|  |
| --- |
| **机器学习** |
| VSB Power Line Fault Detection |
| 主 研 人：岳天昕  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2018/10/31 | A | 初稿 |  |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/4/15 | VSB Power Line Fault Detection | 此次挑战是使用在VŠB的ENET中心设计的新仪表检测从这些电力线获取的信号中的局部放电模式。 使用此数据的有效分类器可以连续监控电力线路是否有故障。 | 分类 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 5](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 5](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 6](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 6](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 10](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一数据预处理及特征工程部分方案 10](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计、建立部分方案 13](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 18](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 19](#_Toc4787_WPSOffice_Level3)

[3.2 方案二 20](#_Toc29922_WPSOffice_Level2)

[20](#_Toc4787_WPSOffice_Level2)

[4. 算法比较 22](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 22](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 23](#_Toc16174_WPSOffice_Level1)

[5.1 总结 23](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 23](#_Toc16619_WPSOffice_Level2)

1. 背景描述

中压架空电力线路运行数百英里，为城市供电。 这些较大的距离使得手动检查线路的损坏（例如树枝撞击线路或绝缘体中的缺陷）变得昂贵。 这些损坏现象称为局部放电的现象 - 放电，其不完全桥接绝缘系统之间的电极。 局部放电会慢慢损坏电源线，因此如果没有修复，它们最终会导致停电或起火。

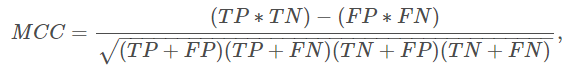
* 1. **竞赛赛题描述**

您面临的挑战是使用在VŠB的ENET中心设计的新仪表检测从这些电力线获取的信号中的局部放电现象。 使用此数据的有效分类器可以连续监控电力线路是否有故障。  
ENET中心研究和开发可再生能源，旨在减少或消除有害的环境影响。 他们的工作重点是围绕能源原材料的运输和加工开发技术解决方案。  
通过开发检测局部放电的解决方案，您将有助于降低维护成本，并防止停电。

**1.2 评估指标描述**

竞赛中评估模型优劣的指标。

选手提交结果与实际检测到的结果进行对比，以Matthews相关系数（MCC）为评价指标，计算公式如下：



其中TP是真正例的数量，TN是真反例的数量，FP是假正例的数量，FN是假反例的数量

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

数据来自VŠB的ENET中心设计的新仪表检测到的信号数据。这里是数据的超链接：<https://www.kaggle.com/c/vsb-power-line-fault-detection/data>

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

电力传输线中的故障会导致称为局部放电的破坏性现象。 如果单独放置，局部放电可能会损坏设备，使其完全停止工作。 您的挑战是检测局部放电，以便在发生任何持久性伤害之前进行修复。  
每个信号包含800,000次电源线电压测量，测量时间超过20毫秒。 由于底层电网工作在50 Hz，这意味着每个信号覆盖一个完整的电网周期。 电网本身采用三相电源方案，同时测量所有三相。

**2.2.2 数据字段介绍：**

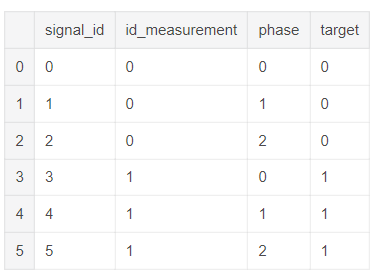
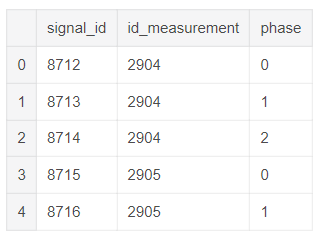
做成表格形式（如果有多个数据表，请做多个数据字段介绍表格），格式如下：

**表2-1 训练/测试数据表字段介绍**

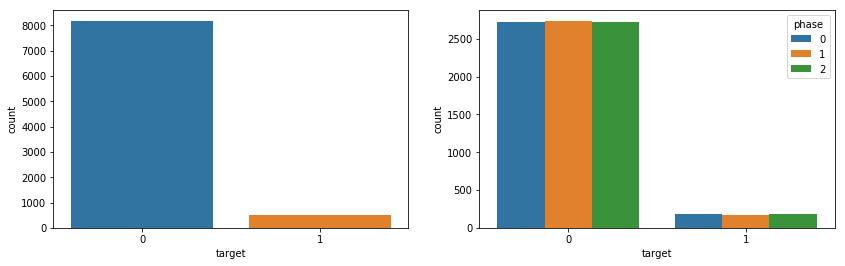
|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| id\_measurement | 同时记录的三个信号的ID代码。 | 离散 | 0% |
| signal\_id | 信号数据的外键。 每个信号ID在列车和测试中都是唯一的，因此训练集中的第一个ID为“0”，但测试集中的第一个ID为“8712”。 | 离散 | 0% |
| phase | 信号trio中的相位ID码。 这些阶段可能会也可能不会受到线路故障的影响。 | 离散 | 0% |
| target | 如果电源线未损坏，则为0;如果有故障，则为1。（测试集中无） | 离散 | 0% |

