南京大学 NANJING UNIVERSITY | 工程管理学院 SCHOOL OF MANAGEMENT & ENGINEERING

数据分割与强规则过滤

汇报人：杨超凡

小组成员：王一伦、钱创杰、温子祺、陈波

PART ONE

目录

CONTENTS

工作回顾


特征选择-过滤式
概念
规则
特征的发散性
特征与目标的相关性
方差选择法
概念
缺陷
适用于离散型变量
效果不佳
场景：分类问题
实现
VarianceThreshold
自行实现
挑选：方差大的特征
相关系数法
相关系数理解
计算公式
$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$
局限：线性关系以及原因
场景：回归问题
实现
corr(method='pearson')
scipy.stats.pearsonr()
自行实现
挑选：相关系数值绝对值大的特征
卡方检验法
原理
场景：分类问题
实现：SelectKBest 和 chi2
挑选：卡方值大的特征
互信息法
概念
与相关性的区别
场景：分类、回归
离散型
$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$
计算公式
连续型
$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy$
挑选：I(X, Y) 值大的特征
基于模型的特征排序
思路
步骤
场景：分类、回归
实现
Relief
思路
算法

工作回顾

以 15 号风机各变量方差的百分之二十五分位数作为阈值进行过滤，剔除掉方差小于阈值的变量，保留21 个显著变量。

```
['wind_speed',
 'generator_speed',
 'power',
 'wind_direction',
 'wind_direction_mean',
 'yaw_position',
 'pitch1_angle',
 'pitch2_angle',
 'pitch3_angle',
 'pitch1_moto_tmp',
 'pitch2_moto_tmp',
 'pitch3_moto_tmp',
 'acc_x',
 'acc_y',
 'environment_tmp',
 'int_tmp',
 'pitch1_ng5_DC',
 'pitch2_ng5_DC',
 'pitch3_ng5_DC',
 'group',
 'timestamp']
```

如果属性变量太多容易导致过拟合，为提高模型泛化能力通过卡方检验进一步筛选，最后保留16个原始变量。

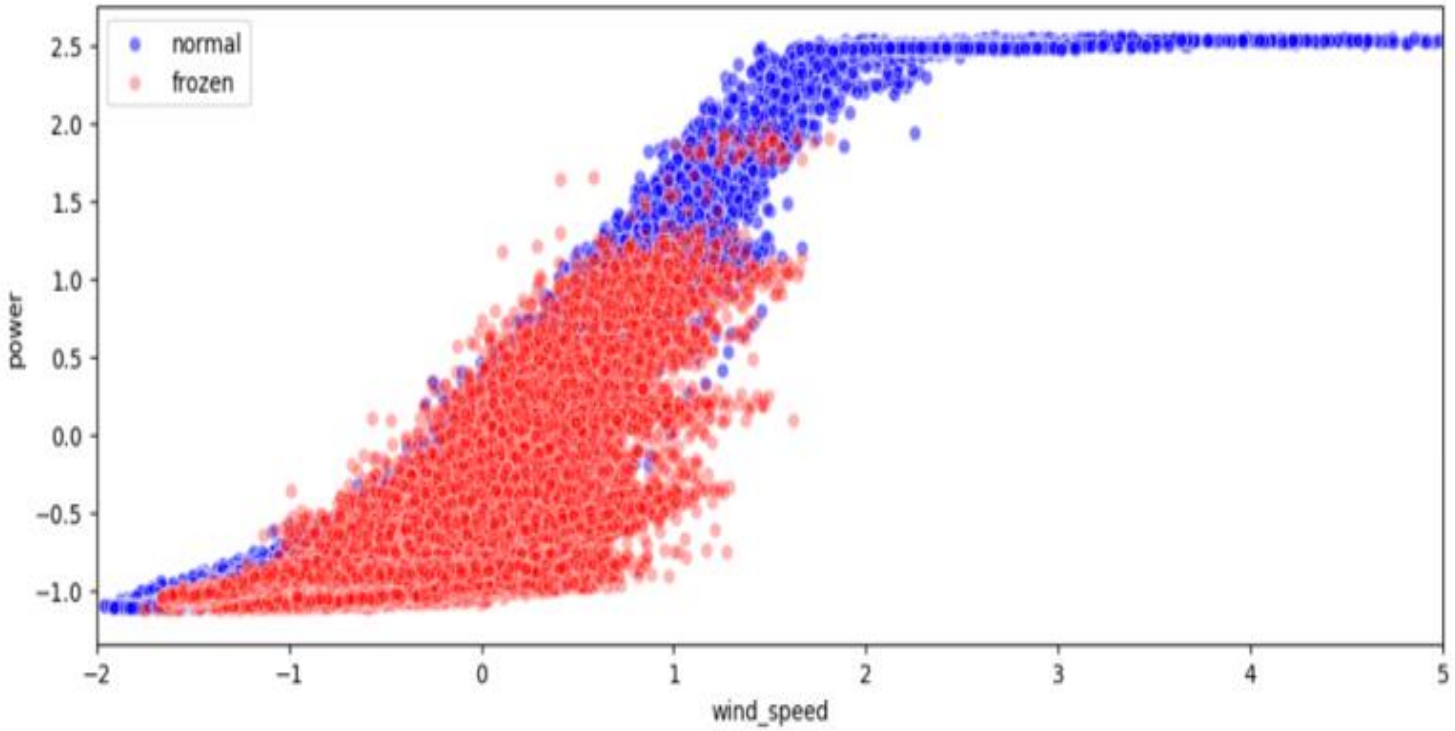
```
['wind_speed',
 'generator_speed',
 'power',
 'wind_direction',
 'wind_direction_mean',
 'yaw_position',
 'pitch1_angle',
 'pitch1_moto_tmp',
 'pitch2_moto_tmp',
 'pitch3_moto_tmp',
 'acc_x',
 'acc_y',
 'environment_tmp',
 'int_tmp',
 'pitch1_ng5_DC',
 'group']
```

