2019数字中国创新大赛

“海上风场SCADA数据缺失智能修复”答辩

团队：NXmeah

初赛B榜：第六名

小队成员

匡祯辉，秦金绪， 丁青，刘世欢， 陈博文

01

团队简介

成员介绍

02

赛题分析

赛题任务

赛题方案

数据分析

验证集构建

03

数据修复

统计量填充

模型填充

机组相互填充

04

总结展望及致谢

总结与展望

致谢

团队简介

匡祯辉

福州大学机械工程学院硕士在读、数据挖掘爱好者、工业数据应用

研究方向：工业数据挖掘

秦金绪

浙江工商大学统计学理学硕士、数据分析与数据挖掘爱好者、多次获得国内外各类数据挖掘比赛中获取过名次

近年获奖：ATEC蚂蚁开发者大赛--支付风险识别 一等奖

银联银杏大数据竞赛--信贷预期预测 二等奖

丁青

浙江工商大学统计学理学硕士、从事风控策略相关工作

近年获奖：融360天机智能算法挑战赛--拒绝推断 三等奖

刘世欢

浙江工商大学硕士

近年获奖：天池--美年AI大赛Rank9

融360天机智能算法挑战赛--拒绝推断 二等奖

陈博文

福州大学机械工程学院硕士在读、数据挖掘爱好者、自然语言处理

研究方向：自然语言的仿生设计运用

赛题任务&赛题方案

赛题分析

- 赛题任务：要求参赛团队对海上风场机组缺失数据做出修复。

- 赛题方案：根据数据分析，不同列选用下述不同修复方法，线上结果表明效果还不错！

a) 插值算法

b) 统计值填充

c) 模型填充

赛题分析

赛题方案

1)当$x_{i,j}$为浮点数值型变量时:

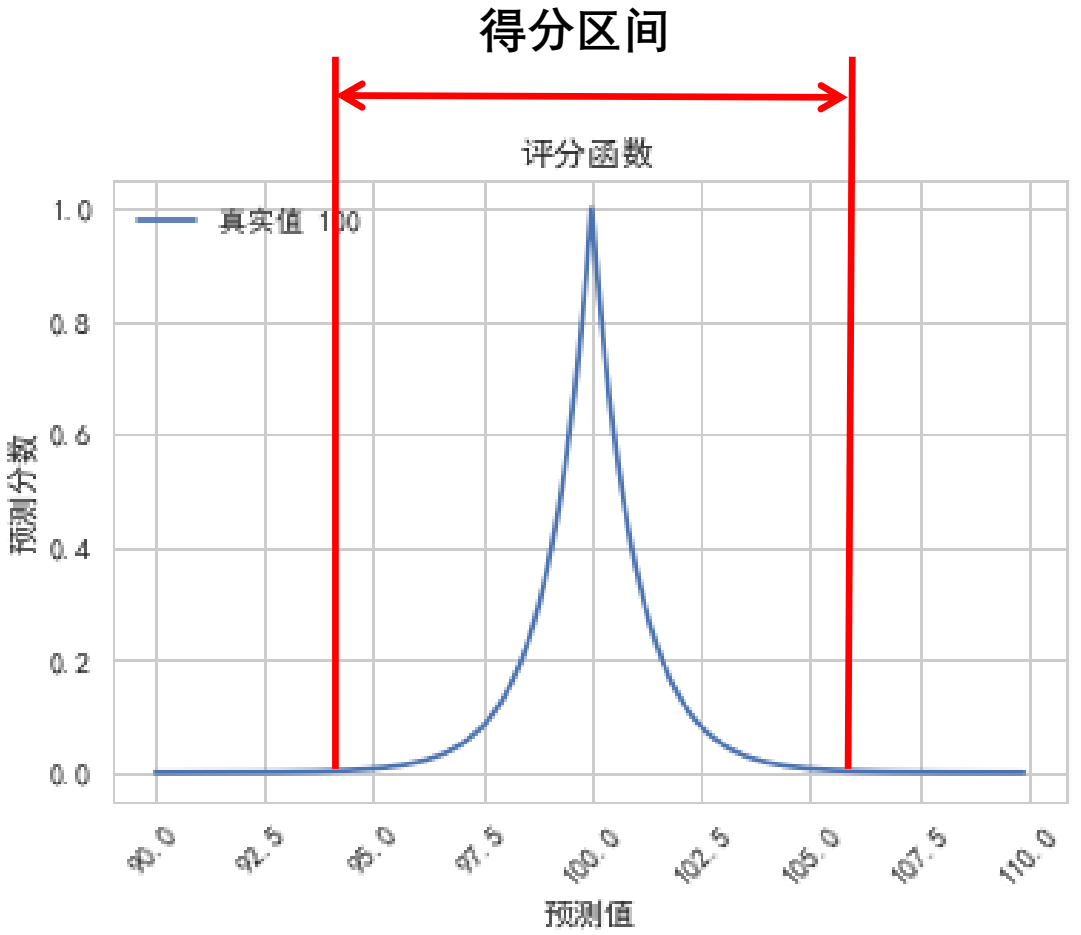
$$f_{i,j}(x_{i,j},\hat{x}_{i,j}) = e^{-\frac{100|x_{i,j}-\hat{x}_{i,j}|}{max(x_{i,j},10^{-15})}}$$