**2.2.3 数据描述性统计**

训练集/测试集预览：（无缺失值）

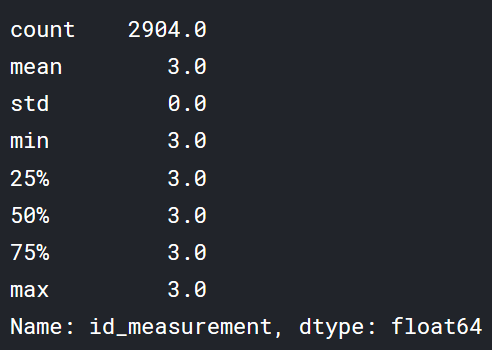
 

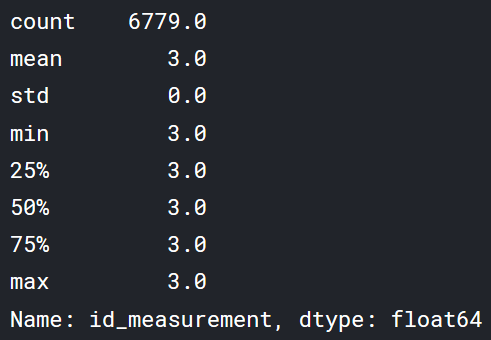
查看目标值：

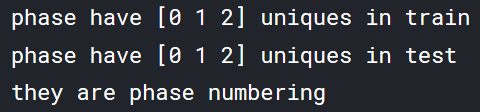


看出电线有故障的数据太少了，目标值在每个相位分布几乎均匀

查看元数据：

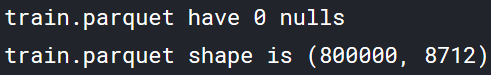
训练集 

测试集

相位

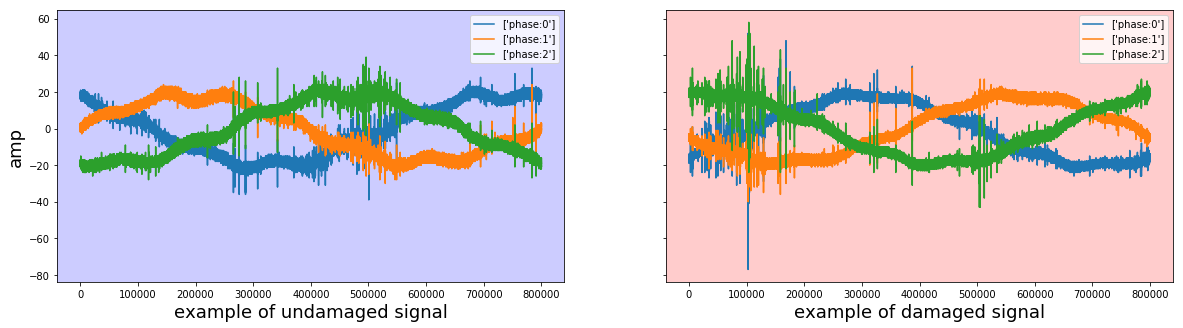
查看信号数据：

每列包含一个信号; 使用pyarrow.parquet版本0.11导出的800,000个int8测量值。  
请注意，这与我们通常每次观察一行的数据方向不同;  
开关可以有效地加载信号的子集。

