

特征工程与特征选择

汇报人：MG1915029-钱创杰
小组成员：王一伦，杨超凡，温子祺，钱创杰，陈波

一、**数据描述：**数据采集与监视控制系统的数据通常有上百个监控变量，本次项目的数据经过筛选保留了其中 28 个连续数值型变量，涵盖了风机的工况参数、环境参数和状态参数等多个维度。变量的名称及说明如下表所示：

字段名	说明
time	时间戳
wind_speed	风速
generator_speed	发电机转速
power	网侧有功功率(kw)
wind_direction	对风角(°)
wind_direction_mean	25 秒平均风向角
yaw_position	偏航位置
yaw_speed	偏航速度
pitch1_angle	叶片 1 角度
pitch2_angle	叶片 2 角度
pitch3_angle	叶片 3 角度
pitch1_speed	叶片 1 速度
pitch2_speed	叶片 2 速度
pitch3_speed	叶片 3 速度
pitch1_moto_tmp	变桨电机 1 温度
pitch2_moto_tmp	变桨电机 2 温度
pitch3_moto_tmp	变桨电机 3 温度
acc_x	x 方向加速度
acc_y	y 方向加速度
environment_tmp	环境温度
int_tmp	机舱温度
pitch1_ng5_tmp	ng5 1 温度
pitch2_ng5_tmp	ng5 2 温度
pitch3_ng5_tmp	ng5 3 温度
pitch1_ng5_DC	ng5 1 充电器直流电流
pitch2_ng5_DC	ng5 2 充电器直流电流
pitch3_ng5_DC	ng5 3 充电器直流电流
group	数据分组标识

表 1：SCADA 数据变量名称及表述

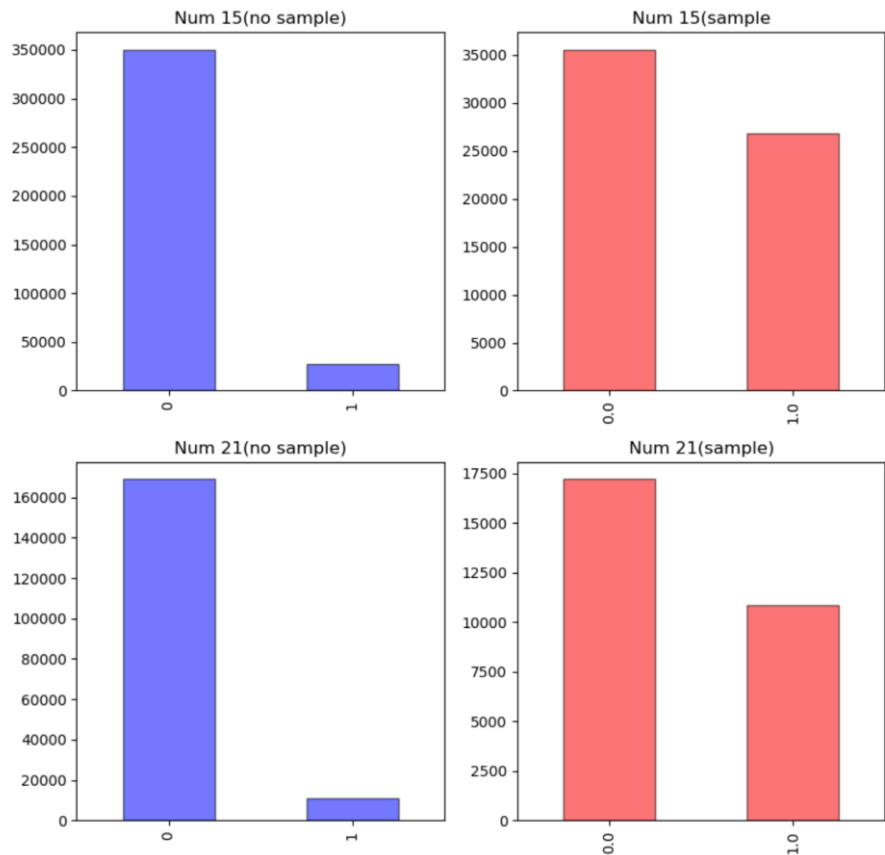
我们通过时间匹配，给数据增加了 Label 标签，Label=0 说明此状态风机未结冰，Label=1 说明此状态风机结冰。此项目旨在通过 28 个特征或者新构造的特征对风机的结冰状态进行预测。我们进行特征工程的目的就是从 28 个特征中寻找最优特征子集。特征选择能剔除不相关或冗余的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的。另一

方面，选取出真正相关的特征简化模型，协助理解数据产生的过程。

二、数据预处理：

2.1 缺失值：经过 python 的统计，我们 15 和 21 号风机数据均不存在缺失值；

2.2 样本平衡：原始数据的正常风机与结冰风机的样本数量比在 12:1 左右，样本量失衡，对正常风机进行欠采样。统计检测装置的连续检测时间大概为 7 秒，考虑到正常风机在 1min 内的状态变化较小，所以对每一组的风机进行移动平均，步长为 10。最终得到的 15 号风机和 21 号风机的数据样本比例控制在 1.4:1，提高处理速度，也便于模型更好拟合数据特征。