2)当$x_{i,j}$为枚举或布尔型变量时:

$$f_{i,j}(x_{i,j},\hat{x}_{i,j}) = \begin{cases} 1, \hat{x}_{i,j} = x_{i,j} \\ 0, \hat{x}_{i,j} \neq x_{i,j} \end{cases}$$

1：浮点型变量评分函数，关注相对误差；同时评分函数非线性，一旦预测值超出真值某个范围，得分等同于0。

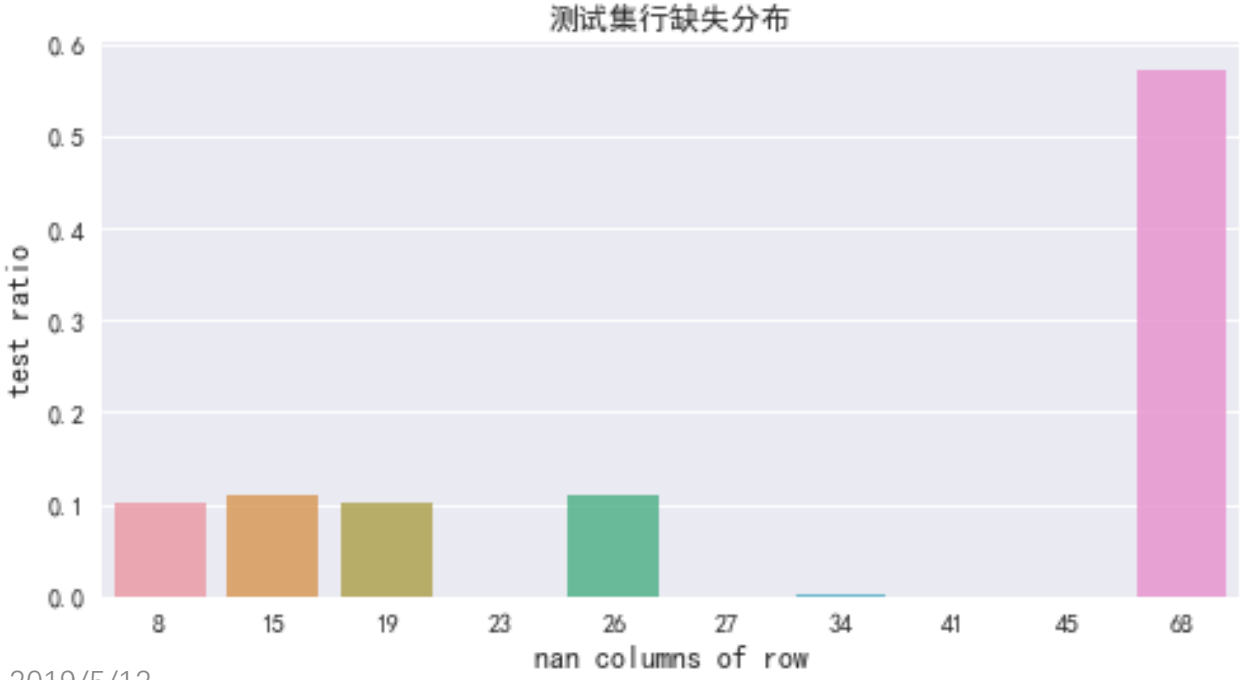
2：枚举和布尔型变量采用分类问题精确率作为评分函数，表面上看预测错误就没有任何收益，实际上由于类别极度不均衡，全部填大类就有不错的收益。

