwrBox126', 'pwrBox131132', 'pwrBox133134', 'pwrBox135136', 'pwrBox137140', 'pwrBox175', 'pwrBox180'

### 数据集切分

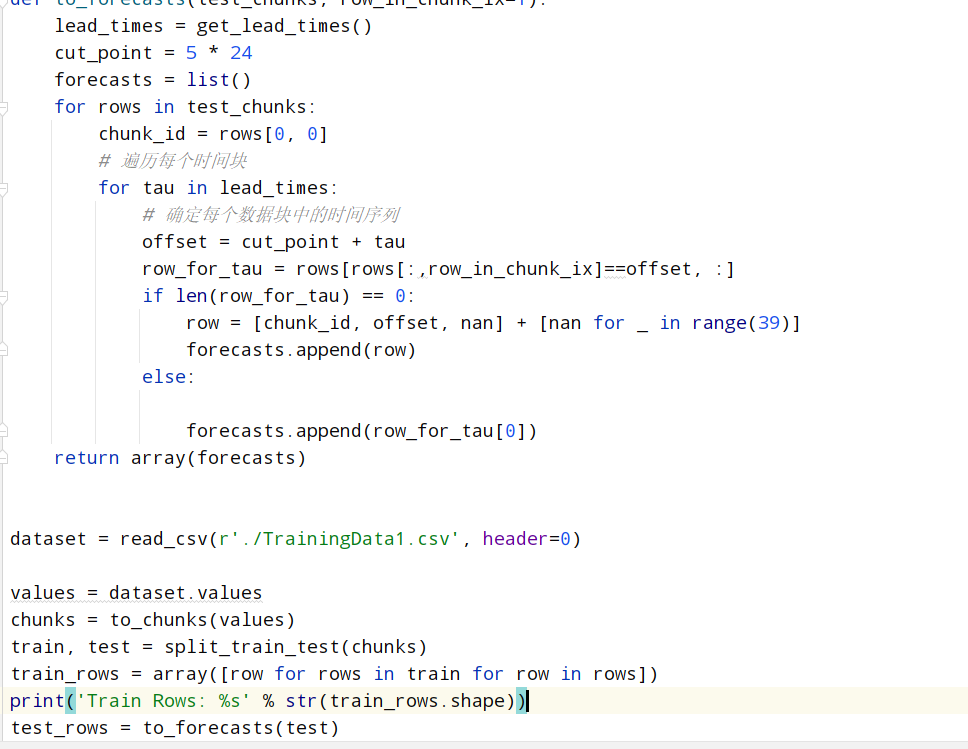
为了保证模型训练结果的可靠性，需要将数据集切分为训练集、测试集、验证集三个部分，通常三个部分数据的比为6:2:2。训练集主要用于模型的训练，测试机用于模型训练中模型的矫正调优，验证集主要用于验证码模型最终预测效果。

数据切分的主要代码如下：



### 数据构造

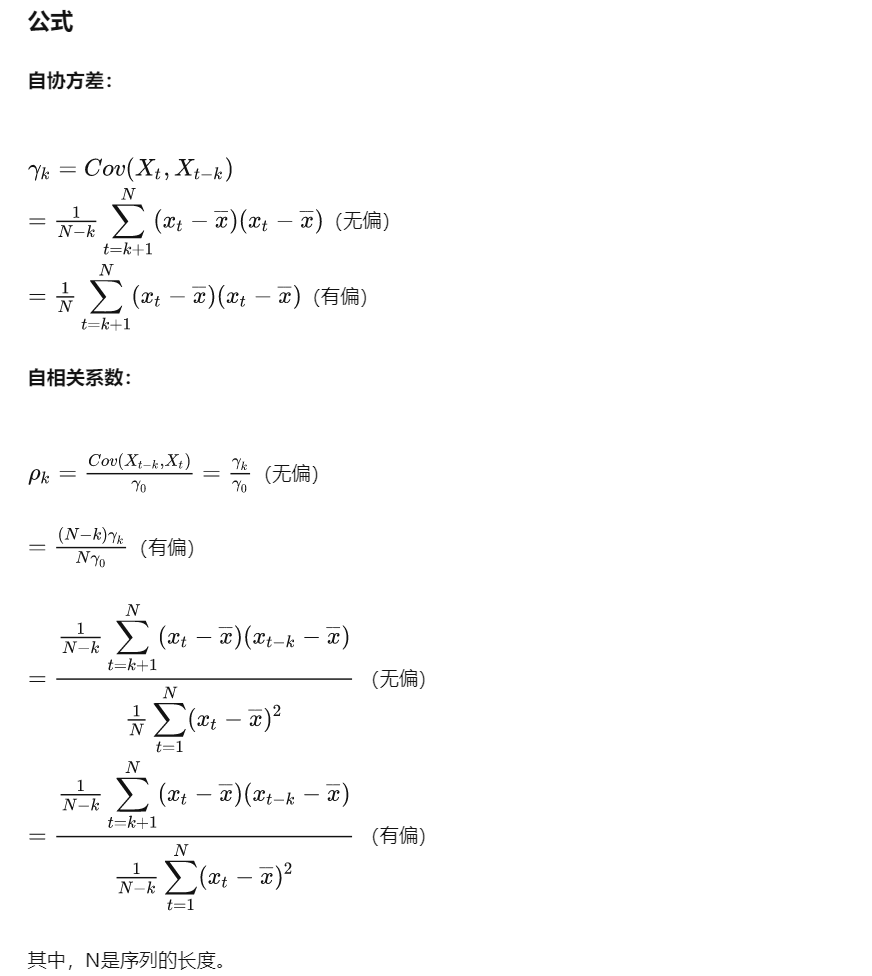
时序序列模型无法直接直接使用二维数组建模，需要对数据进行二次加工处理重新建数据构造成三维数组，由于当前数据还需要进行优化构造，当前文档中暂时将目前最优的模型效果构造成[16, 8,6]一个批次进行建模。数据构造需要大量探索才能得到最优模型，其中部分代码如下:



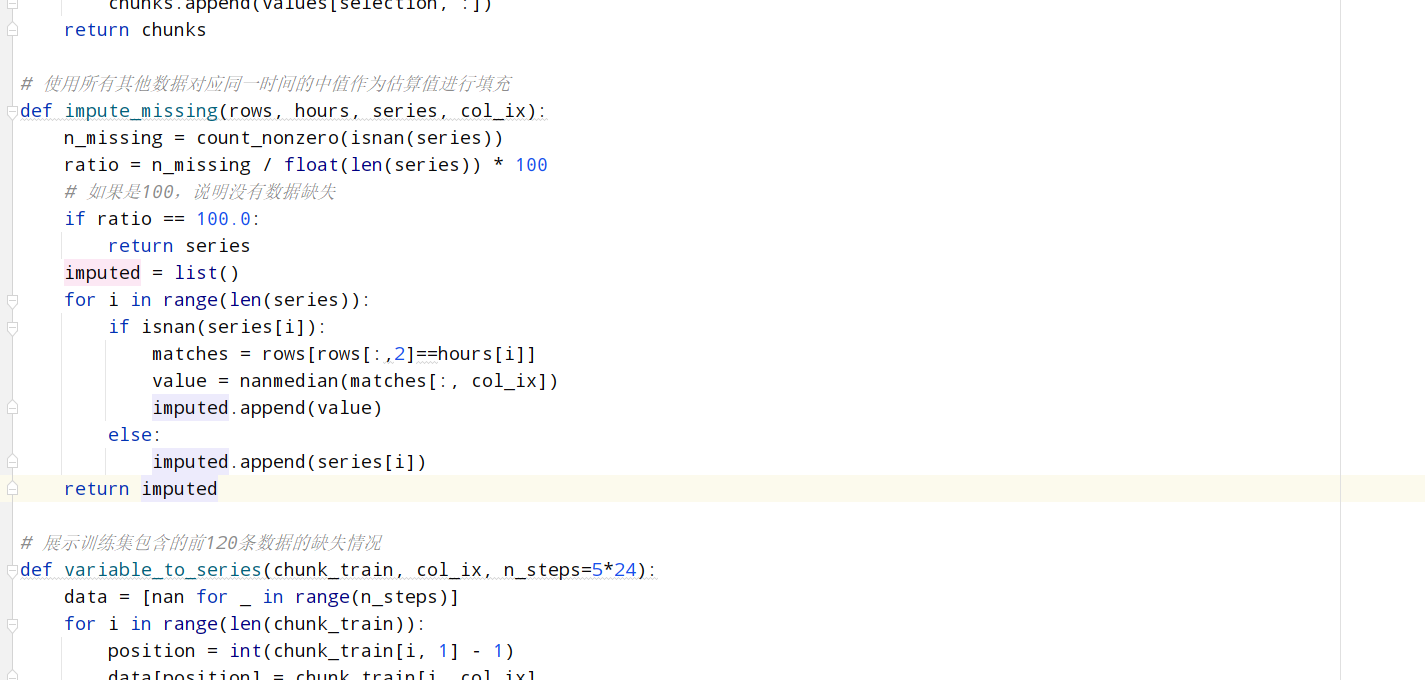
### 计算发电机前轴承自相关系数和协方差

时序序列中为了保证特征的平稳性需要对数据做平稳性校验和协方差计算，计算公式如下：

ACF（Autocorrelation Function 自相关函数）：自相关系数构成的序列。



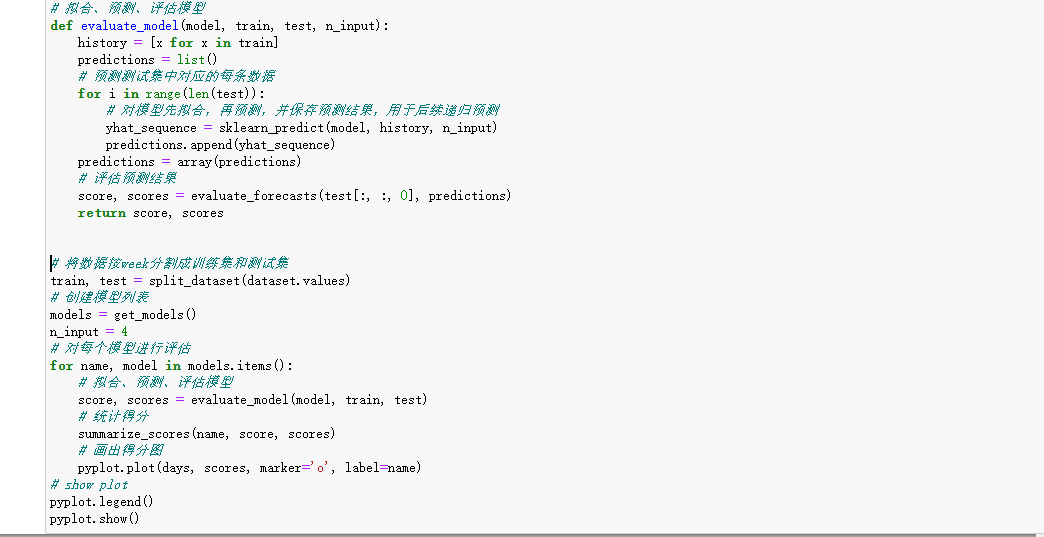
部分代码如下：



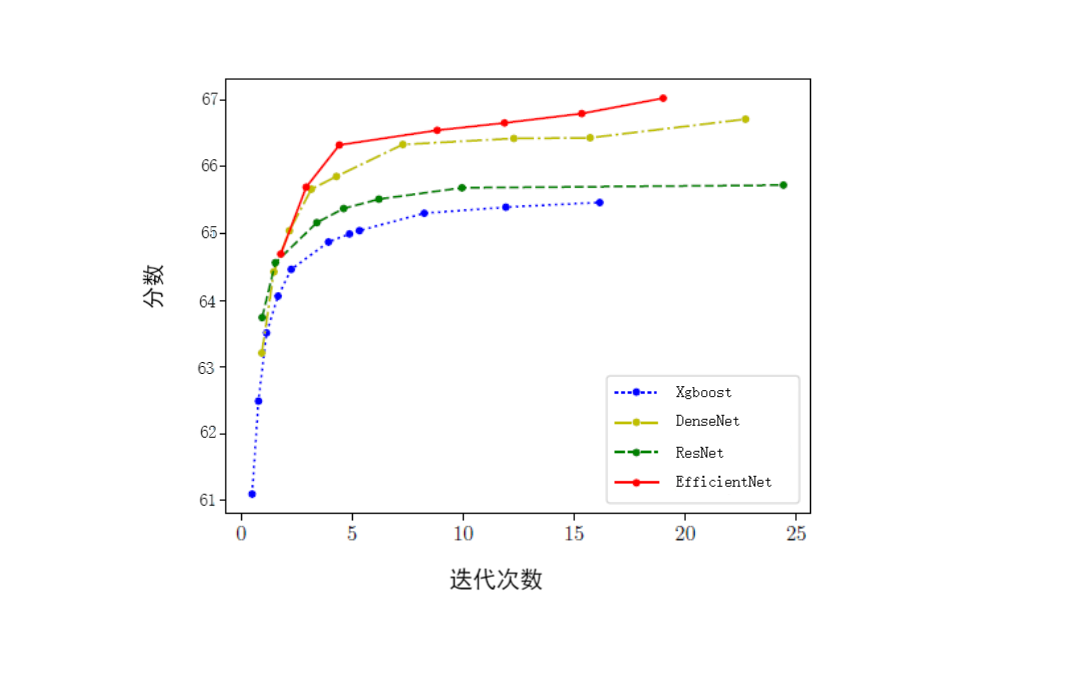
保证数据的特征稳定之后，就可以先使用传统机器学习算法对模型进行训练。

### 模型训练与评估

将数据处理为三维数据后，将数据分为训练集和测试集，其中训练集用于训练模型，测试集用于评估模型。分别使用Xgboost、DenseNet、ResNet、EfficientNet等算法训练模型，训练好模型后使用测试集验证模型效果，预测未来10分钟的故障情况，其中部分代码如下：



最后使用pyplot绘制出不同模型的分数曲线，最终的结果如下：



以上结果均为对算法参数进行调优，同时探索阶段的数据处理流程较为简单，后续需要通过大量探索得到最优的算法参数与结果，目前还有很大的提升空间。