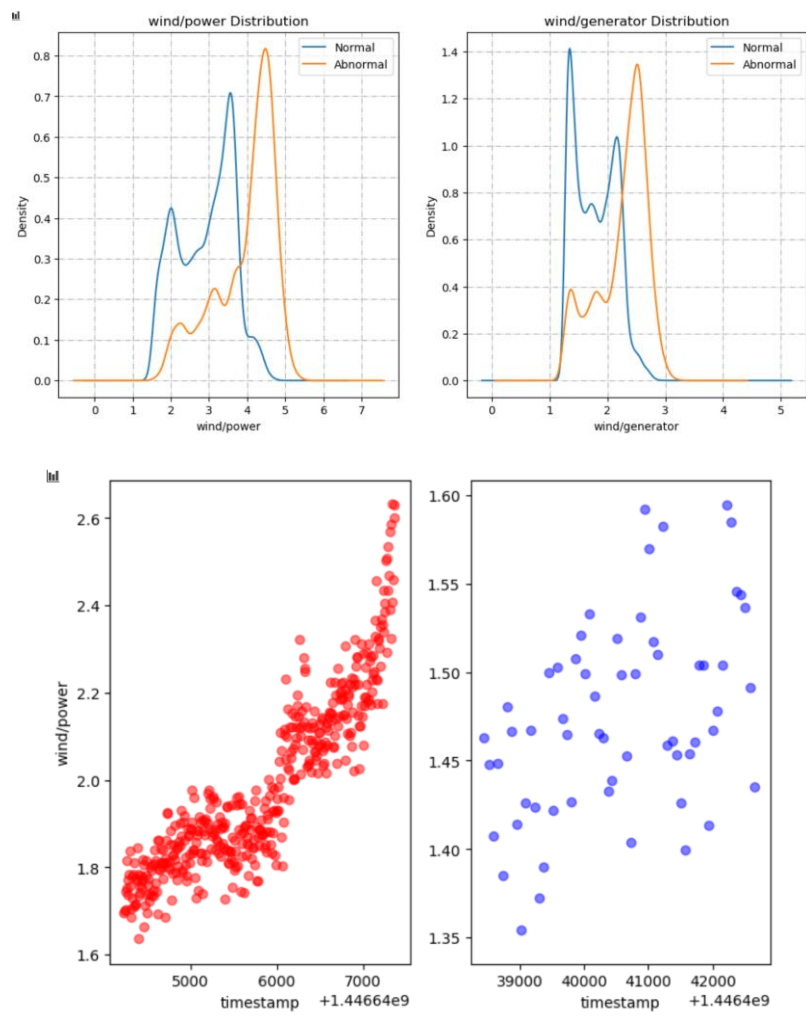


三、物理基础：

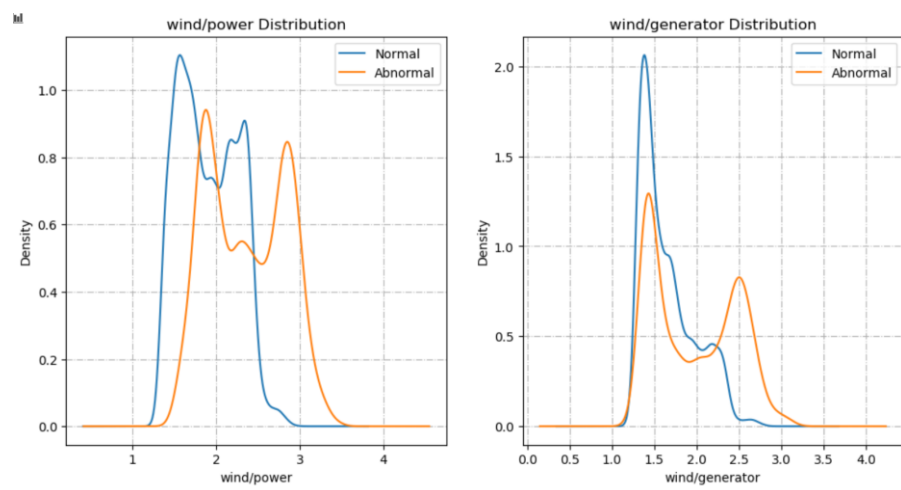
叶片结冰可以看做一个缓慢的能量累积和转化过程，结冰的程度、影响度与环境条件（温度、湿度、风速等）、以及风机参数（叶型、高度、额定功率等）有关。在给定温度条件下，空气含水量、空气密度、水滴形状均不变的情况下，单位时间内叶片结冰质量取决于风速的大小，风速越大，结冰质量越大。

经过数据预处理后将选择和构建风机叶片结冰预测的敏感表征特征。首先，分析各原始特征之间的相互关系。除了风速与功率之外，环境温度与机舱温度也对风机叶片结冰较为敏感。尝试分析其原因，可能是因为叶片结冰影响了风机风速与功率之间的关系，导致机舱内主传动链发热增加，使机舱温度处在较高水平，机舱内外温度差较低。

