DataFountain

DCIC 2019数字中国创新大赛 Digital China Innovation Contest

海上风场SCADA数据缺失智能修复

硬刚队

刘明桓

成员介绍 - 硬刚队

戴玮

中科院自动化所模式识别国家重点实验室博士生

研究方向：遥感目标的检测与分割

Team Leader

许凡凯

华南理工大学机器学习实验室硕士生

研究方向：多目标追踪

刘明桓

西南交通大学大四本科生
上海交通大学预录取博士

研究方向：强化学习、多智能体系统

邓金红

西南交通大学大四本科生
电子科技大学预录取硕士

研究方向：迁移学习、计算机视觉

最终成绩 - A / B

排名	排名变化	队伍名称	最高得分
1		上来就是一串代码	0.71018840
2	↓ 1	太帅了，求你报警	0.70856076
3		一方通行	0.70820690
4	↑ 2	三蹦子	0.70513830
5	↑ 2	DaciLab	0.70146054
6	↑ 3	反正我扛不住	0.70044020
7	↓ 6	UnionT	0.69487340
8		xili	0.69304890
9		宇智波斑·砖侠	0.67894280
10	↓ 4	lzhexp	0.67470586
11	↑ 4	硬刚队	0.67211250

A-11-0.67211250

排名	排名变化	队伍名称	最高得分
1		太帅了，求你报警	0.71077900
2	↓ 1	上来就是一串代码	0.71073280
3	↓ 1	一方通行	0.70943660
4		三蹦子	0.70561100
5	↓ 1	DaciLab	0.70187020
6		反正我扛不住	0.70102364
7	↓ 5	xili	0.69829345
8		UnionT	0.69535100
9	↑ 1	dcic	0.68348885
10	↑ 1	硬刚队	0.68113250
11		宇智波斑·砖侠	0.67941760

B-10-0.68113250 ↑1

目录

- 问题描述
- 基本思路 – 3 way
- 数据分析
  - 变量关系分析
  - 站点关系分析
  - 缺失情况分析
  - 异常值分析
- 最终思路 – 2 way
- 方案描述
  - 实验细节
  - 模型选择
  - 特征构造
  - 实验结果
- 总结展望

问题描述

- **数据**：数据：海上风电场监控数据（数据名称已脱敏），数据记录周期约为7秒（部分为3秒）
- **预测**：所有变量缺失处的数据
  - 包括短期缺失和长期缺失
  - 包括部分变量缺失和所有变量缺失
- **评价标准**：所有缺失行的分数平均值
  - 某一缺失行的分数是所有缺失变量的分数平均值
  - 部分变量缺失和整行缺失的收益相同
  - 只预测少量部分缺失也会有不错的分数

$$F = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i} f_{i,j}(x_{i,j}, \hat{x}_{i,j})$$

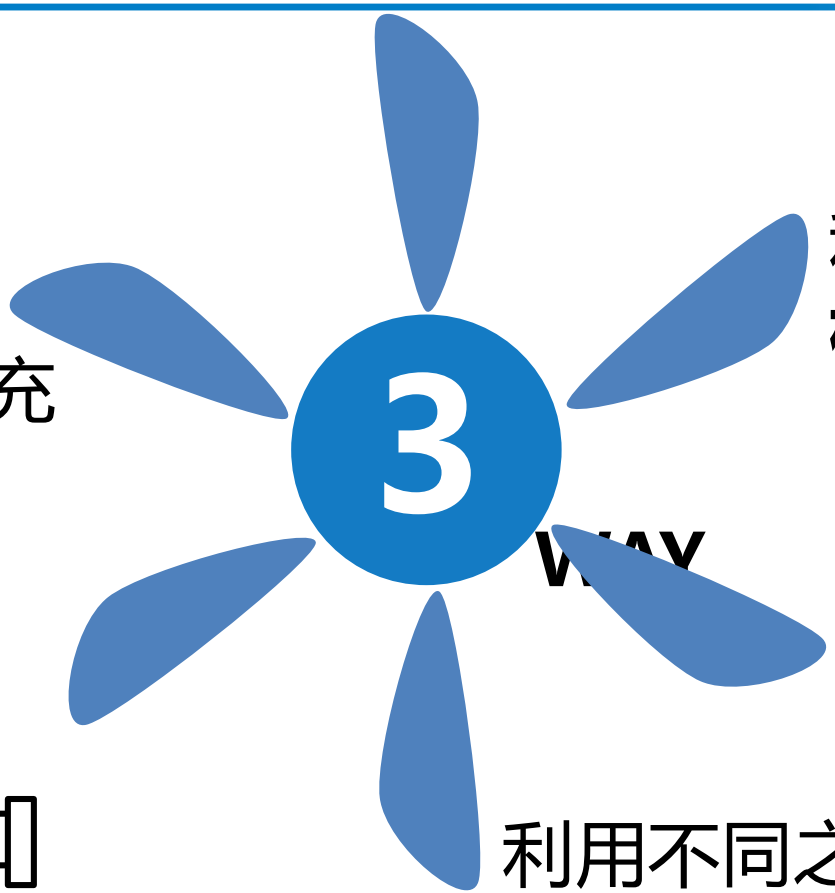
$$f_{i,j}(x_{i,j}, \hat{x}_{i,j}) = e^{-\frac{100|x_{i,j}-\hat{x}_{i,j}|}{max(|x_{i,j}|, 10^{-15})}}$$

