化工产品品质智能预测算法说明书

————by 冬天很冷

本赛题目数据跨时 2 个月，第一个月的数据用于训练，第二个月的数据用于测试。数据可以描述为 3 类：

1. 生产参数记录表。（36 个测点的传感器数据）
2. 产品检测结果。（需要预测的结果）
3. 生产工艺流程。（生产参数和结果的相互关系）

思路（一）：尝试提取生产数据当特征建立回归模型

1，将生产参数记录表中的数据当成训练特征，产品检测结果当成标签数据，建立回归模型。

2，根据生产结果中的时间段，在生产参数记录表中特征在规定时间的平均值、和、检测次数作为特征

3，建立 LR、BayesianRidge、SVR、GradientBoostingRegressor 等模型训练预测。

结论：LR 的效果超出其他模型很多，说明数据具有线性关系。再加上用这种提取生产参数数据的方式建立模型，对于特征提取的要求高，难度较大，所以建议利用时序回归模型。

思路（一）优缺点

优点：

1，抗噪音。因为（1）生产参数特征较多且全面，（2）根据生产参数预测检测结果，类似端到端的训练模型，能捕捉特征反映的结果。

2，查生产过程问题。可以得到每个生产参数的权重 ，如果产品质量出现问题，可以很快很好定位到哪个生产参数记录表出现问题，排查改善。

缺点：

1，生产参数特征不好提取。有缺失不规律，很难提取关键特征。

2，模型准确率低。正因为无法提取较好的特征，所以模型一般很差。

思路（二）：直接利用检测结果建立 ARMA 模型

1，将检测结果按照“product_batch”列为时序，建立时间序列数据

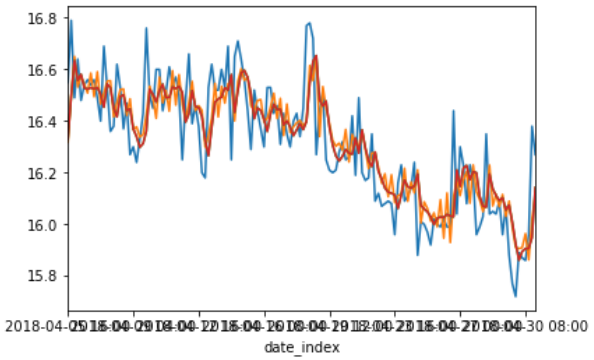
2，对数据进行平稳性检测：基本上是一阶差分或本身平稳。

3，随机性检测，原数据是非白噪声，所以可以直接用原数据建立 ARMA 模型。

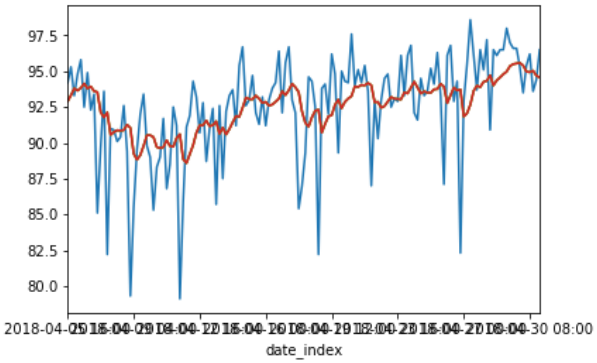
4，确定 ARMA 阶数，这里采用两种方法，一是画出 ACF、PACF 观察 p，q 值，锁定。

5，ARMA 模型参数 max_ar=6,max_ma=4。而是根据锁定的参数训练模型依据 aic、bic、hqic 指标搜索最优 p,q 值。(训练结果见下组图)