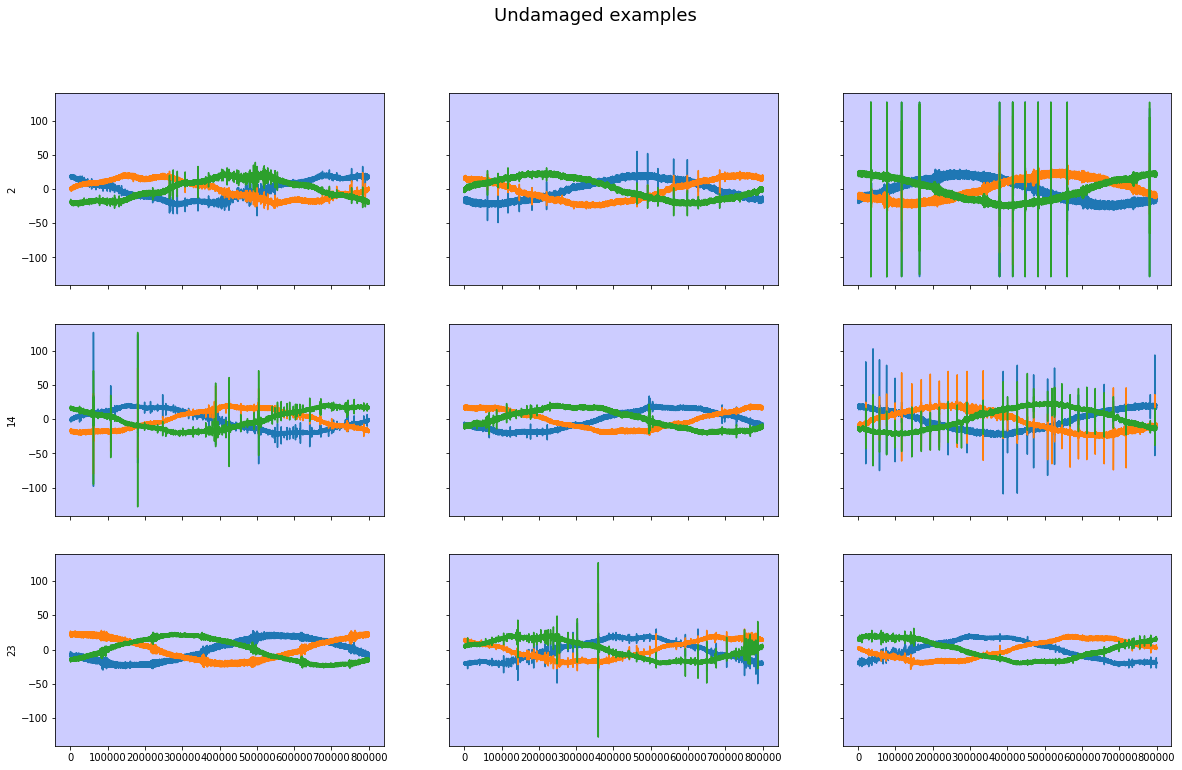
训练信号数据

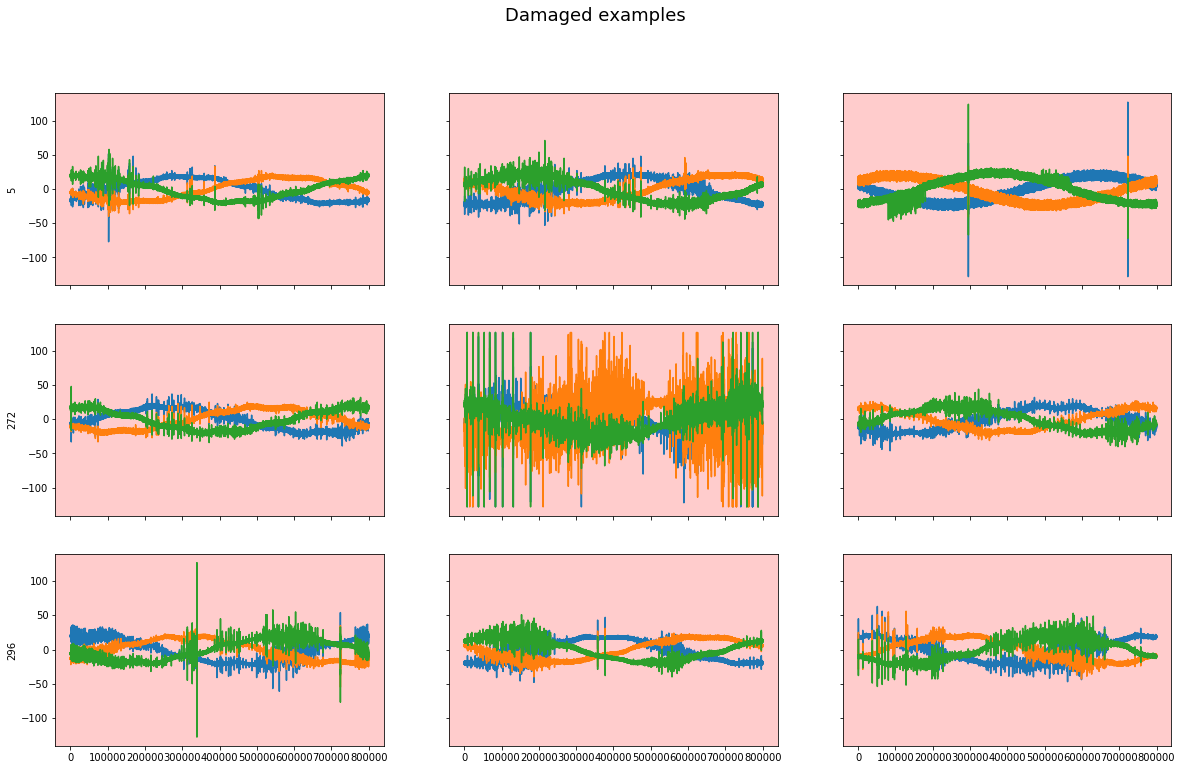


查看信号波：



我们可以看到三个相位和一些噪音点，没有看到很大的区别  
这种差异是大噪音点吗？让我们看看更多的数据





3. 优秀算法思路

这个部分麻烦寻找竞赛结果排名至少top10的思路分享（天池大赛会在技术圈置顶，kaggle的话需要自己在交流区搜索一下）优秀算法思路可以包含多个，可以选择2-3个算法方案进行描述。