工作回顾

Model	RandomForest	DecisionTreeClassifier	GBDT	GBDT+LR	RF+LR
Cross_val_score	0.756	0.808	0.844	0.899	0.869

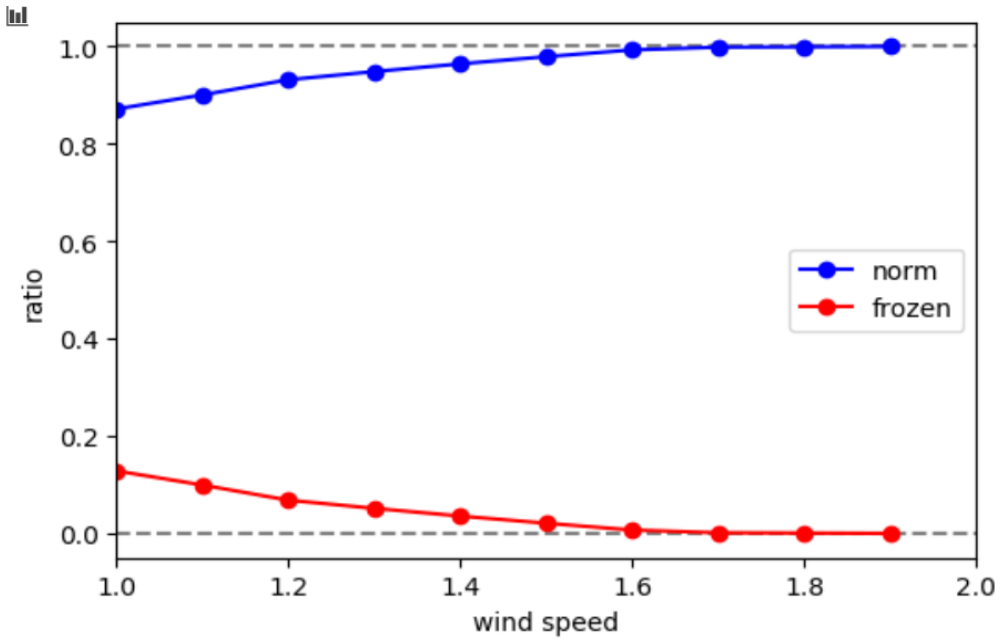
强规则过滤

基于特征属性