浮点

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

布尔/枚举

基本思路 - 3 way

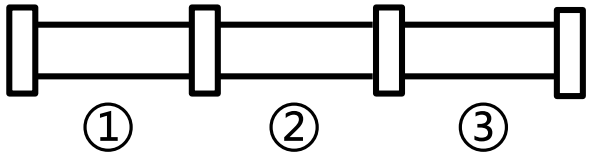


利用变量自身的
时间序列关系填充

利用站点之间的
相似关系填充

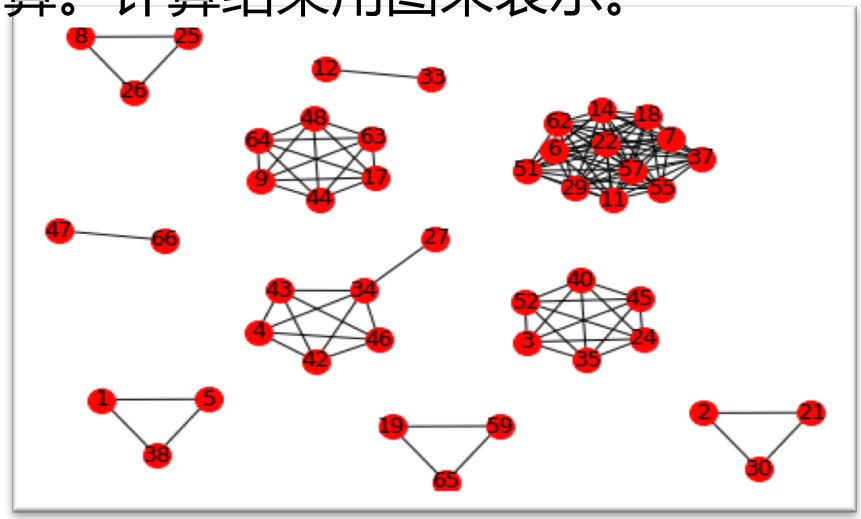
利用不同之间的
特征关系填充

Design a Pipeline

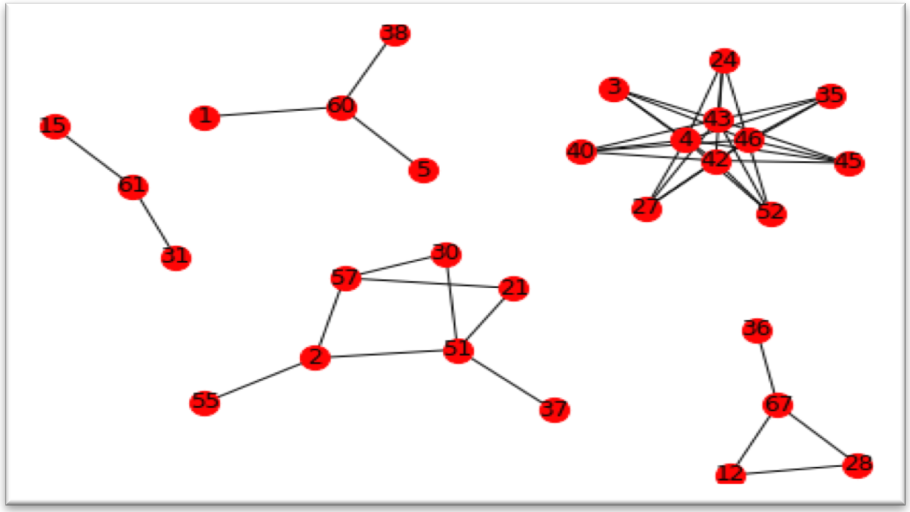


数据分析 - 变量关系分析

- **相关性（ Correlation ）**：在概率论和统计学中，相关性显示两个随机变量之间线性关系的强度和方向。可以用来衡量两个变量相对于其相互独立的距离。
- 我们对**同一个发电机**风阻的不同变量进行相关性的计算。因为每一个发电机组收集到的数据都极为相似，这里，我们取了wtid为1的数据进行相关性的计算。计算结果用图来表示。



相关性≥0.9



0.8≤相关性<0.9

数据分析 - 站点关系分析