**3.1 方案一**

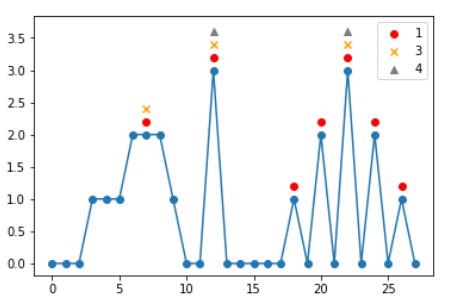
**3.1.1** 方案一数据预处理及特征工程部分方案

预处理每个信号迹线以识别峰值并计算特征。  
具体步骤：

1.使用flatiron\_function展平迹线  
2.使用local\_maxima\_1d\_window函数识别每条迹线中的局部最大峰值  
3.过滤步骤2中识别的峰值，将信号与噪声分开  
4.使用calculate\_peak\_features计算步骤3中识别的每个峰的特征

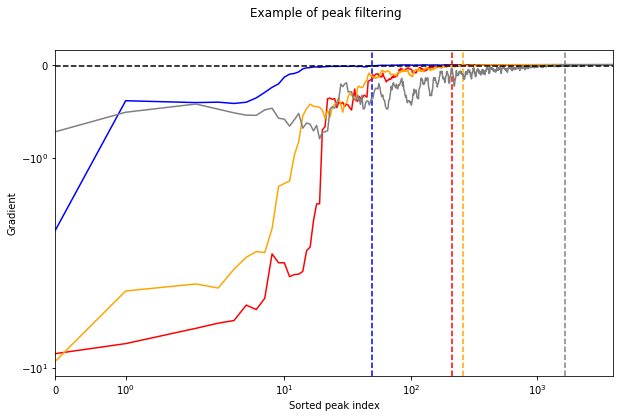
步骤2.为了识别局部最大值，使用函数local\_maxima\_1d\_window。 此函数采用窗口长度参数，该参数是用于比较的每一侧的点数。

图示如下：

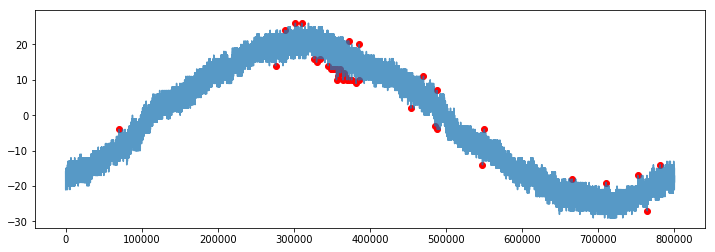


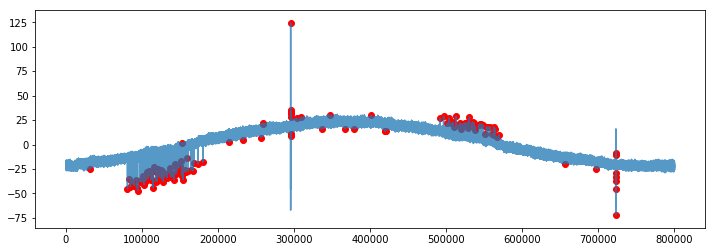
步骤3.一旦识别出迹线中的所有峰值，就需要消除由信号中的噪声引起的峰值。这是在get\_peaks函数中执行的。当峰值按高度排序时，执行拐点检测以识别由于达到噪声基底而停止变化时的点。

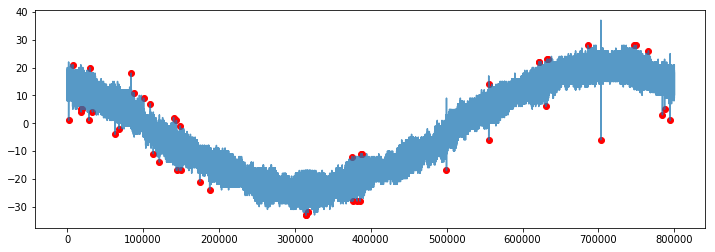
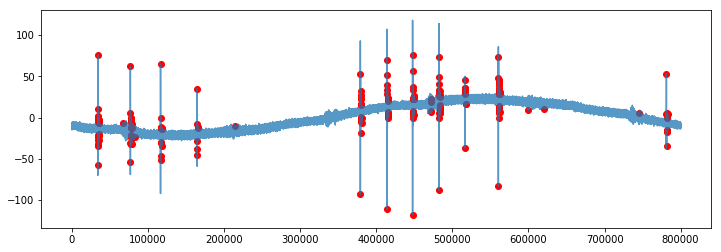
图示：