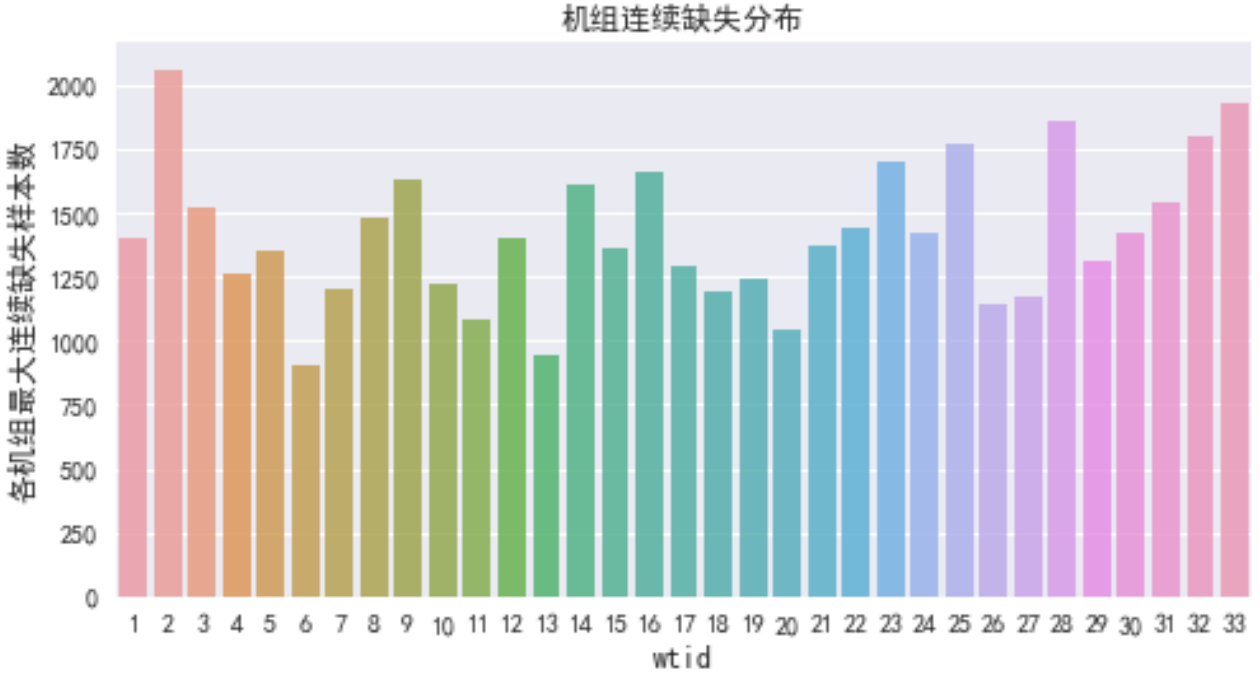


赛题分析

赛题难点

- 高相关性列一起缺失，并且采用匿名变量
- 稳定的线下验证集的建立
- 数据全部缺失的行在测试集中占比大
- 数据时序上最大连续缺失约3.5小时

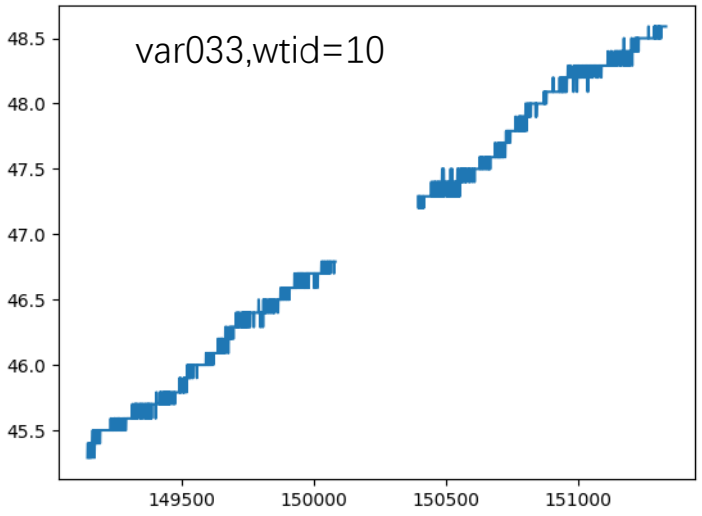


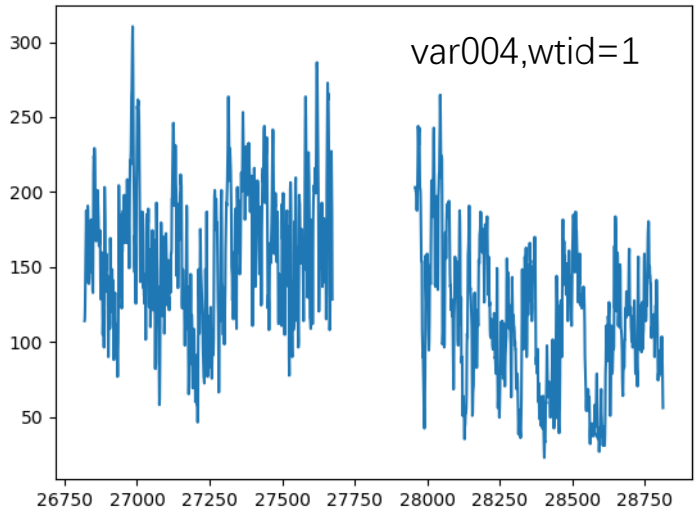