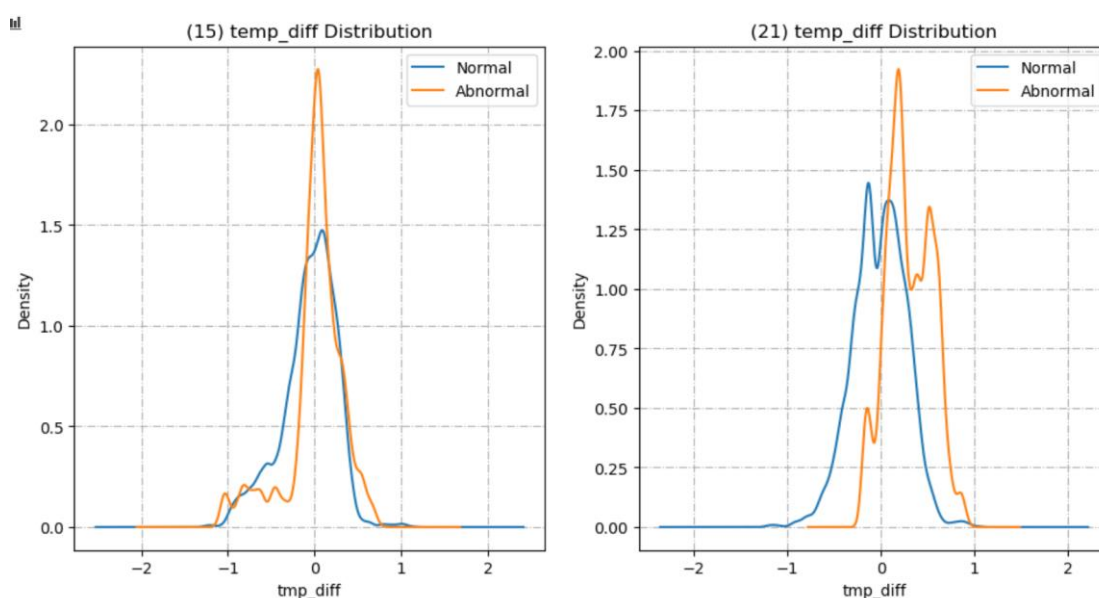
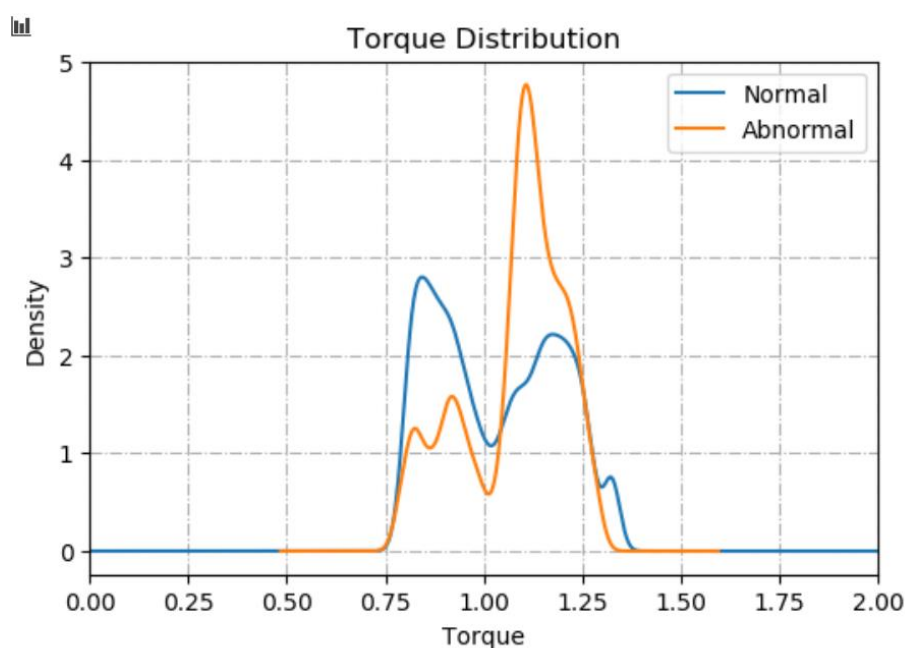
添加新特征： `wind/power`, `wind/generator_speed`, `power/generator_speed`, `tmp_diff`



15 号风机



21 号风机



通过上述的数据探索和机理研究，构建出大量有利于算法识别的特征。这些特征经过算法多轮次的数据可视化分析、建模后的特征重要性评估、特征选择，有一些新的特征保留下来。根据实际问题，提取了风速与网侧有功功率、发动机转速之间的比值等特征。特征能够作为建立预测模型的输入参数，能够最大程度地为分类算法提供相互独立的信息，在保证预测精度的前提下，降低分类模型的复杂度，提高模型的泛化性能。

四、特征选择

1. 特征选择的搜索策略：完全搜索策略、启发式策略以及随机搜索策略。
2. 大规模数据集-过滤式方法：先按照某种规则对数据集进行特征选择，然后再训练学习器，特征选择过程与后续学习器无关，这相当于先用特征选择过程对初始特征进行“过滤”，再用过滤后的特征来训练模型。
3. 特征的发散性：如果一个特征不发散，例如方差接近于 0，也就是说样本在该特征上基

本没有差异，那么这个特征对于样本的区分并没有什么用。

- 特征与目标的相关性：特征与目标的相关性越高说明特征的变动对目标的影响较大，因此我们应当优先选择与目标相关性高的特征。



五、过滤式特征选择

1. 方差过滤

计算各个特征的方差，然后根据阈值选择方差大于阈值的特征，或者指定待选择的特征数 k ，然后选择 k 个最大方差的特征。我们旨在通过方差过滤剔除例如那些在所有样本中都一样的特征。原数据共包含 28 个变量。首先计算 15 号风机所有变量的方差：

wind_speed	6.151976e-01
generator_speed	8.800969e-01
power	6.670247e-01
wind_direction	1.120073e+00
wind_direction_mean	1.049015e+00
yaw_position	8.567647e-01
yaw_speed	1.797369e-02
pitch1_angle	3.350457e-01
pitch2_angle	3.313231e-01
pitch3_angle	3.336161e-01
pitch1_speed	1.172634e-02
pitch2_speed	1.161098e-02
pitch3_speed	1.000912e-02
pitch1_moto_tmp	4.858727e-01
pitch2_moto_tmp	4.512023e-01
pitch3_moto_tmp	4.386276e-01
acc_x	3.522074e-01
acc_y	2.478685e-01
environment_tmp	1.009938e+00
int_tmp	1.088290e+00
pitch1_ng5_tmp	5.439174e-02
pitch2_ng5_tmp	5.370099e-02
pitch3_ng5_tmp	4.600535e-02
pitch1_ng5_DC	5.362647e-01
pitch2_ng5_DC	4.682992e-01
pitch3_ng5_DC	6.364880e-01
group	1.049464e+06
timestamp	2.033347e+12

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，保留方差大于阈值的变量（21 个），结果如下：

```
['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']
```