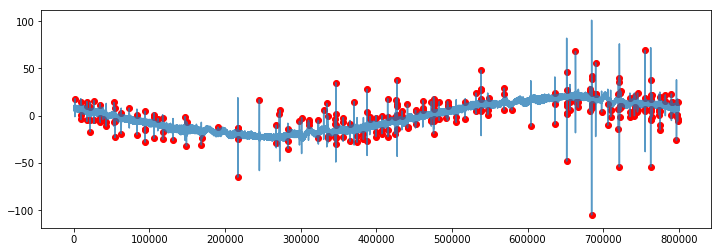
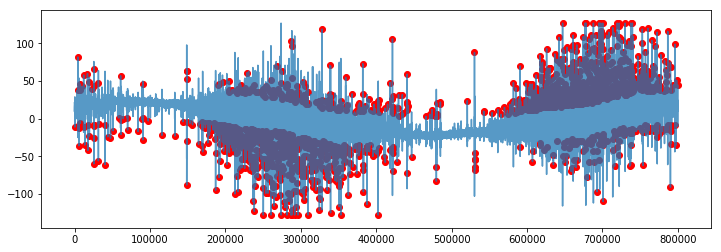


使用上述步骤检测到的峰的一些示例如下所示：



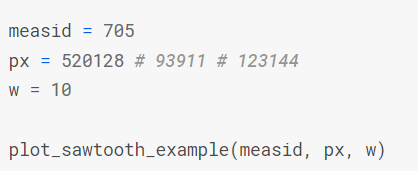
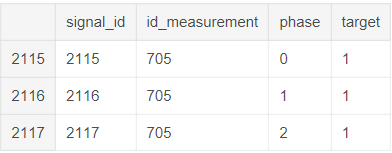


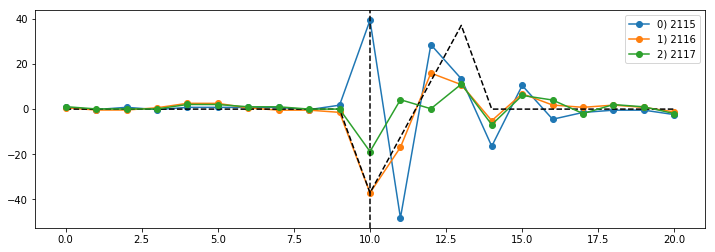




步骤4.一旦消除了噪声峰值，就计算每个剩余峰值的特征。

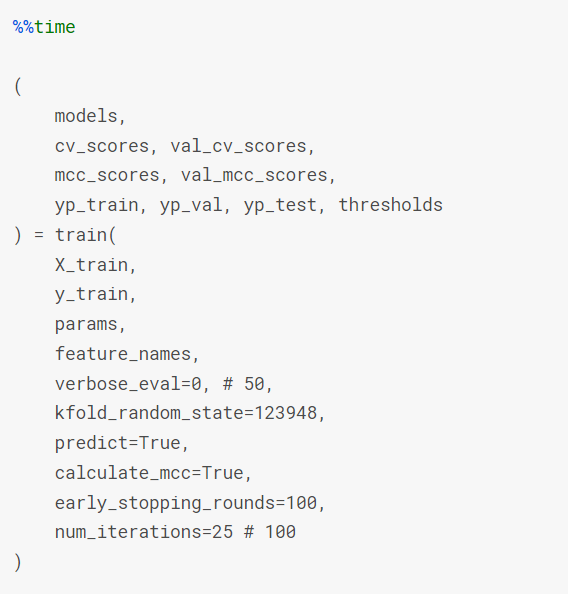
这在calculate\_peak\_features中执行。其中最有趣的功能是sawtooth\_rmse功能。这是值窗口之间的RMSE，包括峰值和峰值两侧的25个数据点以及锯齿状模板。下面是一个例子：