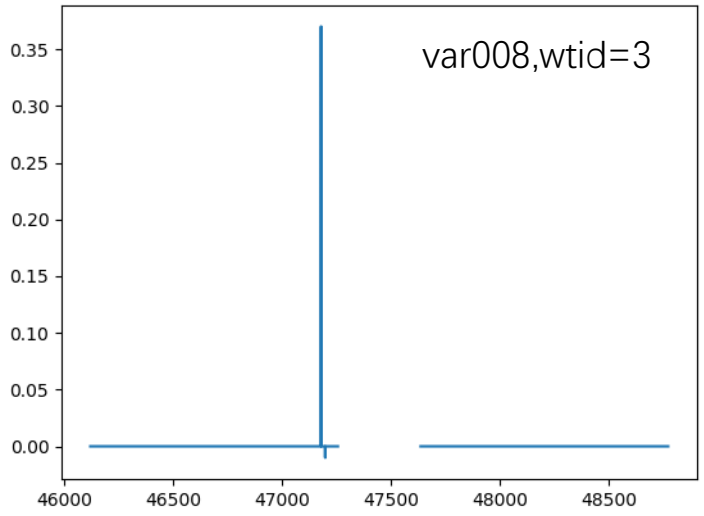
数据分析

赛题分析

单机组缺失形式：





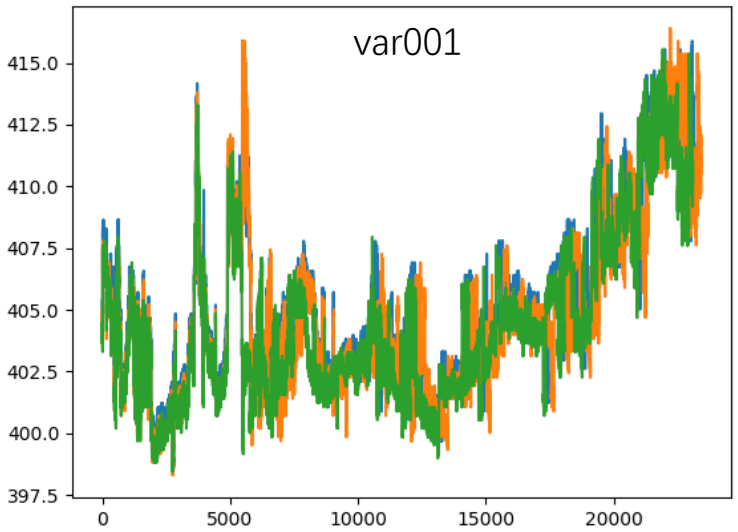


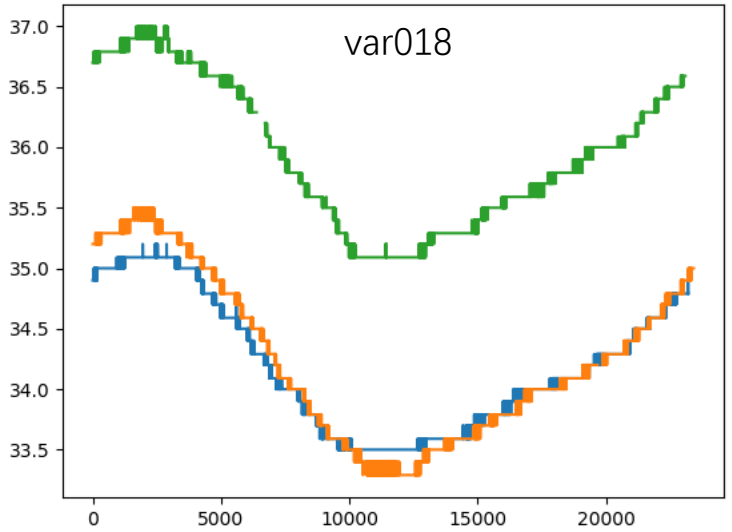
容易发现：线性插值适合var033系列，而最邻近插值法或者统计值（众数）填充适合var008系列，var004系列波动太大，可能对上述修复方法都不敏感。