方差过滤后得到如上图的 21 个变量，将该 21 个变量作为预测变量放入到决策树模型中。对方差过滤后的 15 号风机数据采用交叉验证进行预测。

```
clf=DecisionTreeClassifier(random_state=20,criterion='entropy')
score_15=cross_val_score(clf,data,ytrain,cv=10)
score_15.mean()
```

0.8118828845207006

```
score_15.min()
```

0.7049679487179488

```
score_15.max()
```

0.9770833333333333

预测结果均值、最小值、最大值分别为 81.1%、70.5%、97.7%

2.卡方过滤

卡方检验适用于分类问题。卡方值越大，表明特征与预测结果的相关性也越大，同时 p 值也相应较小，因此我们优先选择 p 值小的特征。在方差过滤的基础上，计算每个非负变量和标签间的卡方统计量，按照卡方统计量由高到低排名。

由于需要保证变量非负，首先对方差过滤后的 15 号风机数据进行归一化。

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.015484	0.016712	0.455540	0.955513	0.607946	0.418899	0.458891	0.220380	0.349586	0.480839	0.636071	0.448052	0.648456	0.015576	0.050005
1	0.015486	0.016714	0.442668	0.957407	0.588878	0.434364	0.544702	0.220380	0.349586	0.480839	0.636071	0.883117	0.783848	0.015576	0.050006
2	0.015489	0.016717	0.444153	0.959300	0.589969	0.393436	0.417402	0.221576	0.349586	0.480839	0.636071	0.512987	0.912114	0.015576	0.050008
3	0.015492	0.016720	0.460491	0.957407	0.585066	0.403244	0.668736	0.220380	0.349586	0.477662	0.636071	0.506494	0.484561	0.015576	0.050009
4	0.015494	0.016722	0.495148	0.959300	0.584518	0.397774	0.534030	0.220380	0.349586	0.480839	0.636071	0.519481	0.954869	0.015576	0.050011
...
62386	0.094284	0.994946	0.422468	0.959678	0.588279	0.634534	0.666416	0.230181	0.173413	0.212955	0.279169	0.575325	0.619240	0.995067	0.997367
62387	0.094287	0.994972	0.445440	0.960151	0.590676	0.634308	0.669062	0.229702	0.172953	0.213562	0.279566	0.769481	0.517340	0.995067	0.997380
62388	0.094290	0.994997	0.410140	0.960341	0.567633	0.643927	0.668757	0.229583	0.172493	0.214515	0.278772	0.553247	0.526603	0.995067	0.997394
62389	0.094292	0.995023	0.415685	0.959773	0.572100	0.641098	0.668246	0.229343	0.172493	0.213562	0.278772	0.611688	0.626366	0.995067	0.997407
62390	0.094295	0.995040	0.419398	0.955040	0.561641	0.623067	0.682293	0.228446	0.172493	0.214356	0.278772	0.727273	0.655582	0.995067	0.997416

62391 rows × 15 columns

之后用 卡方过滤对归一化后数据进行特征筛选：得到各个变量的 P 值

```
pvalues_chi
```

```
array([3.55761821e-002, 5.04428153e-226, 0.00000000e+000, 1.56335854e-001,
        4.27052933e-001, 3.88636780e-003, 8.88673720e-036, 4.59286040e-035,
        1.48846529e-036, 1.62985617e-012, 1.80853955e-009, 2.60412051e-012,
        1.58758262e-005, 6.99615382e-001, 2.16386510e-145, 1.18724288e-117,
        2.58447462e-002, 3.05141501e-001, 7.30624308e-002, 4.59346038e-010,
        7.04575851e-007])
```

以 0.05 为显著性水平，认为大于显著性水平的变量为相关性低的变量，对其进行剔除。

```
k=chivalue.shape[0]-(pvalues_chi>0.05).sum()
k
```

16

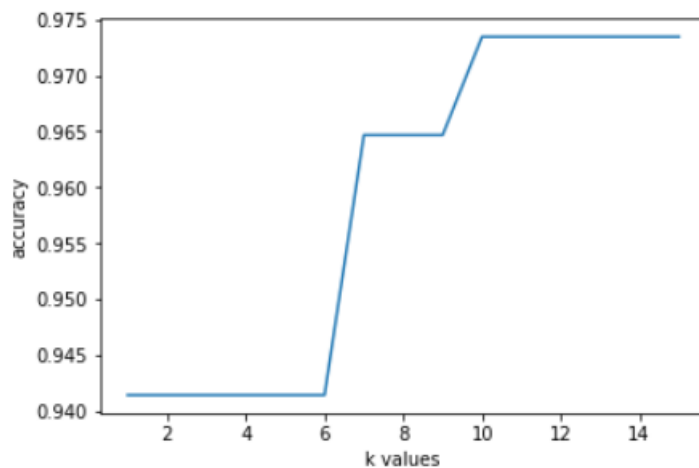
从而将变量个数降低到 16 个。

```
: score_15_k=cross_val_score(clf,data15_k,ytrain,cv=10)
score_15_k.mean()
```

```
: 0.8104871406939669
```

结果为 81.0%，说明用筛选出的 16 个变量进行预测，得到的结果和方差过滤中略有下降。如下图所示，原因在于，当变量取 K 方最高的前 10 个时，预测效果已经达到饱和无法增加，因此选择 15 个变量或 12 个变量的结果都相同。

```
for i in range(1,16):
    data15_k=SelectKBest(chi2,k=i).fit_transform(result_,ytrain)
    once=cross_val_score(clf,data15_k,ytrain,cv=10).mean()
    score.append(once)
plt.plot(range(1,16),score)
plt.xlabel('k values')
plt.ylabel('accuracy')
plt.show()
```