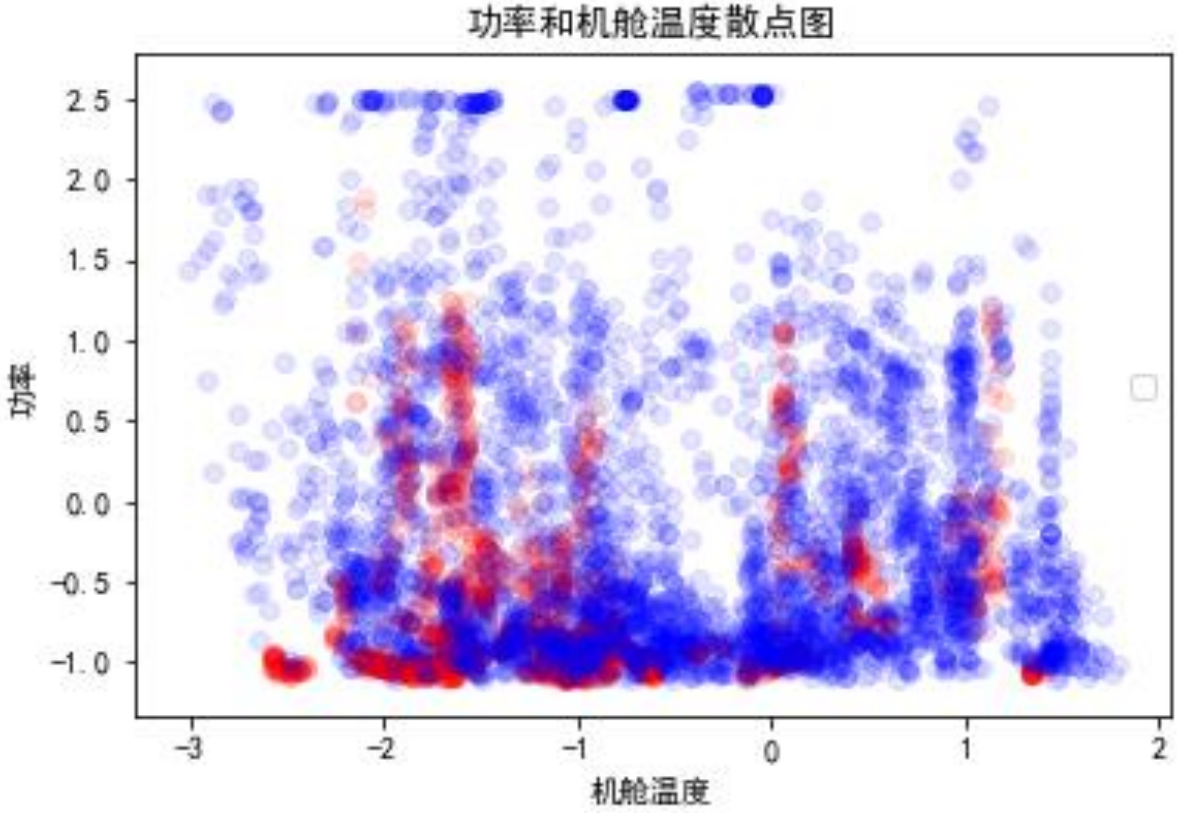
可以看出在风速和功率<2的范围内，模型对两类样本的区分度不大。

强规则过滤



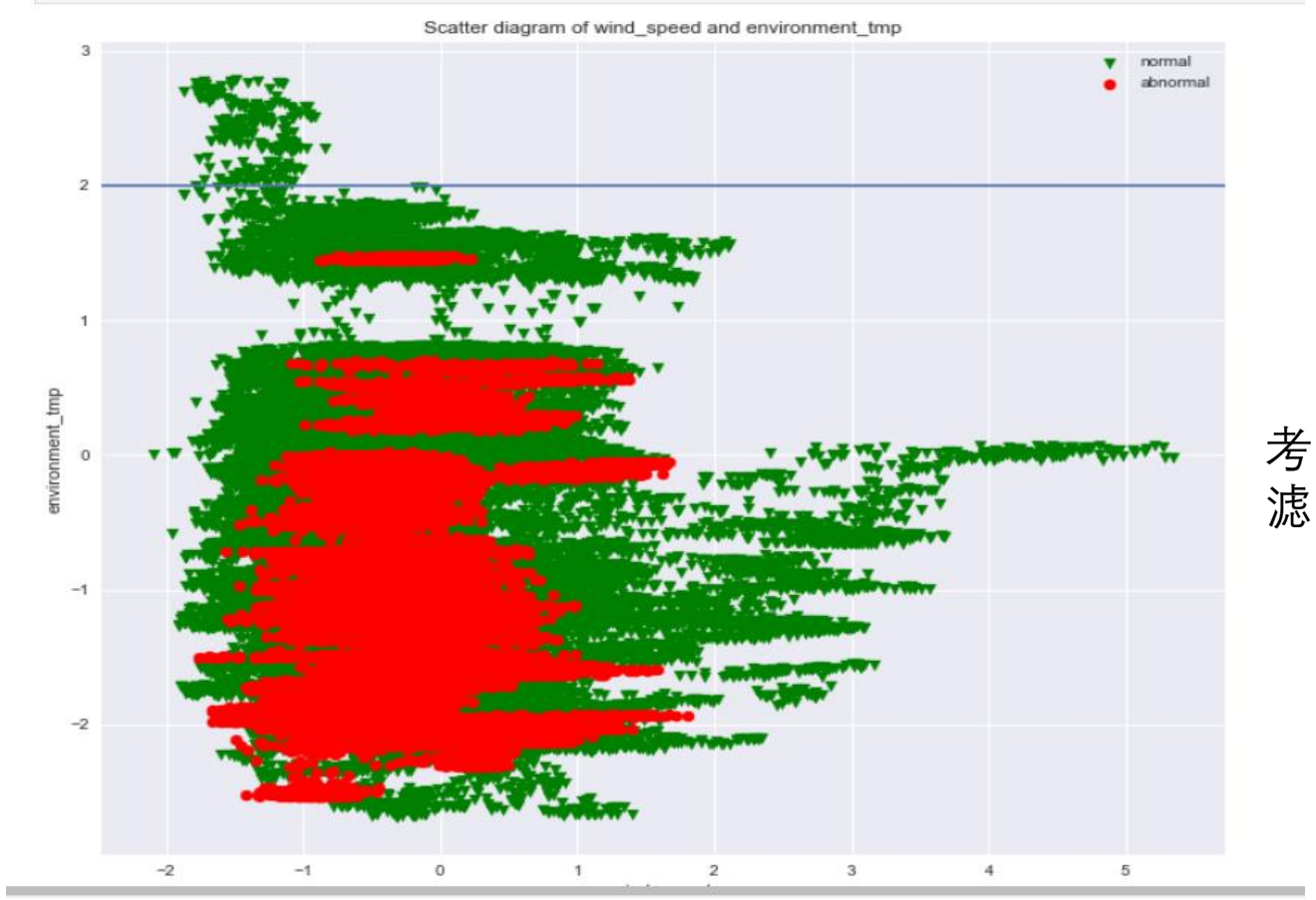
当风速较大时，风机为正常的比重较高，当风速为1.以上的时候正常风机的比例为100%，认为可剔除风速>1.9的数据。