- 由于33个发电机组是位于**同一个**区域，那么每一个机组记录的数据也会有比较强的**相关性**。在这里，我们对同一个变量，不同发电机组记录的数据进行分析。如下图，我们统计变量“var004” 在33个不同机组的分布情况。

var004 * 33

- 从图中不难发现站点之间的相似性。
- **但是**由于每一个机组采集数据的**时间**都不一样，甚至时间间隔也**不稳定**。而且，从不同站点中查找相关变量的**计算成本**较高。
- 在进行尝试之后发现效果并**不理想**。在找不到有效利用33个发电机组之间关系的方法的情况下，我们**暂时不考虑**这个突破口。

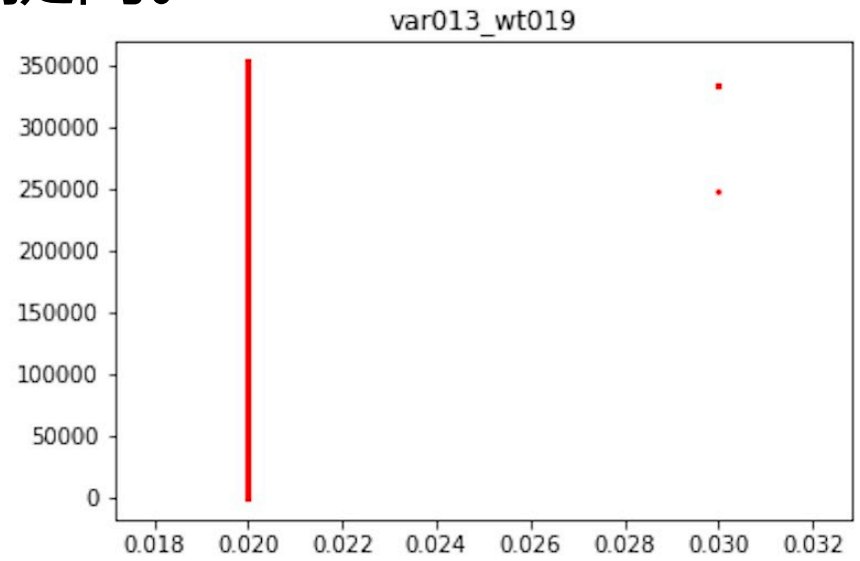
数据分析 - 变量缺失分析

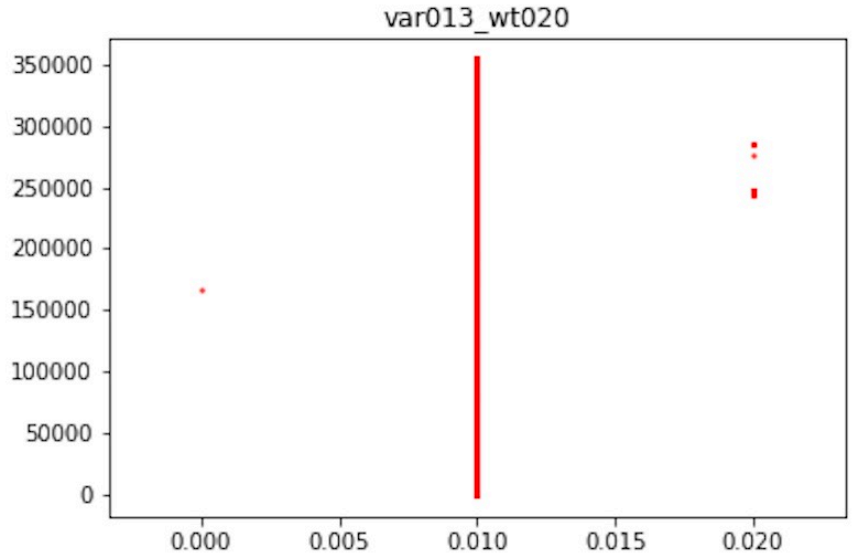
- **横向缺失信息**：提交的49万多条数据中，有大约**29万条**是全部缺失，21万条是部分缺失。
  - **部分缺失**的数据，我们可利用**变量之间的关系**进行计算。
  - **全部缺失**的数据，无法通过同一个风车组变量之间的关系进行填充值的计算，只能利用数据，**构造相关特征**，通过训练模型来对空的数据进行预测。