因此，根据 P 值，选取 P 值较小的 8 个变量即可。

```
[('Unnamed: 0', 0.0),
 ('power', 0.0),
 ('generator_speed', 5.0442815338183355e-226),
 ('environment_tmp', 2.1638650987649144e-145),
 ('int_tmp', 1.187242878224715e-117),
 ('pitch1_moto_tmp', 1.629856170055471e-12),
 ('Unnamed: 0.1', 4.104469567826702e-10),
 ('group', 4.593460380344482e-10),
 ('timestamp', 7.045758508298403e-07),
 ('yaw_position', 0.0038863677950941055),
 ('pitch1_ng5_DC', 0.02584474620830755),
 ('wind_speed', 0.03557618206007333),
 ('pitch3_ng5_DC', 0.07306243078366069),
 ('wind_direction', 0.15633585435743683),
 ('wind_direction_mean', 0.4270529329363869)]
```

选择的特征分别为：

功率，发动机转速，环境温度，机舱温度，变桨电机 1 温度，组别，timestamp,偏航位置。

六、模型嵌入特征选择

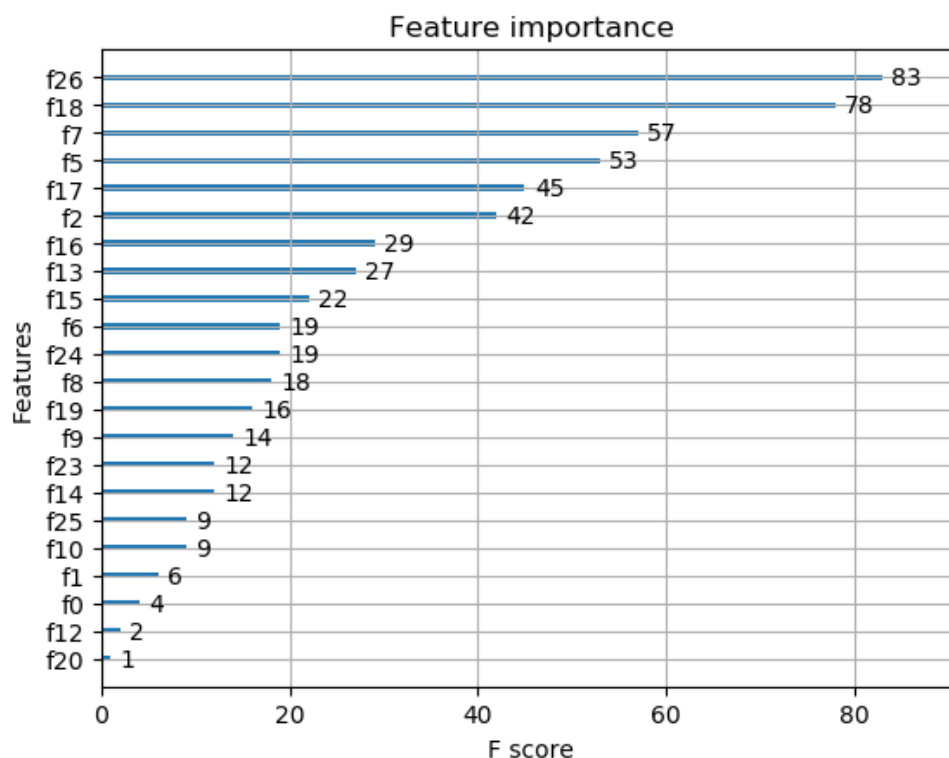
嵌入法，即基于模型的特征排序。先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于过滤式方法，但是是通过训练来确定特征的优劣。

1.用 XGboost 模型对特征重要性进行排序

使用梯度提升算法的好处是在提升树被创建后，可以相对直接地得到每个属性的重要性得分。一般来说，重要性分数，衡量了特征在模型中的提升决策树构建中价值。一个属性越多的被用来在模型中构建决策树，它的重要性就相对越高。

属性重要性是通过对数据集中的每个属性进行计算，并进行排序得到。在单个决策书中通过每个属性分裂点改进性能度量的量来计算属性重要性，由节点负责加权和记录次数。也就说一个属性对分裂点改进性能度量越大（越靠近根节点），权值越大；被越多提升树所选择，属性越重要。最终将一个属性在所有提升树中的结果进行加权求和后然后平均，得到重要性得分。

通过 `plot_importance()`函数进行特征重要性得分排序，结果如下图：



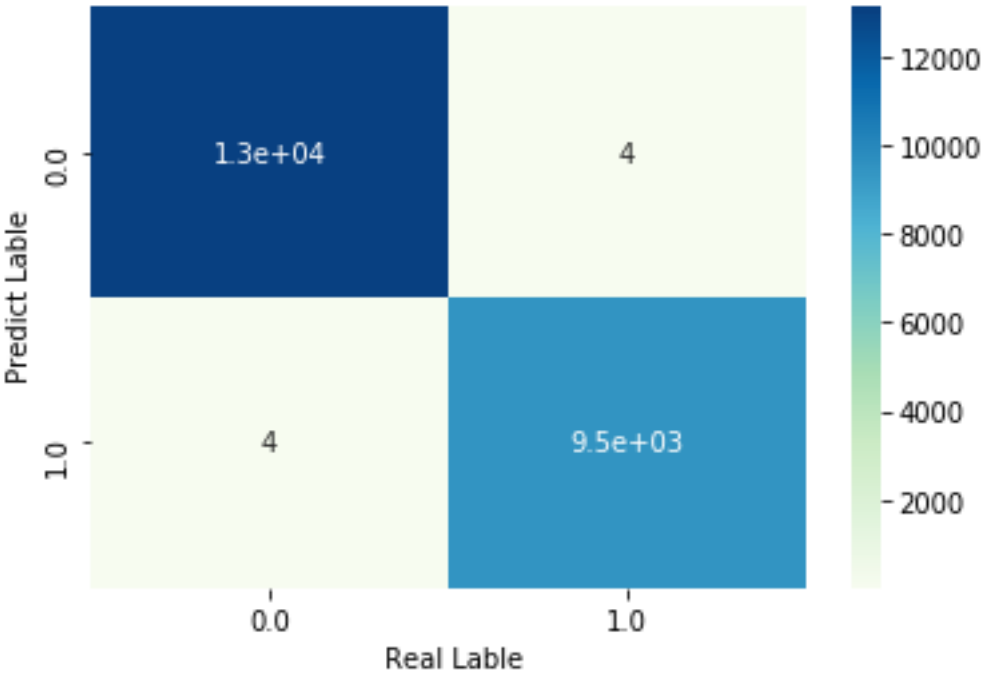
根据其在输入数组中的索引，特征被自动命名为 `f0-f27`。在问题描述中手动的将这些索引映射到名称，我们可以看到，`f26`（功率）具有最高的重要性。