强规则过滤



当机舱温度大于1.5或小于-2.5时，数据基本为不结冰数据，可以考虑将“int_tmp>1.5 & int_tmp<-2.5”作为强规则对数据进行过滤。

强规则过滤



当环境温度大于2时，数据基本为不结冰数据，考虑将“environment_tmp>2”作为强规则对数据过滤。

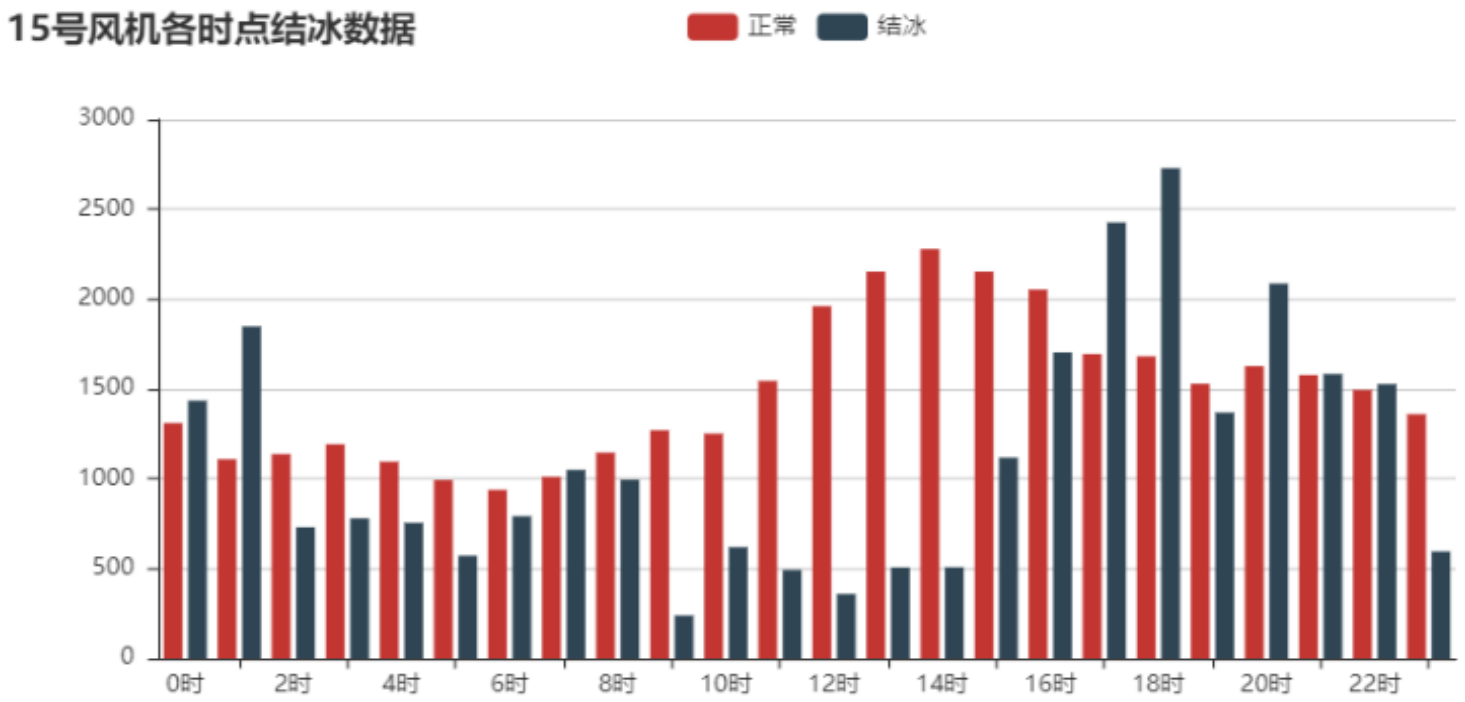
强规则过滤

综上所述，构建出四条强规则对数据进行过滤，保留满足以下四个条件的数据

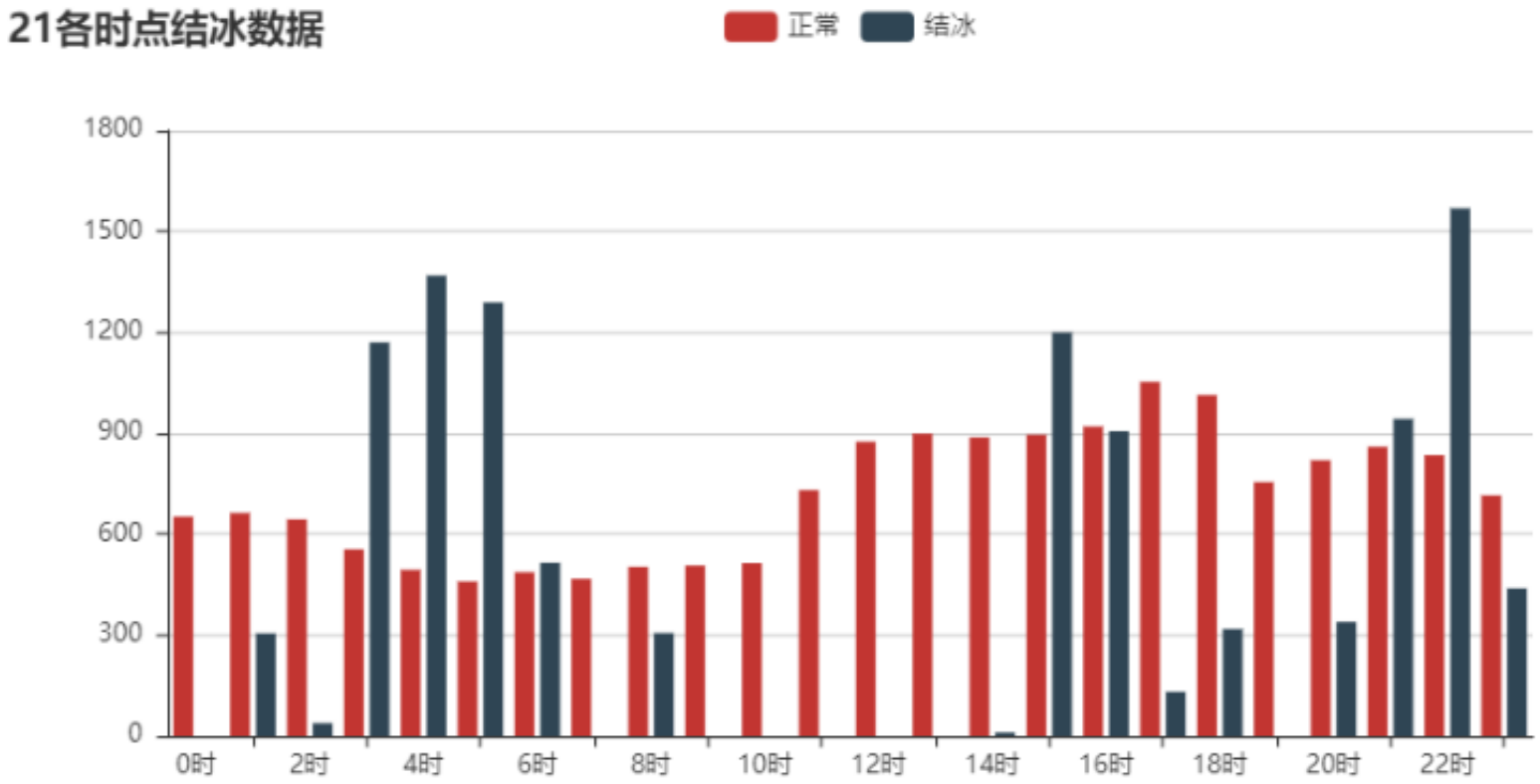
1. power<2
2. wind_speed<1.9
3. environment_tmp<2
4. -2.5 <int_tmp<1.5

数据分割

基于时间节点


15号风机各时点结冰数据
正常
结冰
3000
2500
2000
1500
1000
500
0
0时
2时
4时
6时
8时
10时
12时
14时
16时
18时
20时
22时

数据分割


21各时点结冰数据
正常
结冰
1800
1500
1200
900
600
300
0
0时
2时
4时
6时
8时
10时
12时
14时
16时
18时
20时
22时

数据分割

综上所述，我们尝试按数据时间点是否属于9-14时段对数据分割
两种策略：把period作为标签属性作为特征学习或分割后各自训练

```
data15['period']=1
data21['period']=1
for i in range(len(data15)):
    if data15.loc[i,'time']<9 or data15.loc[i,'time']>14:
        data15.loc[i,'period']=0
for i in range(len(data21)):
    if data21.loc[i,'time']<9 or data21.loc[i,'time']>14:
        data21.loc[i,'period']=0
```

优化效果与总结

强规则过滤优化效果

```
data15=data15[(data15['power']<2) & (data15['wind_speed']<1.9) &(data15['environment_tmp']<2)&(data15['int_tmp']<1.5)&(data15['int_tmp']>-2.5)]
```

```
data21=data21[(data21['power']<2) & (data21['wind_speed']<1.9) &(data21['environment_tmp']<2)&(data21['int_tmp']<1.5)&(data21['int_tmp']>-2.5)]
```

```
data15.shape
```

```
(62391, 33)
```

```
data21.shape
```

```
(28053, 33)
```

过滤前

```
data15.shape
```

```
(59109, 32)
```

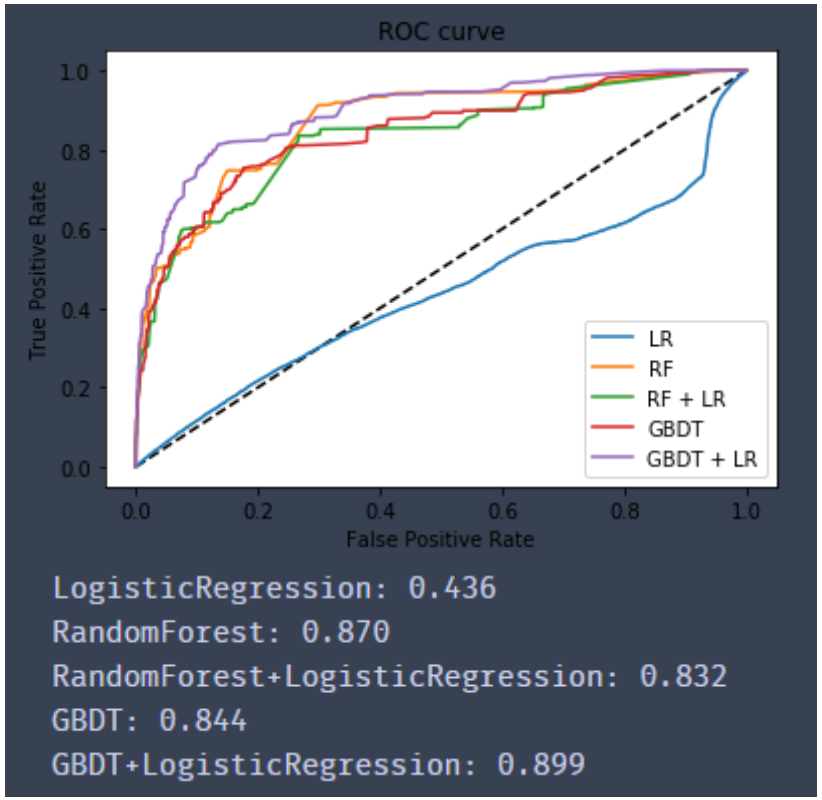
```
data21.shape
```

```
(24493, 32)
```