- **纵向缺失信息**：最多有**11400条**约为**22.16小时**的连续整行缺失。
  - 因此在利用**纵向历史时序**特征时最少要构造一天前的特征。

数据分析 - 异常值分析

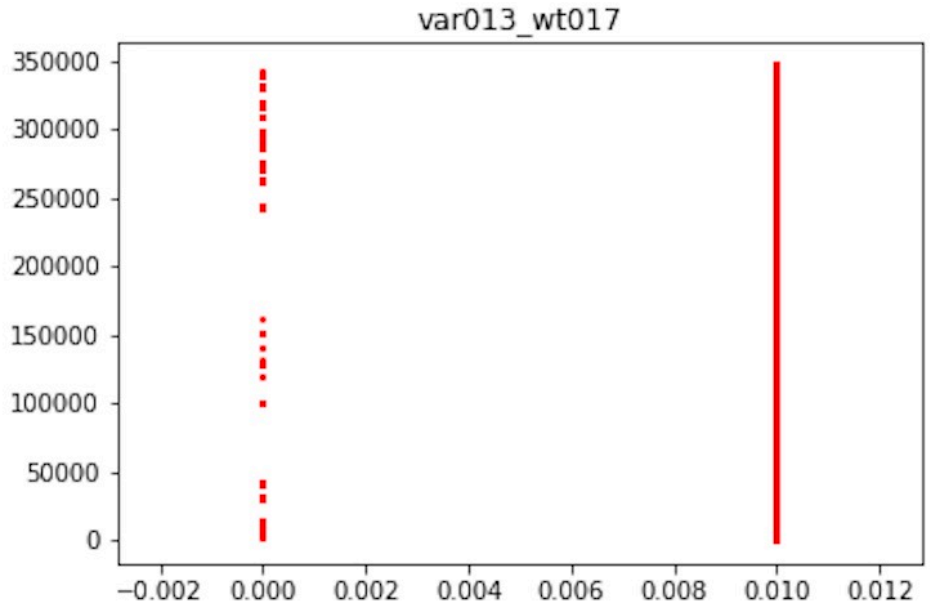
- 在数据分析时，我们发现数据存在**孤立点**，在前期处理中我们将其作为异常点进行去除。处理时利用数据的较小四分位点和较大四分位点进行异常值去除。并在初期取得了提高。