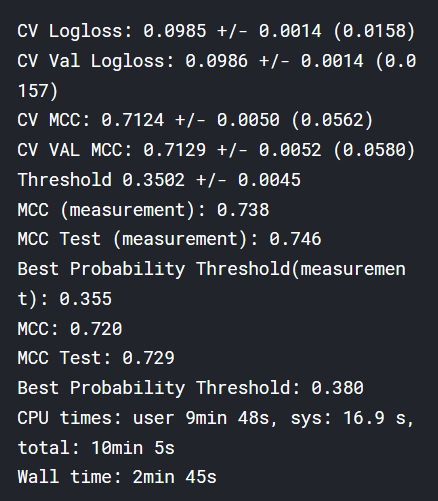
 



**3.1.2** 方案一模型设计、建立部分方案

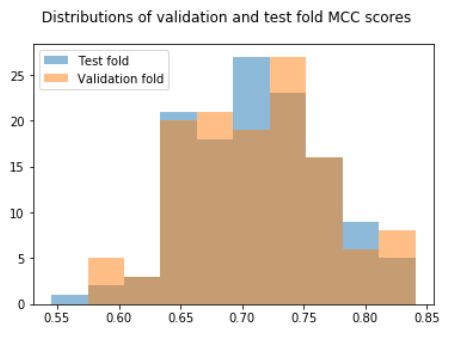
训练模型：





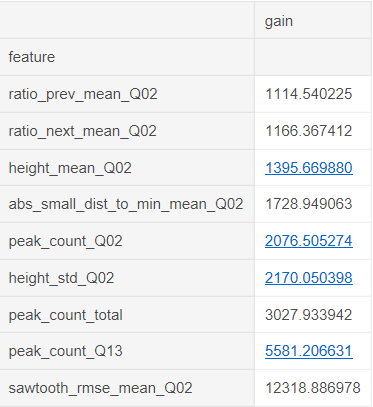
测试模型：

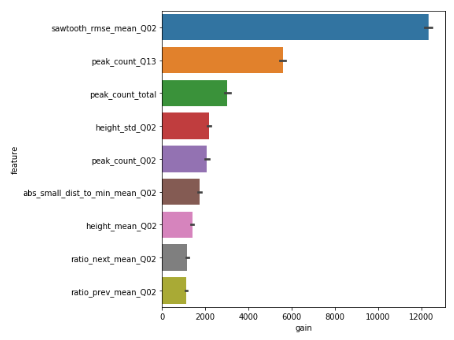
评估：



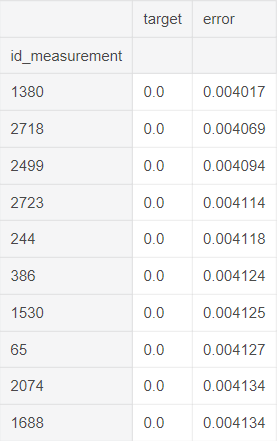
分析：

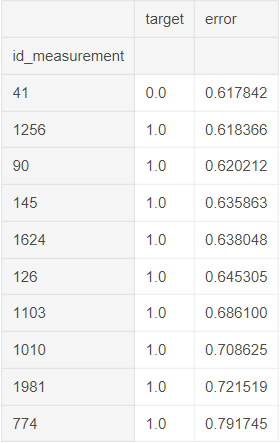
特征重要性





预测错误率：

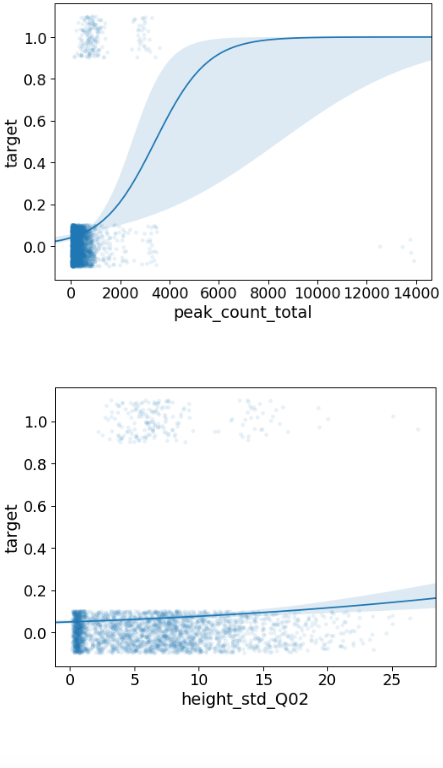
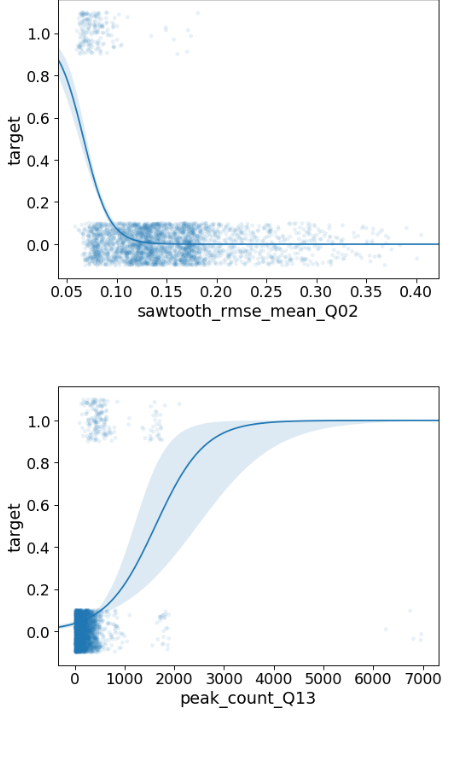
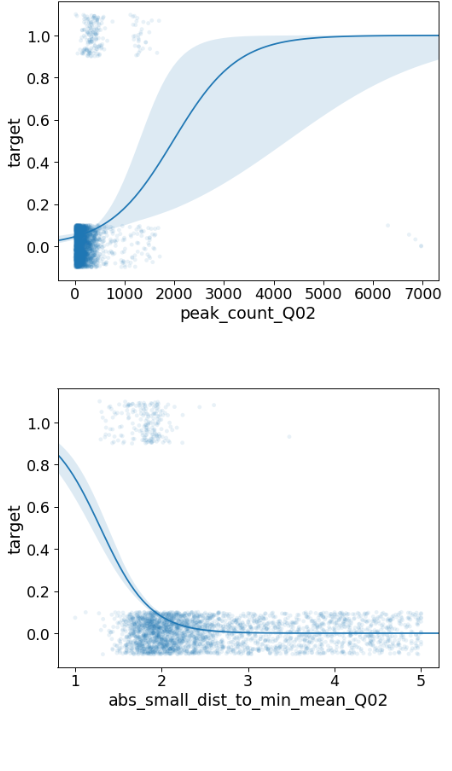


