**强规则过滤**

1. power<2

2. wind\_speed<1.9

3. environment\_tmp<2

4. -2.5 <int\_tmp<1.5

然后在数据集中引入period即时段属性作为特征训练模型，最后AUC得分效果比较好的有随机森林和随机森林+逻辑回归。

保留特征：time只保留小时数，period=1表示时间段在9-14时内，period=0表示时间段不在9-14时内，其余特征由保留的原始特征和构建物理特征组成。

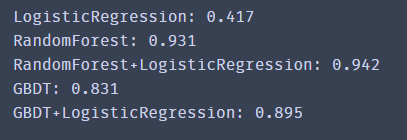
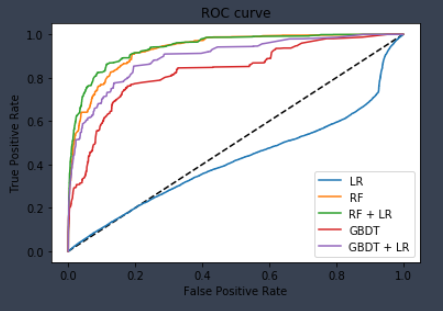
[ 'time','wind\_speed', 'generator\_speed', 'power', 'wind\_direction',

'wind\_direction\_mean', 'yaw\_position', 'pitch1\_moto\_tmp',

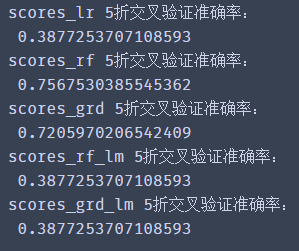
'pitch2\_moto\_tmp', 'pitch3\_moto\_tmp', 'acc\_x', 'acc\_y',

'environment\_tmp', 'int\_tmp', 'pitch1\_ng5\_DC', 'pitch2\_ng5\_DC', 'pitch3\_ng5\_DC', 'temp\_diff', 'torque', 'cp', 'ct', 'pitch\_angle\_ave',

'pitch\_angle\_std', 'spd\_rate', 'pitch\_speed\_ave','period']



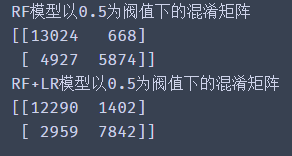
但从训练集15号风机上的五折交叉验证结果来看，使用两种模型结合（随机森林+LR或GBDT+LR）的交叉验证得分明显比单模型训练的得分要低，随机森林的稳定性最好（0.757），因此需要借助两种模型的混淆矩阵，从业务场景的角度辅助判断模型优劣。



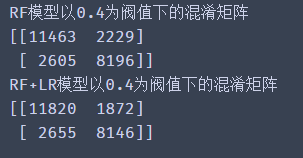
**模型调优**

由于分类模型的输出结果是预测指标可能为1的概率，但混淆矩阵只允许输入的数据全为0-1变量，这里需要人为设置阀值，把大于阀值的概率返回1，小于阀值的概率返回0。分别尝试了以0.4，0.5和0.6作为阀值得到的混淆矩阵结果。

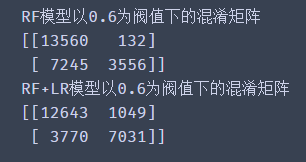
0.5作为阀值



0.4作为阀值



0.6作为阀值



RF

|  |  |  |
| --- | --- | --- |
| 阀值 | FN | FP |
| 0.4 | 0.214 | 0.185 |
| 0.5 | 0.102 | 0.274 |
| 0.6 | 0.036 | 0.348 |

RF+GBDT

|  |  |  |
| --- | --- | --- |
| 阀值 | FN | FP |
| 0.4 | 0.187 | 0.183 |
| 0.5 | 0.152 | 0.194 |
| 0.6 | 0.130 | 0.210 |

实际工业场景中，电力公司认为预测不结冰实际结冰（FN）造成的经济损失远高于预测结冰实际不结冰（FP）带来的损失，基于此可以考虑采用RF模型并将预测概率的判断阀值调至0.6，虽然升高阀值让FN降低的同时导致了FP的提升，但对公司来讲是可以接受的。

以随机森林决策树作为最终模型，同时结合实际业务场景选取0.6作为结冰判断的临界值，以15号风机作为训练集，21号风机作为测试集得到结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | Cross\_val\_score | Accuracy | FN | FP |
| 随机森林 | 0.757 | 0.931 | 0.036 | 0.348 |

模型将3.6%的结冰数据预测为不结冰，34.8%的不结冰数据预测为结冰。