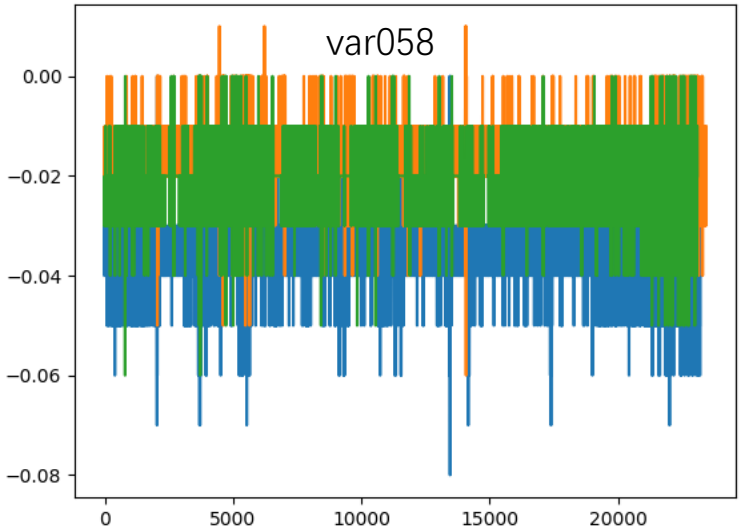
数据分析

赛题分析

机组间规律：由于各机组海上作业区域接近，故可能包含环境因素的变量时序趋势接近，可以相互填充。







通过三个机组某段趋势图发现，var001重合率高，后续可以尝试相互填充，var018,var058重合率低，且滞后趋势不统一。

验证集构建

赛题分析

线下验证集建立：

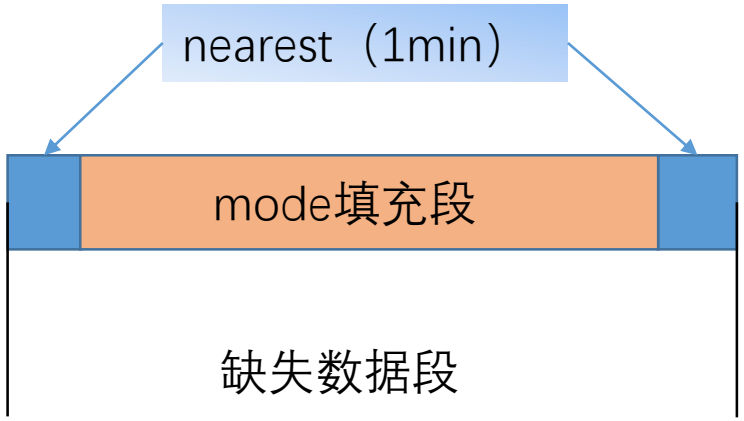
- 统计量填充(median，mode，mean)，线下以全量训练集做验证，结果保守，效果不错。

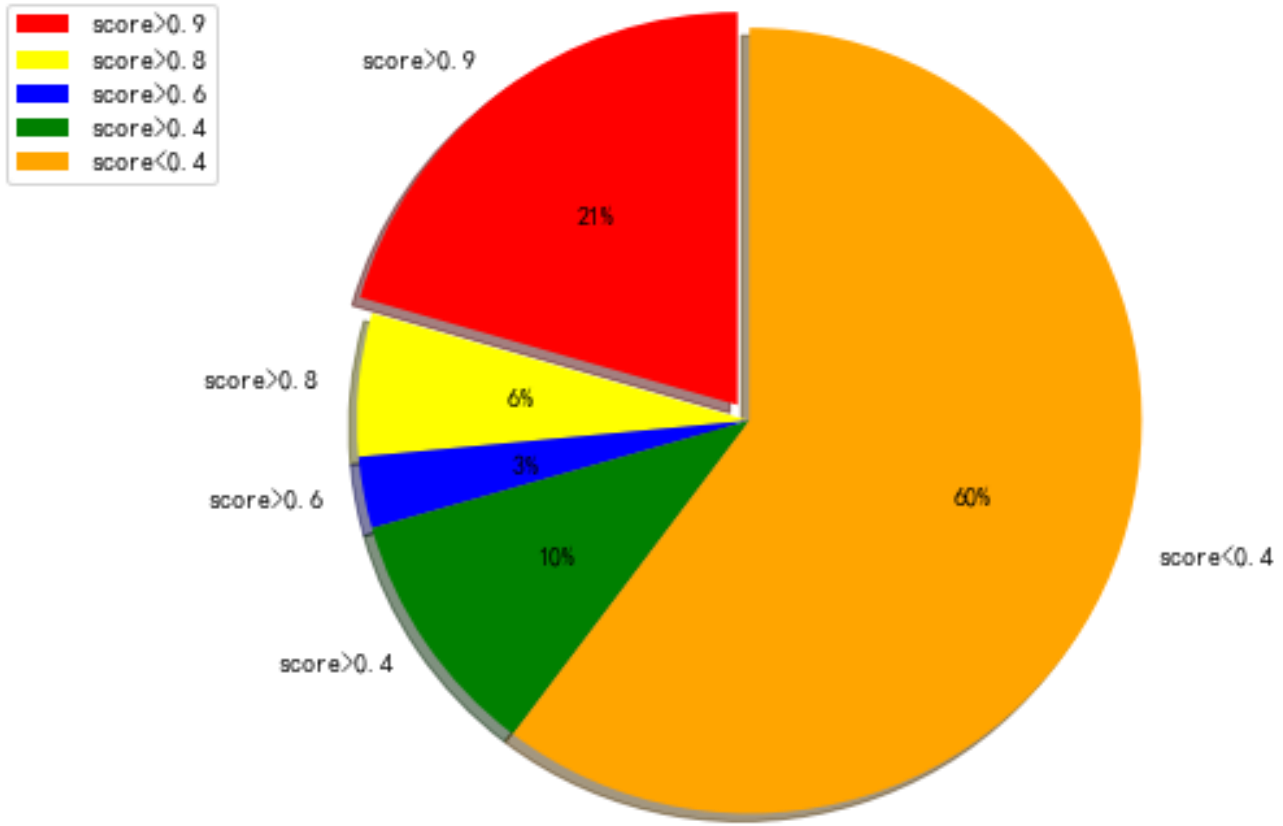
- 插值法以及模型填充：此处采用出题方开源数据删除程序，以训练集数据模拟删除，采用相应算法修复并以计算得分，重复5~10次取平均，保证线下线上一致性。

数据修复

统计量填充

右图为各机组每列众数填充得分均值的分布，有14列得分均值超过0.9，考虑到开源程序nearest插值填充的优异效果，起始和终止缺失位置，取间隔1min采用nearest填充，剩余众数填充。