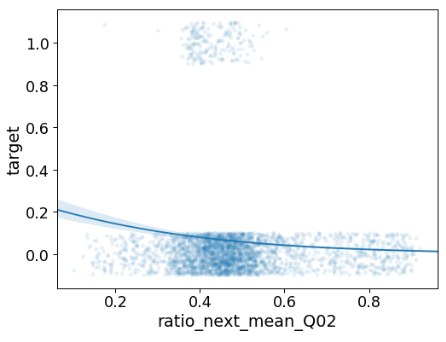
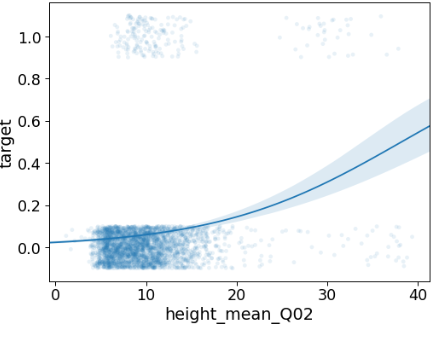
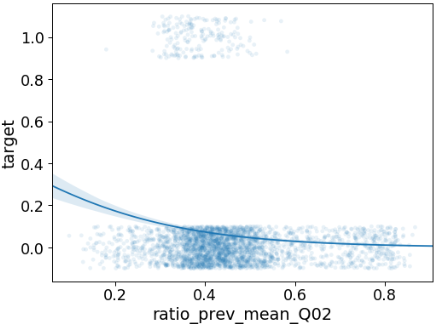


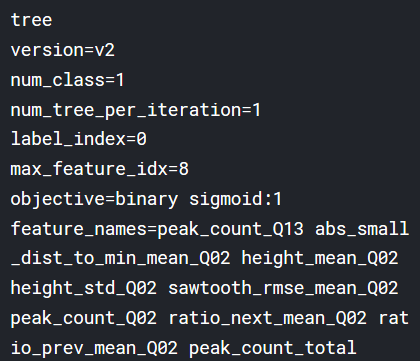
测试：

得到测试结果后分析，检测结果数据完整性准确性

再次查看特征相关性

   
最终模型：

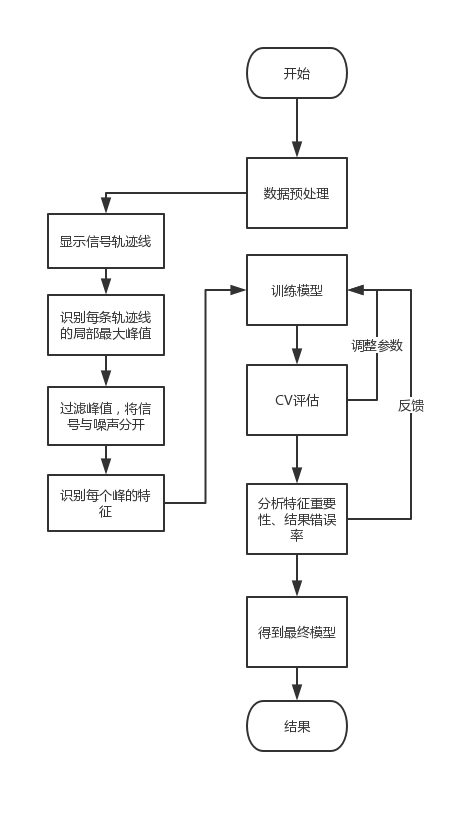


**3.1.3** 方案一结果、排名等

TOP1

Score=0.71501

**3.1.4** 方案一算法流程图



**图3-1**

**3.2 方案二**

**3.2.1** 方案二数据预处理及特征工程部分方案

**预处理：**  
3个信号相位仅用作一个样点。每个信号被对齐（使用四线变换的50Hz相位）以从信号穿过轴从负到正（π/ 2）开始，并且去除信号的最后四分之一。  
在每个模型中，我使用了这个对齐/裁剪信号的一些去噪版本（变化的阈值）：

信号波：  
在MaxHalford的存储库之后：extract\_solo\_features.py  
Fourrier / IQR  
使用傅立叶变换消除低频并使用四分位数范围消除点  
我最终获得了4个信号版本（Wavelet / Fourrier，具有2个阈值电平）  
对于RNN和CNN，我将600000大小的信号欠采样到300000，取得每对点的最大