2.KNN 算法，通过尝试，发现当选取如下特征进行预测分类最准。

```
'wind_speed','generator_speed','power','wind_direction','wind_direction_mean',  
'pitch1_moto_tmp','pitch2_moto_tmp','pitch3_moto_tmp',  
'acc_x','acc_y','environment_tmp','int_tmp','pitch1_ng5_tmp',  
'pitch2_ng5_tmp','pitch3_ng5_tmp','pitch1_ng5_DC','pitch2_ng5_DC',  
'pitch3_ng5_DC','group','pitch_speed_mean','Label'
```

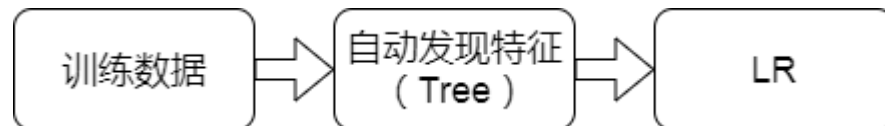
当k 值取 7 时，将 15 号风机和 21 号风机数据合并后，分为训练集和测试集，进行测试的结果如下：

```
# 构建模型，并将最佳的近邻个数设置为 7
knn_class = neighbors.KNeighborsClassifier(n_neighbors = 7, weights = 'distance')
# 模型拟合
knn_class.fit(X_train, y_train)
# 模型在测试数据集上的预测
predict = knn_class.predict(X_test)
# 构建混淆矩阵
cm = pd.crosstab(predict,y_test)
cm
```



	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	13115
1.0	1.00	1.00	1.00	9496
accuracy			1.00	22611
macro avg	1.00	1.00	1.00	22611
weighted avg	1.00	1.00	1.00	22611

3.GBDT/RF + LR 算法



GBDT 由多棵 CART 回归树组成，将累加所有树的结果作为最终结果。GBDT 拟合的目标是一个梯度值（连续值，实数），可以用来做回归预测，调整后也可以用于分类。

学习方式：通过迭代多棵树来共同决策，Gradient Boosting，每一次建立模型是在之前模型损失函数的梯度下降方向。

之前特征选取的依据是在特征工程中我们根据物理意义人为构造出了如力矩、扭矩等可解释的新特征，把这些特征直接用于模型训练的缺点在于忽视了高维度的特征关系，如 time 和 wind_speed 两个特征交互出的新特征（在某个特定时间的风速）是否能对结冰做出更好的解释？基于这样的想法，可以考虑把特征训练的工作交给模型自身，这里分别采用了随机森林和 GBDT 两种决策树学习的方式来构造新特征（缺点是无监督学习出来的新特征的解释性差），然后把新特征交给 LR 模型训练，最后在 21 号风机上测试取得了比较好的效果。在传递训练特征前的数据可视化探索中，考虑到 group 的存在可能导致模型过拟合这里没有选入训练，而通过柱状图发现结冰数据与每天固定的某个时间段有相关性，因此 time 经过切割只保留了小时数。

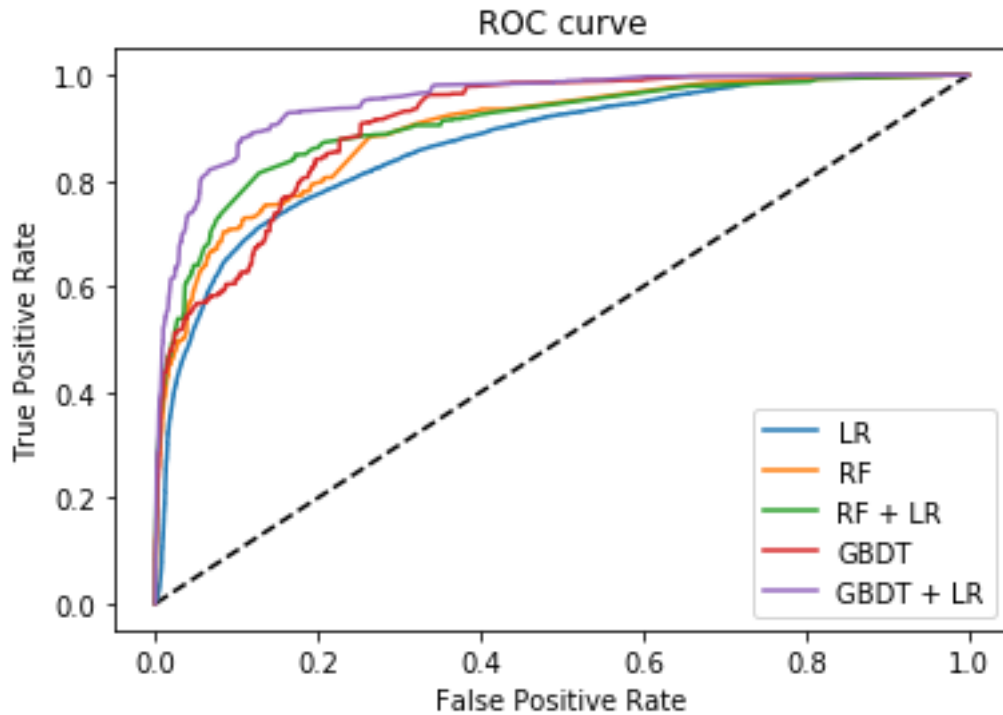
用于训练的特征：

```
train_features=['time', 'wind_speed', 'generator_speed', 'power', 'wind_direction',
               'wind_direction_mean', 'pitch1_moto_tmp', 'pitch2_moto_tmp',
               'pitch3_moto_tmp', 'acc_x', 'acc_y', 'environment_tmp', 'int_tmp',
               'pitch1_ng5_tmp', 'pitch2_ng5_tmp', 'pitch3_ng5_tmp', 'pitch1_ng5_DC',
               'pitch2_ng5_DC', 'pitch3_ng5_DC', 'pitch_speed_ave']
```