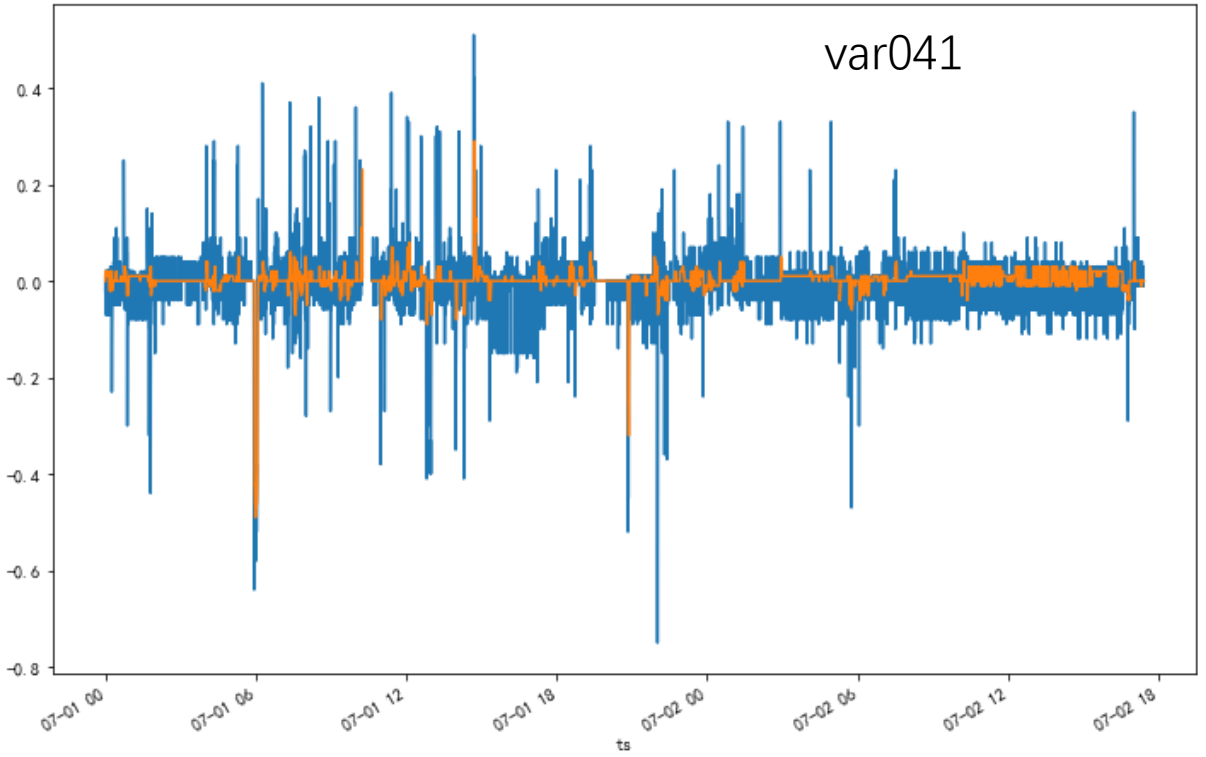




数据修复

统计量填充

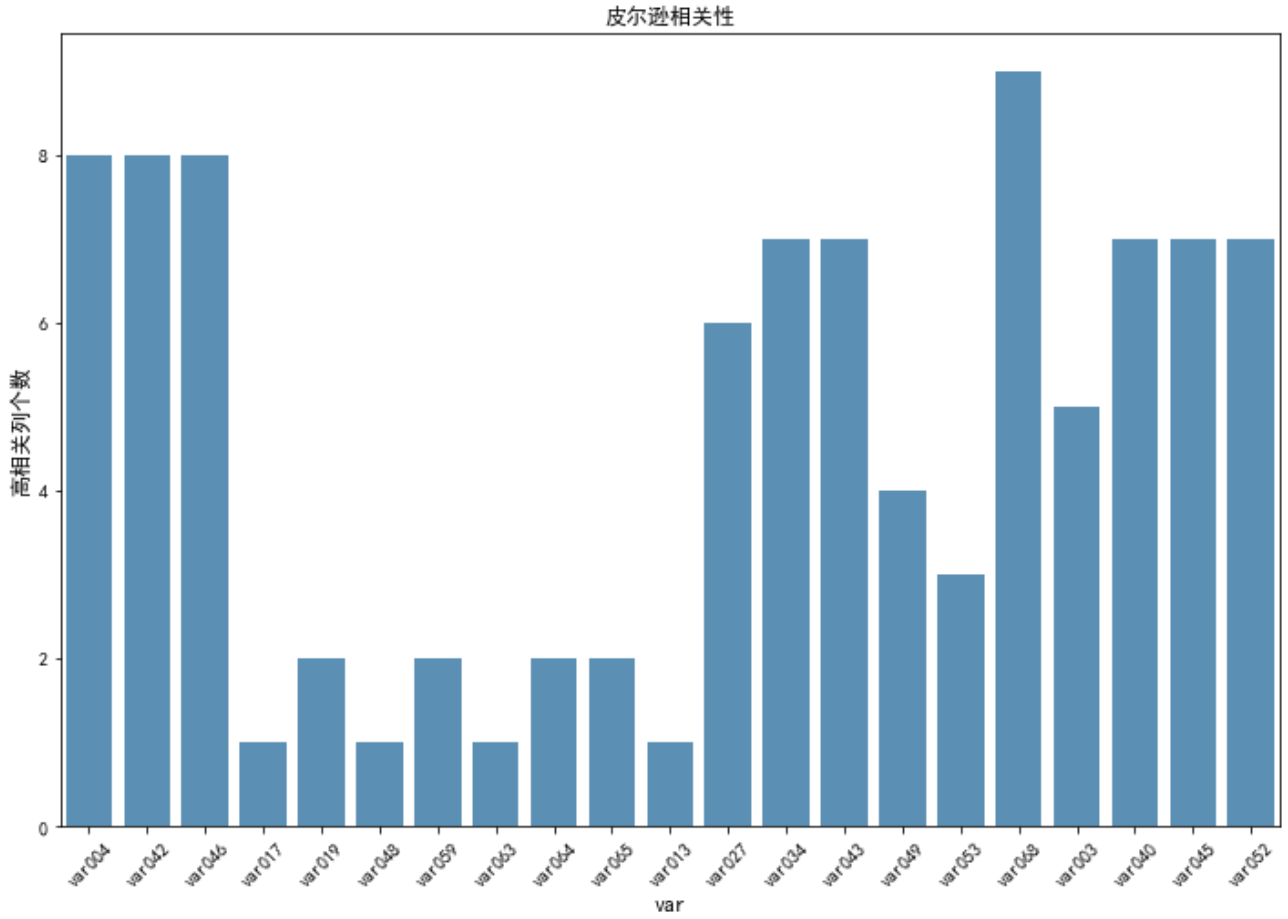
工业数据中，经常存在噪声，而此次数据的采样频率大约7s，故而存在大量噪声，常用的处理方法有移动均值，指数平滑，实质上是滤波。由于统计量填充效果均值<中位数≈众数，同时为了提高效率，选用中值滤波，过滤后的数据通过线下验证选用相应的修补方法，线上得到了很大提升。