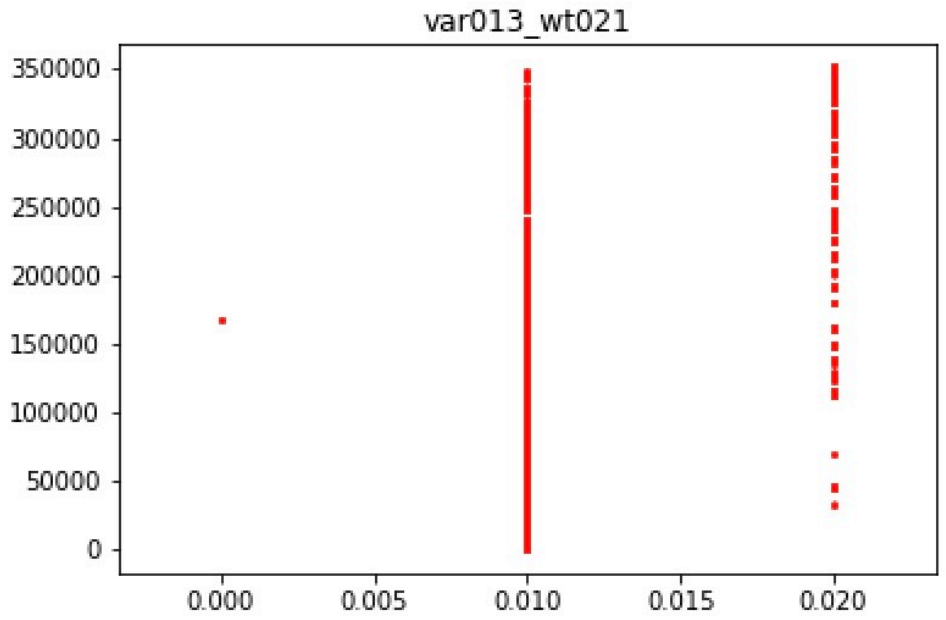




数据分析 - 异常值分析

- 但在后期我们发现，这样的孤立点可能**并不是**异常值，这样处理会带来信息的缺失并影响最终的结果。但后续没有时间进行改进了。

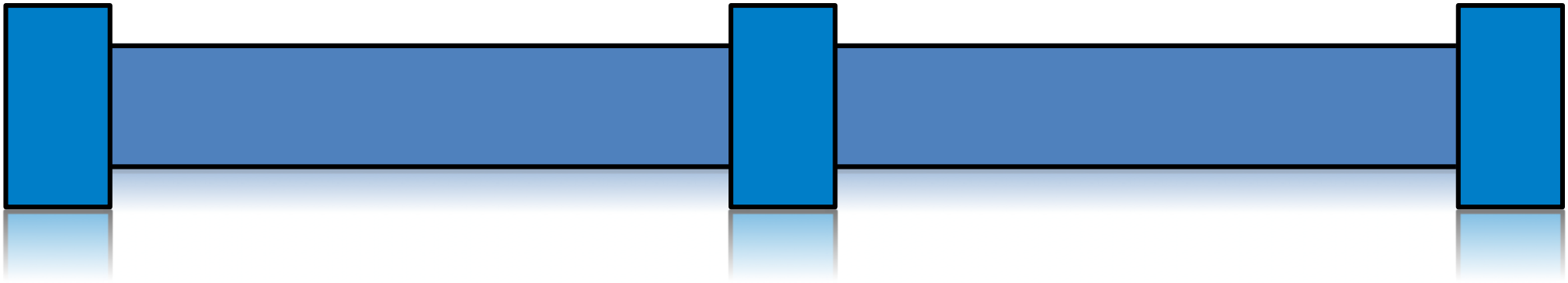




最终思路 - 2 way pipeline

1. 对于相关性较高的变量且没有全部同时缺失的，我们变量之间的关系进行插值。

2. 对于1无法填充的（e.g. 变量同时缺失），我们构造特征模型进行填充。



方案描述 - 实验细节

1

对于相关性较高的变量且没有全部同时缺失的，我们利用变量之间的关系进行插值。

①进行按比例的变化进行插值。 如var004为空值，去相关性>0.9的其它变量中寻找同时间段的，计算它们从上一个值到当前值的变化率，记录下所有变量的变化率，若为空则跳过。最后，对求得的所有变化率求均值。

②利用步骤①变化率和与当前要进行插入的上一个值进行计算，得出当前值，并插入。

③以上步骤先向上搜索，上一个值如果都为空，再向下搜索，如果都为空，那么跳过。

405.37	405.02
405.02	

如现在求空值。设空值为x，则

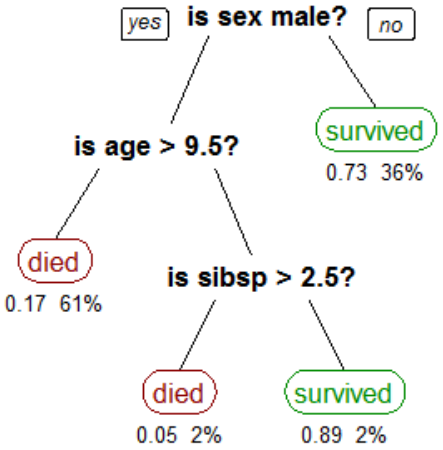
x = (405.02-405.37)/405.37 *405.02 +405.02