对风机 15 数据集一半作为训练集，一半作为验证集，同时为了防止学习过程标签泄露，前 1.7 万个数据只用于 GBDT 或 RF 学习特征划分，后 1.7 万个数据通过 LR 对学习到的特征重要程度排序：

```
# 将样本集分成训练集和验证集
X_train, X_test, y_train, y_test = train_test_split(data15[train_features], data15['Label'], test_size=0.5)
# 再将训练集拆成两个部分 (GBDT/RF, LR)
X_train, X_train_lr, y_train, y_train_lr = train_test_split(X_train, y_train, test_size=0.5)
```

对比 RF, RF+LR, GBDT, GBDT+LR 预测结果的 ROC 曲线：



AUC 得分:

```
LogisticRegression: 0.8681961417833913
RandomForest: 0.8952829508908803
RandomForest+LogisticRegression: 0.9069724825688029
GBDT: 0.9074537145483896
GBDT+LogisticRegression: 0.951154576326235
```

4.合并特征，分组实验

(1)在上周初步选择特征的基础上，对如下的特征进行进一步的特征选择分析。

```
['power',
 'generator_speed',
 'pitch_angle_avg',
 'pitch_tmp_avg',
 'wind/power',
 'wind/generator',
 'torque',
 'tmp_diff',
 'hour']
```

(2)首先在 15 号风机，21 号风机，以及两个风机合并的三个数据集中进行方差的检验，发现 `pitch_tmp_avg` 的方差小于 0.1。

(3)而后进行上述特征的 f 检验，得到结果如下，每个特征显著不为零。

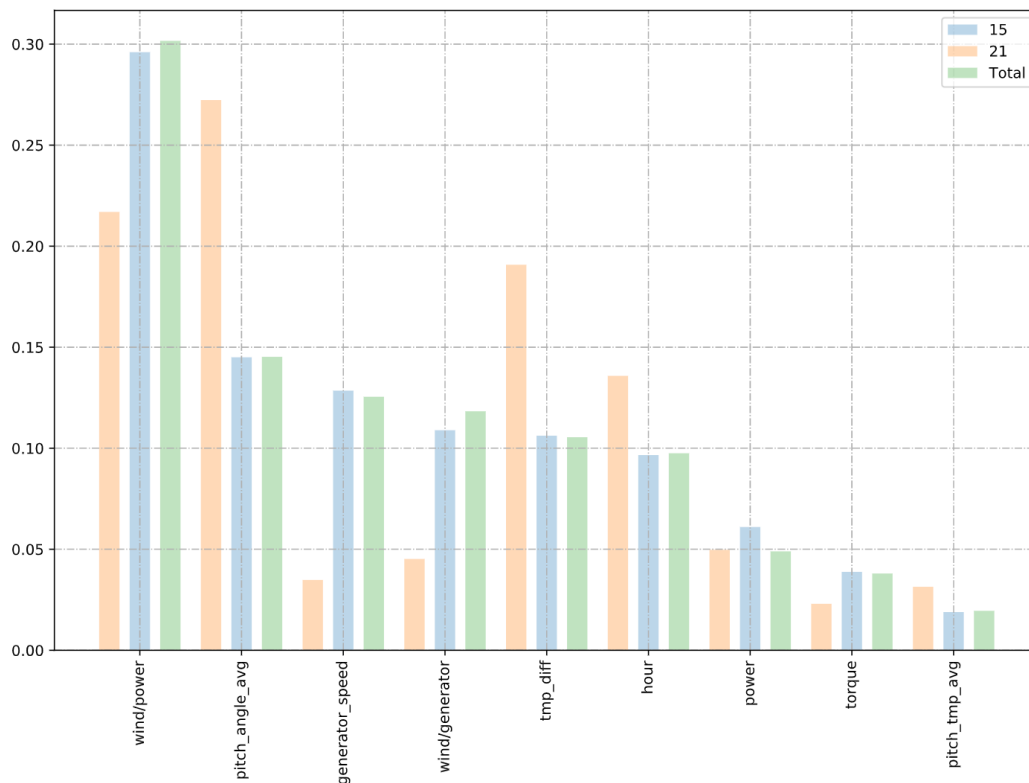
	15	15score	15p	21	21score	21p	total	totalscore	totalp
0	power	8351.926042	0.000000e+00	power	4050.594343	0.000000e+00	power	8351.926042	0.000000e+00
1	generator_speed	7886.358900	0.000000e+00	generator_speed	2652.265485	0.000000e+00	generator_speed	7886.358900	0.000000e+00
2	pitch_angle_avg	1006.364755	4.149922e-219	pitch_angle_avg	909.883724	9.703218e-197	pitch_angle_avg	1006.364755	4.149922e-219
3	pitch_tmp_avg	366.515333	1.841021e-81	pitch_tmp_avg	55.065073	1.199111e-13	pitch_tmp_avg	366.515333	1.841021e-81
4	wind/power	30857.556351	0.000000e+00	wind/power	8989.487597	0.000000e+00	wind/power	30857.556351	0.000000e+00
5	wind/generator	13650.796048	0.000000e+00	wind/generator	3434.586596	0.000000e+00	wind/generator	13650.796048	0.000000e+00
6	torque	7547.948011	0.000000e+00	torque	2177.251446	0.000000e+00	torque	7547.948011	0.000000e+00
7	tmp_diff	667.476889	2.091538e-146	tmp_diff	9227.559626	0.000000e+00	tmp_diff	667.476889	2.091538e-146
8	hour	44.124074	3.107286e-11	hour	22.617658	1.986410e-06	hour	44.124074	3.107286e-11