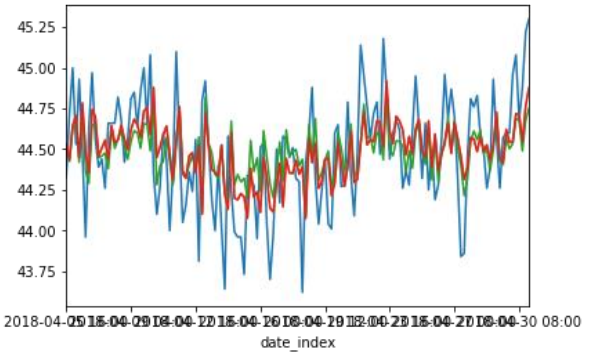
nitrogen_content 建模效果：



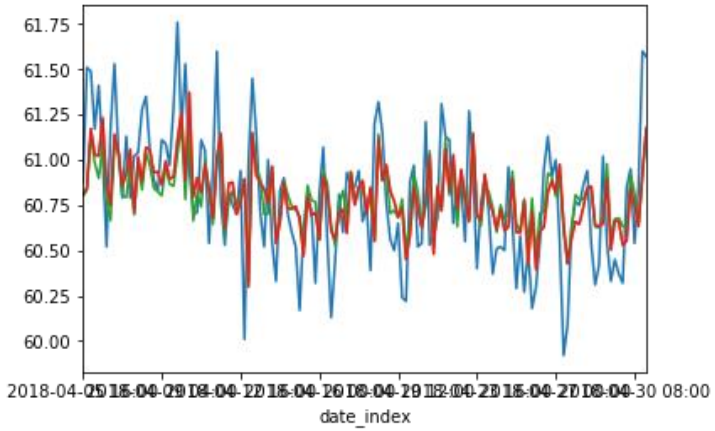
particle_size 建模效果：



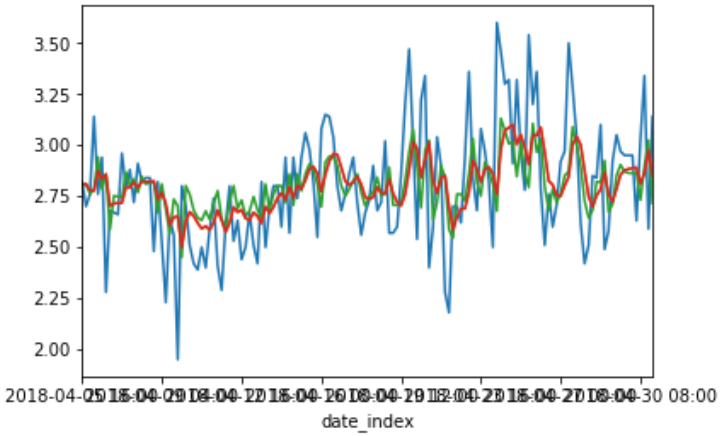
phosphorus_content 建模效果：



total_nutrient 建模效果：



water_content 建模效果：（**蓝色是 4 月真实值，其他颜色是依据 aic bic hqic 最优指标预测 4 月**）



6，根据最优 p,q 值训练模型，线下评价。

7，噪音调优部分：

模型预测后，可能 5 月份数据有噪音的加入，使得模型不够完善，预测 5 月份品质数据不够准确。

这里采用了一个假设：每月品质均值波动变化不大，简单的操作就是 5 月份和 4 月份的均值一样

具体加噪音调优的思路有三个：

（1），在每个时间的数据上，加上 4 月均值和 5 月均值差的正态分布数据数，这样就会让 5 月预测结果加了一个正态分布噪声 。

（2），直接在每个数据上加上均值差，这样相当于在 5 月份结果上加上了值为均值差的偏置。

（3），控制预测值加上均值后在 4 月份最大最小值中间，相当于数据压缩。

线上提交结果显示：设置（arma_order_select_ic）选择最优参数 max_ar=6,max_ma=4，选择出 aic 指标下最优的 p，q 值，训练出的 ARMA，在 6 噪音调优中加**偏置**（2）效果最好，B 榜分数 **0.353305，可复现**。

8，各品牌检测分布调优

在分析 5 月份预测结果分布和 4 月份真是分布发现，“nitrogen_content”围绕均值 15.85 呈现 (0，0.5) 之间随机分布情况，所以用 15.85+ (0，0.5) 随机分布来预测“nitrogen_content”指标，B 榜分数 **0.352423。因为代码没有固定随机数产生的序列，不可复现。**

9，**未完成部分**：

（1） 原始数据的异常值检测，去除后训练模型可能效果更好。

（2） 选择 ARMA 模型的 p，q 值可能可以优化，因为只是认为主管设置 max_ar=6,max_ma=4。

（3） 原始数据 log 变换或其他变换后再训练可能效果更好。

思路（二）：优缺点

优点：

1， 对于短期稳定数据预测准确度高。

2， 建模方便容易。

缺点：

1， 对于长期趋势难预测。

2， 异常值对于模型干扰严重。

3， 难以准确预测产品质量哪天会出问题，风控能力小。

数据提供建议

1， 改善生产参数传感数据，最好能获取周期性稳定数据。

2， 积累长期数据，比如半年，一年。

3， 等模型训练稳定后，为节约成本可简化生产参数传感数据周期性稳定的获取方式。