方案描述 - 模型选择

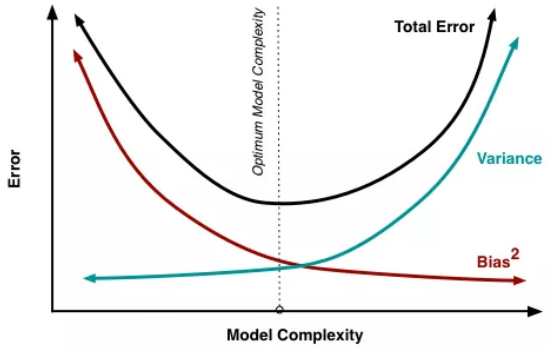
2

对于1无法填充的（e.g. 变量同时缺失），我们构造多维特征模型进行填充。

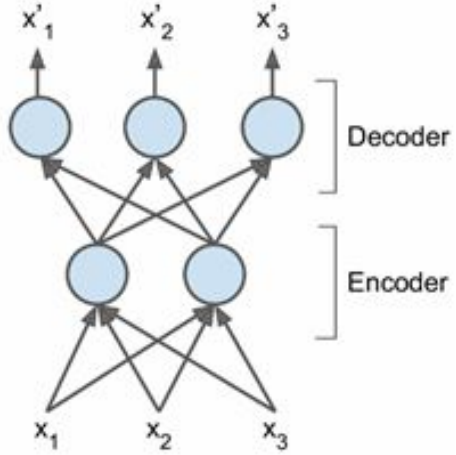
我们选择了几个模型作为候选，并进行了尝试。



Decision Tree



GBDT



AutoEncoder

最终经过比较，我们选择了**Decision Tree**和**GBDT**作为我们采用的模型。

方案描述 - 特征构造

2

对于1无法填充的（e.g. 变量同时缺失），我们构造多维特征模型进行填充。

我们从多个角度，构造了多维特征。

基本特征

1. 距离初始点的时间长度；
2. 星期、小时、分钟的sin/cos特征；
3. 站点序号

其他变量特征

其他变量前后10十条数据的均值

时序特征

该变量前后1分钟的数据特征

方案描述 - 实验细节

2

对于1无法填充的（e.g. 变量同时缺失），我们构造多维特征模型进行填充。

- 在实际实验中，我们从易到难构造了多个特征模型，并进行了融合。

我们采用baseline逐级更新的方式进行融合。

- 在实际实验中，我们使用部分列替换的方式进行融合。

我们利用变量相关性进行整批替换，并在线下和线上验证分数，以判断模型对某列变量的拟合效果。

方案描述 - 实验结果

2

对于1无法填充的（e.g. 变量同时缺失），我们构造多维特征模型进行填充。

在实际实验中，我们从易到难构造了多个特征模型，并进行了融合。我们采用baseline逐级更新的方式进行融合。详细如下：

名称	方法	A榜分数	特征
baseline1	最近邻插值	0.6433xxxx	-
baseline2	GBDT	0.6648xxxx	基本特征1，3
baseline3	DT	0.66869360	基本特征1，3
baseline4	GBDT	0.67071736	基本特征
baseline5	GBDT	0.67210830	基本特征+时序特征
baseline6	GBDT	0.67211250	基本特征+其他变量特征

方案描述 - 实验结果

1

对于相关性较高的变量且没有全部同时缺失的，我们利用变量之间的关系进行插值。

在B榜上采用该方法的结果如下。

名称	方法	B榜分数	备注
baseline6	GBDT	0.6724xxxx	无
baseline7	相关变量搜索插值	0.6811325	只对[4,27,34,42,43,46]进行了处理

总结展望

总结

我们设计了简单的pipeline对数据进行填充。首先，对于相关性较高且没有同时缺失的变量，我们利用相关变量插值进行填充；其次，对于一般数据，我们构造了多维特征，并训练了多个GBDT、DT等模型进行融合。

展望

1. 我们的异常值处理有缺陷，导致某些波动列被处理的较差，损失了很多信息，提升空间0.005-0.01。
2. 我们的pipeline设计比较简单，应该再设计一些利用其它站点特征进行插值的方式，提升空间0.001-0.005。
3. 我们对于某些短时剧烈波动列（e.g., var004）预测精度不够，因此我们的模型特征构造还可以更加细化，提升空间0.001-0.01。