数据修复

模型填充

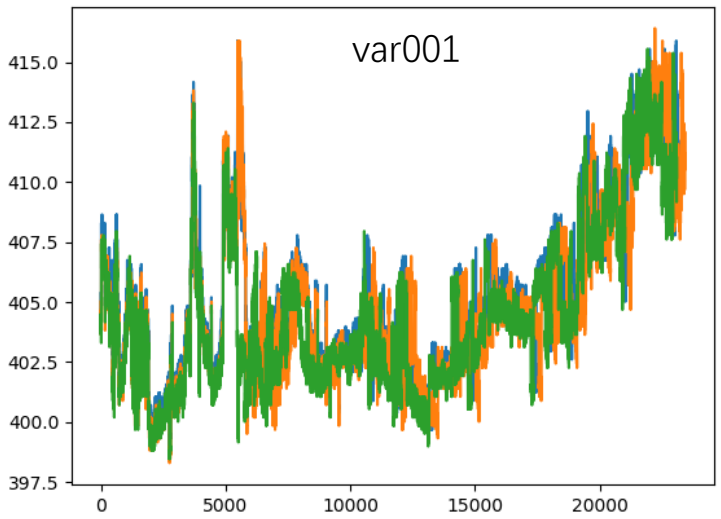
- 机组同时缺失列个数有8，15，19，26四种以及它们的组合缺失。以上述四个缺失分组为基础，选取组内与组外相关性超过0.7且高相关性列数多的列，如var004,var042进行建模。此外考虑到缺失为8的分组，保留了大量完备特征数据，这8列可选为建模列。
- 建模策略：选用lightgbm模型，多机组数据合并建模，提高效率，准确性。