绝对值位置：  
und\_signal = np.where（np.abs（signal [:: 2]）> np.abs（signal [1 :: 2]），signal [:: 2]，signal [1 :: 2]）

**3.2.2** 方案二模型设计、建立部分方案

我的最终解决方案是三个主要分支的集合：  
渐变助推树（Lightgbm）  
递归神经网络（GRU和LSTM）  
卷积神经网络（自定义CNN，Resnet50，DenseNet101）

**GBT：**  
特征是min，max，mean，std，skew和kurtosis：  
峰数，高度，宽度，突出：extract\_solo\_features.py  
熵与分形：vsb-competition-attention-bilstm-with-features  
连接正/负峰值到下一个负/正峰值的线的斜率  
正峰值与下一个最大距离（在Vantuch的论文中定义）点的最小值之比。  
负峰值与下一个最大距离点的最大值之比。  
R²和一些多项式回归的权重适合于正/负峰值

**RNN：**  
预处理信号重新整形为40列

**CNN：**  
预处理信号重新整形为200列

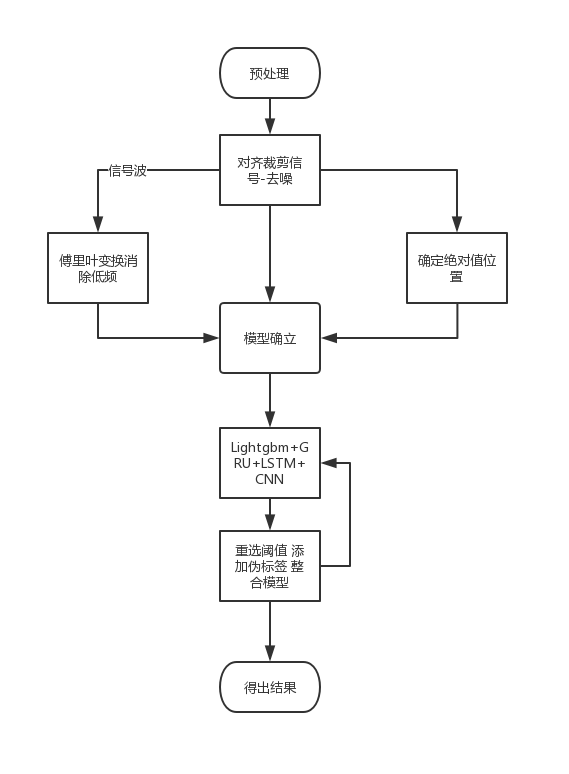
**LB预测**  
在比赛的某个时刻，我建立了一个套索来预测提交的LB分数。我知道它会产生过度拟合的问题所以我试图避免它：  
阈值选择：我将所有阈值固定为0.5  
伪标签：我在所有模型中使用它，我减少了训练有素的套索中重量轻的测试实例的样本重量  
整合：我应该最终得到24个模型（6个模型x 4个预处理信号）但我只有16个（缺少时间......！）  
我无法创建稳定的堆叠，因此我使用了非常精细的策略来集合模型：采用预测的算术平均值...

**3.2.3** 方案二结果、排名等

TOP6

Score = 0.69147

**3.2.4** 方案二算法流程图



4. 算法比较

描述几种算法的基本情况及效果对比，以表格形式：

例如： 特征工程和建模部分可以描述建模用到的基础算法名称（PCA、LSTM、时间衍生等）基本库指用到的基本库（sklearn等）

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **CV评估** | flatiron\_function local\_maxima\_1d\_window  calculate\_peak\_features | **Tree** | **Keras** |
| **算法2** | **没解释** | extract\_solo\_features.py | **Lightgbm+GRU+LSTM+CNN** | **Keras** |

5. 总结与展望

**5.1 总结**

这道题包含信号处理，每个信号包含800,000次电源线电压测量。由于底层电网工作在50 Hz，这意味着每个信号覆盖一个完整的电网周期。电网本身采用三相电源方案，同时测量所有三相。优秀方案中，大家都处理了信号峰值，并从峰值中提取了特征，这样不容易过拟合。

因为此题和物理电压紧密联系，所以我认为基于物理系统理解构建特征应该会使结果更加准确。如果足够了解故障机制，就可能构建一些模型中难以发现的决定性的特征。

**5.2 建模思路**

之前没有接触过信号处理，通过此次kaggle的赛题，我了解到了一些基本的信号的特征处理方法，比如去除噪音点，用平均值代替缺失值，通过信号峰值提取特征。再将这些信号特征放入模型中训练，通过CV评估等评估方法调整模型，最后得出结果。