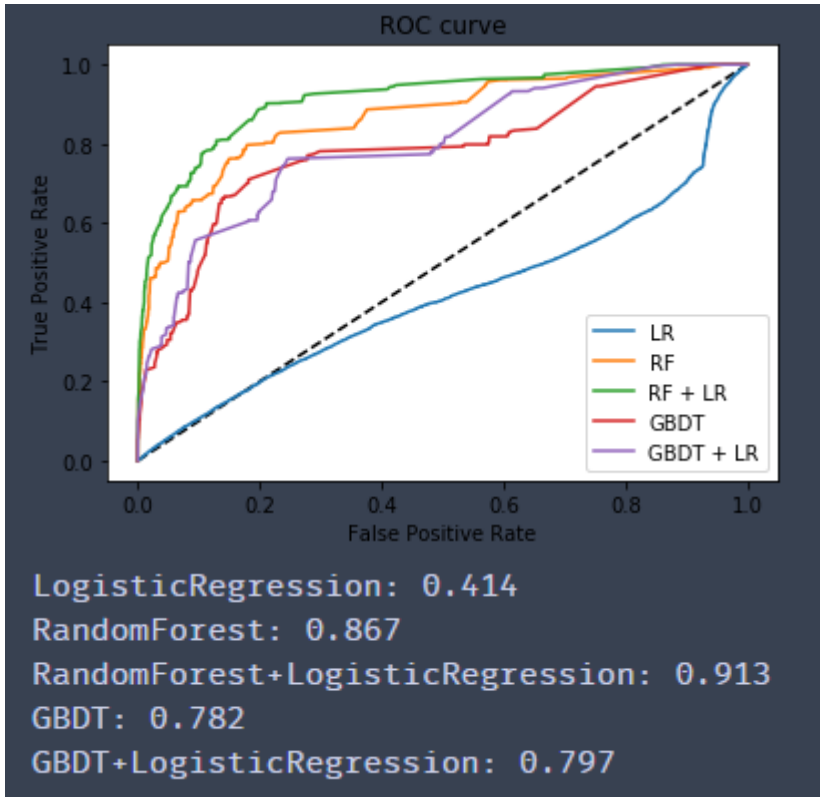
过滤后

优化效果与总结

强规则过滤优化效果



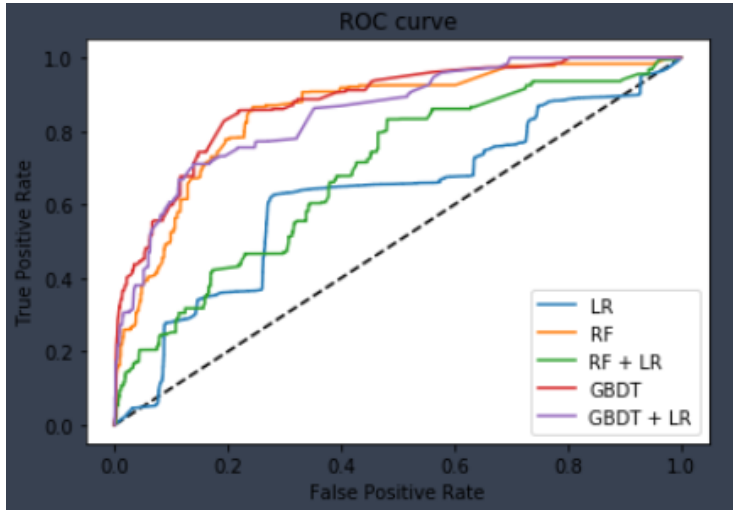
过滤前



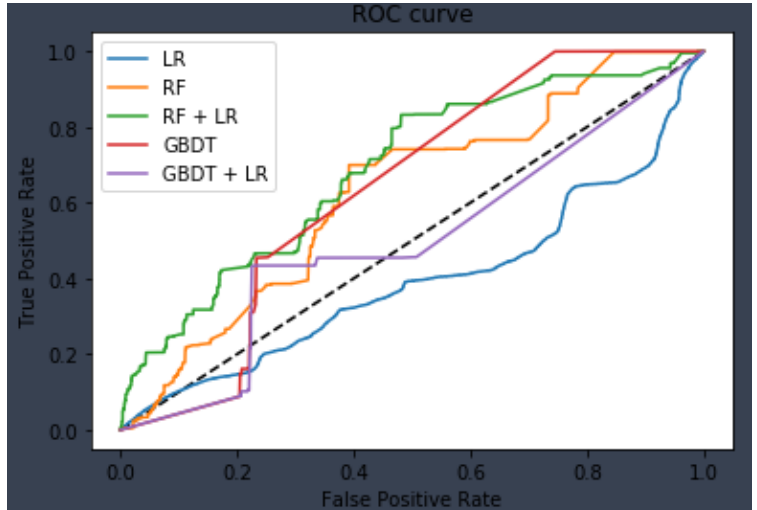
过滤后

优化效果与总结

优化效果（切割后预测）



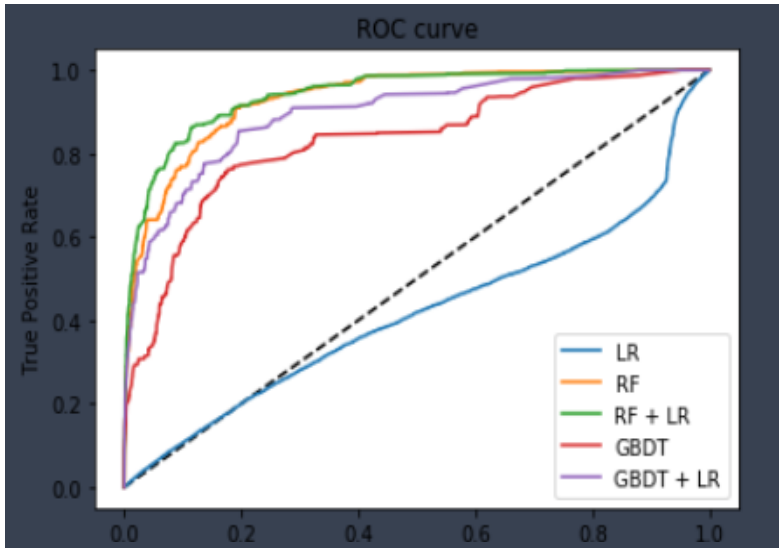
Period=0



Period=1

优化效果与总结

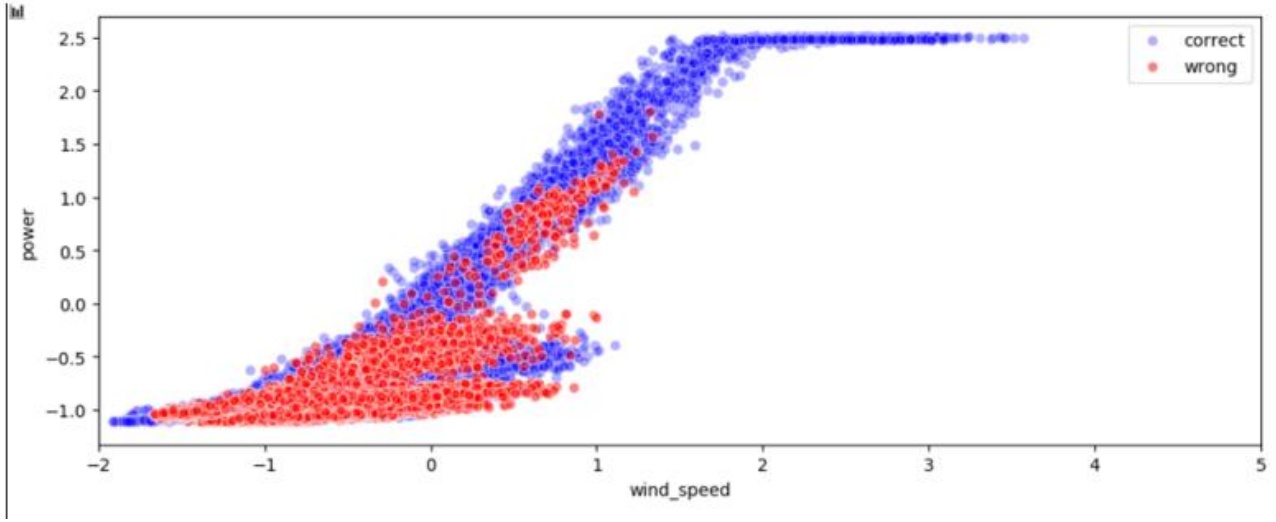
优化效果（把时段作为特征属性）



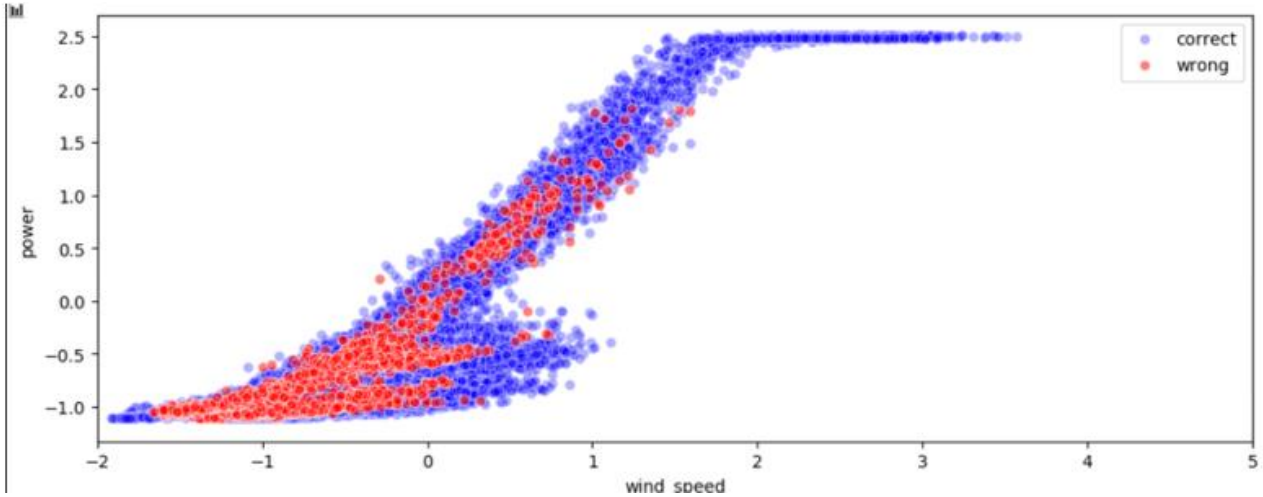