(4)然后进行互信息检验。互信息，即从信息熵的角度分析各个特征和输出值之间的关系评分。在决策树算法中我们讲到过互信息（信息增益）。互信息值越大，说明该特征和输出值之间的相关性越大，越需要保留。互信息计算的不是数据序列，而是数据的分布，因此互信息可以用于检测特征间的非线性关系。

可以看出 torque（power/generator_speed）和 wind/power 两个特征的得分较少，因为这两个特征与 power 呈现线性关系。总体上，选择的特征与 Label 的关系较强。

	15	15score	21	21score	total	totalscore
0	power	0.578307	power	0.454942	power	0.577217
1	generator_speed	0.568374	generator_speed	0.516951	generator_speed	0.569121
2	pitch_tmp_avg	0.462321	pitch_angle_avg	0.461417	pitch_tmp_avg	0.463703
3	wind/power	0.267881	pitch_tmp_avg	0.342679	wind/power	0.267852
4	torque	0.335295	wind/power	0.201027	torque	0.335028
5	tmp_diff	0.507030	tmp_diff	0.489737	tmp_diff	0.507652

(5)使用随机森林对三个数据集的特征进行重要度评估可以看出三个数据集的 pitch_tmp_avg 的重要度都偏低，15 号风机的 power, torque 偏低，21 号风机的 generator_speed, wind/generator, torque, pitch_tmp_avg 偏低。



(6) 使用随机森林对三个数据集进行扰动检验（左图），发现 `pitch_tmp_avg`, `torque`, `power` 受扰动时在两个数据集中对准确度的影响较少，所以考虑将这三个特征进行剔除。结合单变量的随机森林训练分析（右图），得出 15 号风机的 `power` 特征，`pitch_tmp_avg` 特征对区分目标的影响较低，21 号风机的 `torque`, `pitch_tmp_avg` 特征对区分目标的影响较低。

15power	0.0034		
15generator_speed	0.0385		
15pitch_angle_avg	0.0984		
15pitch_tmp_avg	0.0029		
15wind/power	0.2727		
15wind/generator	0.0015		
15torque	0.0123		
15tmp_diff	0.0794		
15hour	0.0937		
21power	0.0130		
21generator_speed	0.0008		
21pitch_angle_avg	0.1907		
21pitch_tmp_avg	0.0074		
21wind/power	0.0940		
21wind/generator	0.0003		
21torque	0.0003		
21tmp_diff	0.1305		
21hour	0.1090		
totalpower	0.0034		
totalgenerator_speed	0.0449		
totalpitch_angle_avg	0.0978		
totalpitch_tmp_avg	0.0034		
totalwind/power	0.2585		
totalwind/generator	0.0006		
totaltorque	0.0146		
totaltmp_diff	0.0786		
totalhour	0.0914		
dtype:	float64		

	Accu
15power	0.669521
15generator_speed	0.840363
15pitch_angle_avg	0.497940
15pitch_tmp_avg	0.571573
15wind/power	0.829512
15wind/generator	0.758412
15torque	0.663222
15tmp_diff	0.543461
15hour	0.479750
21power	0.643958
21generator_speed	0.778953
21pitch_angle_avg	0.603851
21pitch_tmp_avg	0.590382
21wind/power	0.752683
21wind/generator	0.681212
21torque	0.646777
21tmp_diff	0.705163
21hour	0.630872

(7) 对风机的原始特征再进行重要度分析，发现 `acc_x`, `acc_y` 和 `environment_tmp`，的重要度较高，考虑再加入这三个特征。

(8) 对于 15 号风机选择如下的特征进行随机森林的交叉差验证，得到的准确度优于可视化分析得到的初始特征：

```
'wind/power','hour','pitch_angle_avg','tmp_diff','generator_speed','wind/generator','torque',  
'acc_x','acc_y'
```

```
before:  
[0.62008174 0.86544318 0.7932361 0.84252284 0.83603142]  
after:  
[0.70991265 0.94189774 0.90743709 0.96698189 0.92050008]
```

(9) 对于 21 号风机选择如下的特征进行随机森林的交叉验证，得到的准确度优于可视化分析得到的初始特征：

```
'wind/power','hour','pitch_angle_avg','tmp_diff','power','wind/generator','generator_speed',  
'acc_x','acc_y','environment_tmp'
```

```
before:  
[0.66868651 0.65425058 0.79950098 0.71604278 0.78484848]  
after:  
[0.75654963 0.74906434 0.98003921 0.89982175 0.91871658]
```

七、下一步方向：

1. 分析不同风机的结构性差异，提升特征选取后模型的泛化能力。
2. 针对不同模型的特点，匹配最合适的特征进行计算。
3. 风机结冰故障会造成哪些现象、以及由哪些原因导致这种结冰故障，均需要深厚的风机行业知识才能了解。根据基准模型，结合物理意义进一步寻找高级特征进行优化。