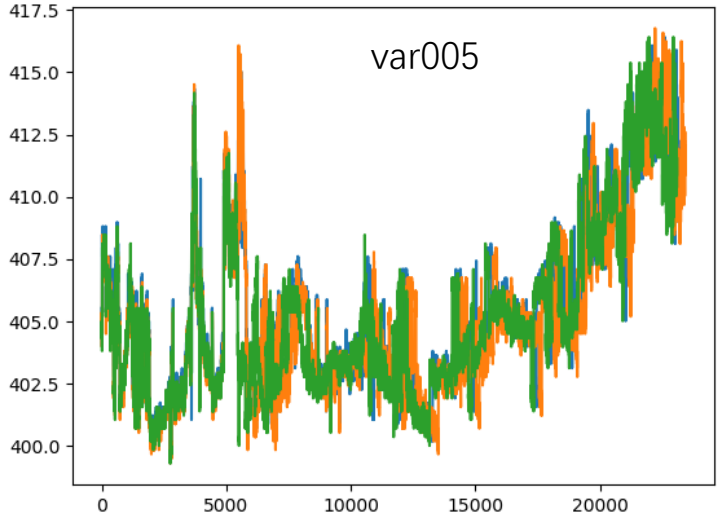


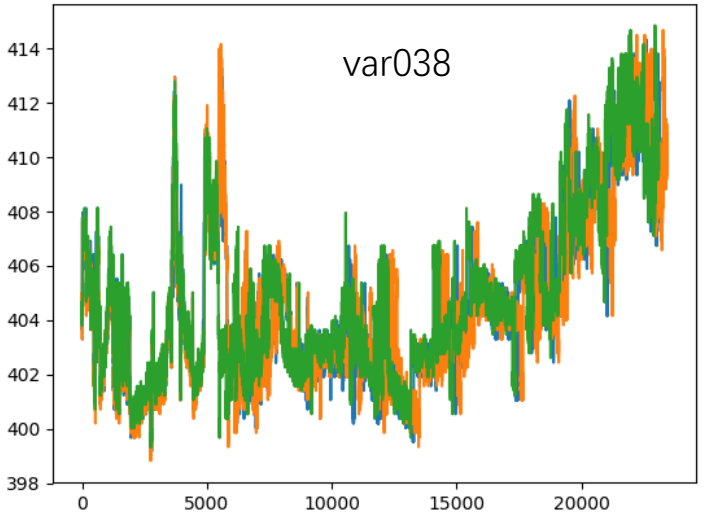
数据修复

机组相互填充

分析不同机组同一变量相关性，选取适合相互填充的列，以剩余机组该变量作为特征，进行模型填补，下图几列线上提升明显。



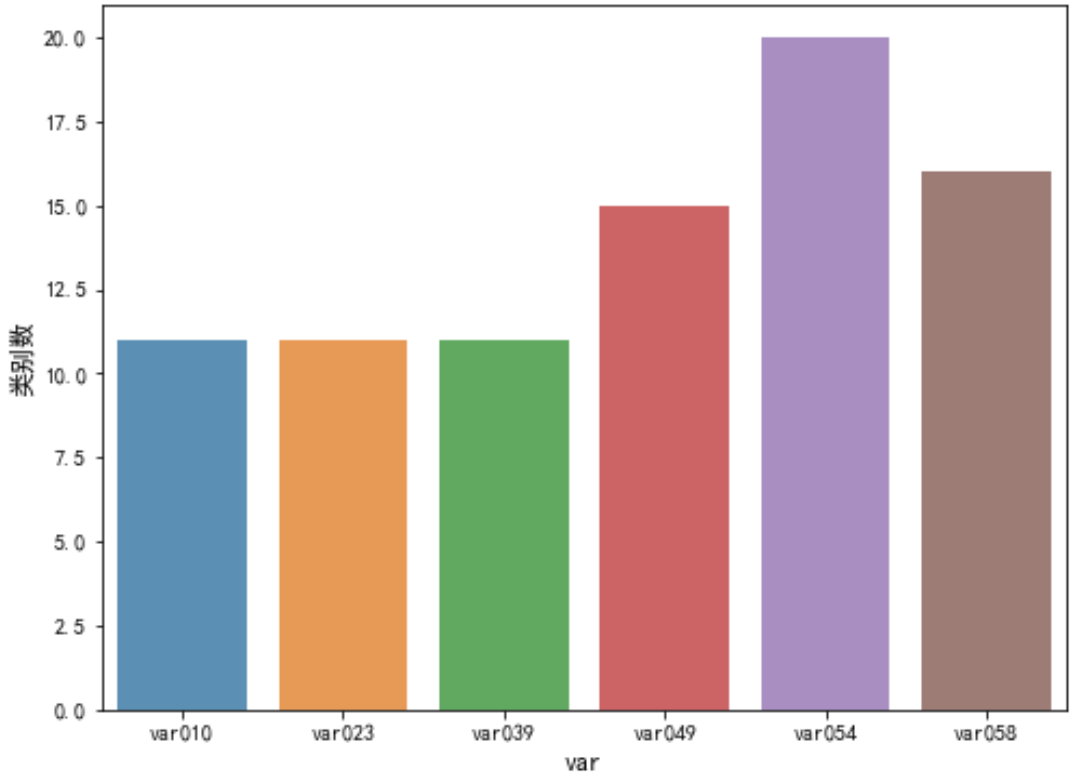




数据修复

- 对模型填充列采用平均融合，有小幅提升
- 通过对评分函数的分析，采用替换法，针对特定列结合线上线下效果使用好的结果替换差的

数据修复



- 对于取值类别少并且值域分布在0附近的浮点型列，预测错误就不得分相当于类别列，通过分析发现先以小时众数修复部分缺失，剩余缺失填充天众数，线上提升很大。
- 尝试对这些列做多分类填充，少数机组线下效果不理想。
- 众数为0的列：0预测错误得分相当于0，一般要后处理或者直接采用众数填充。

总结展望及致谢

- 赛题新颖，要多方探索解决方案!
- 数据的分析处理至关重要!
- 有效的线下验证，是稳步提升的关键!
- 不到最后一刻坚决不会放弃!

感谢主办方以及DataFountain平台提供的参赛机会！