```
LogisticRegression: 0.417
RandomForest: 0.931
RandomForest+LogisticRegression: 0.942
GBDT: 0.831
GBDT+LogisticRegression: 0.895
```

优化效果与总结

从数据分割前后模型在训练集和测试集上的结果来看，通过强规则过滤和引入新的二分类变量可以一定程度上提高模型的预测能力，提升模型的泛化能力。



Before



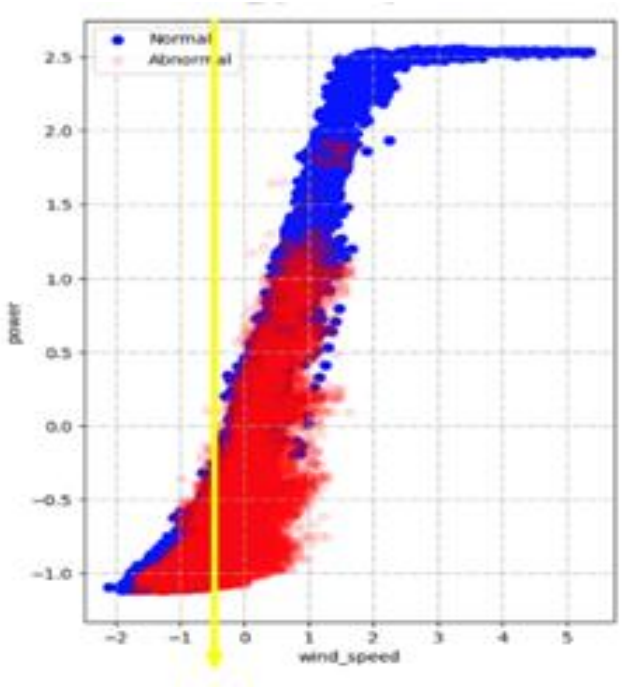
After

优化效果与总结

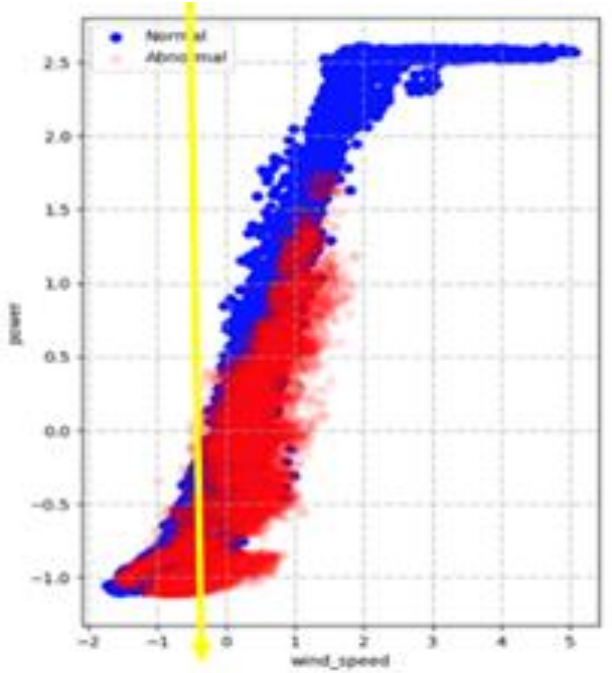
通过数据分割，我们的模型在15号风机wind_speed<-0.5时取得了85.6%的准确率，wind_speed>-0.5时取得了83.1%的准确率；在21号测试集wind_speed<-0.5时取得了79.3%的准确率，wind_speed>-0.5时取得了73.8%的准确率。

模型的预测能力有了明显的下降，我们猜测是因为不同风机的数据强分割界线可能有所不同，我们要选取不同的分割标准。

另外，我们接下来还可以对其他特征考虑进行数据分割，并进一步结合强规则过滤、特征选择来综合提升模型的能力。



15号风机wind_speed与power关系



21号风机wind_speed与power关系

工程管理学院

SCHOOL OF MANAGEMENT & ENGINEERING

非常感谢您的观看