2025-06-12

问题：

- [ ] SARIMIA模型是不是每次调用api都需要重新训练，或者，以隔天训练一次。每次喂一周的数据
- [ ] 特征工程加在哪个模型合适，都加?

设计：

1.一阶段

- 训练数据（完整数据测试，一周）
- 输入长度为一周
- 输出长度为144（=48 * 3 天预测）
- 重采样频率= 10Min
- 趋势分解（特征）
- 时频转换（特征）
- mse计算=（按工作制度，分为3段，分别计算）
  2.二阶段
- 输入（SARIMIA预测值）
- 输出（真实值）
- 模型选择：LGB（基准）
- 考虑采用标准化

基于mse找最优产参数

每天数据点数: 144 (应为144)
开始参数搜索: 125 × 27 = 3375 种组合

🔥 发现新最优模型: SARIMA(0, 0, 0)x(0, 2, 2, 144)
MSE: 0.0152, MAE: 0.0962, MAPE: 22.03%

一阶段

1.一阶段：

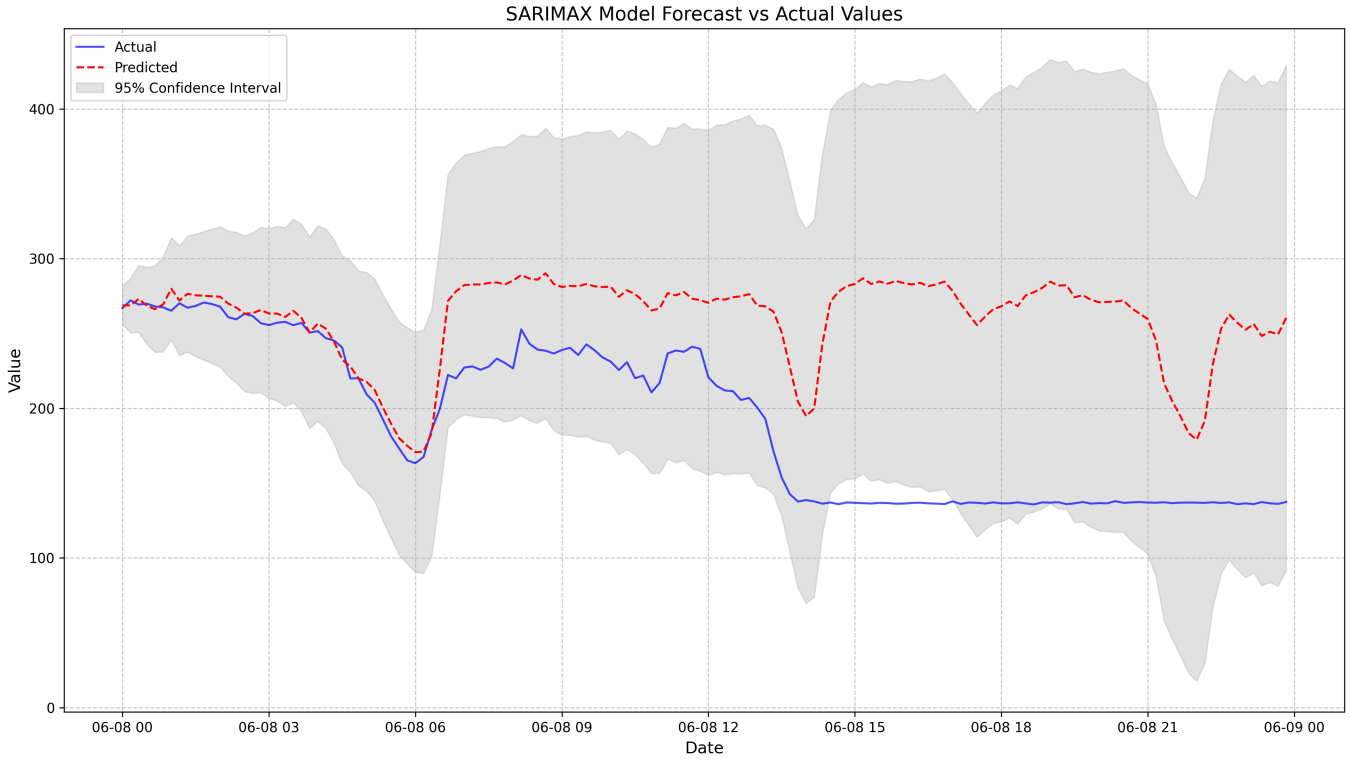
- 训练数据（完整数据测试）
- 输入长度为一周（预测阶段）
- 输出长度为144（= 48 * 3 天预测）
- 重采样频率= 10Min
- mse计算=（按工作制度，分为3段，分别计算）

2.二阶段：

- 输入（SARIMIA预测值）
- 趋势分解（特征）
- 时频转换（特征）
- 输出（真实值）
- 模型选择：LGB（基准）
- 考虑采用标准化

采用7天数据训练模型

成功处理数据! 数据范围: 2025-06-01 00:00:00 到 2025-06-08 23:50:00
总计 1152 个10分钟间隔的数据点
每天有 144 个数据点
训练集大小: 1008 个点 (约 7.0 天)
测试集大小: 144 个点 (1天)
训练用时3分钟



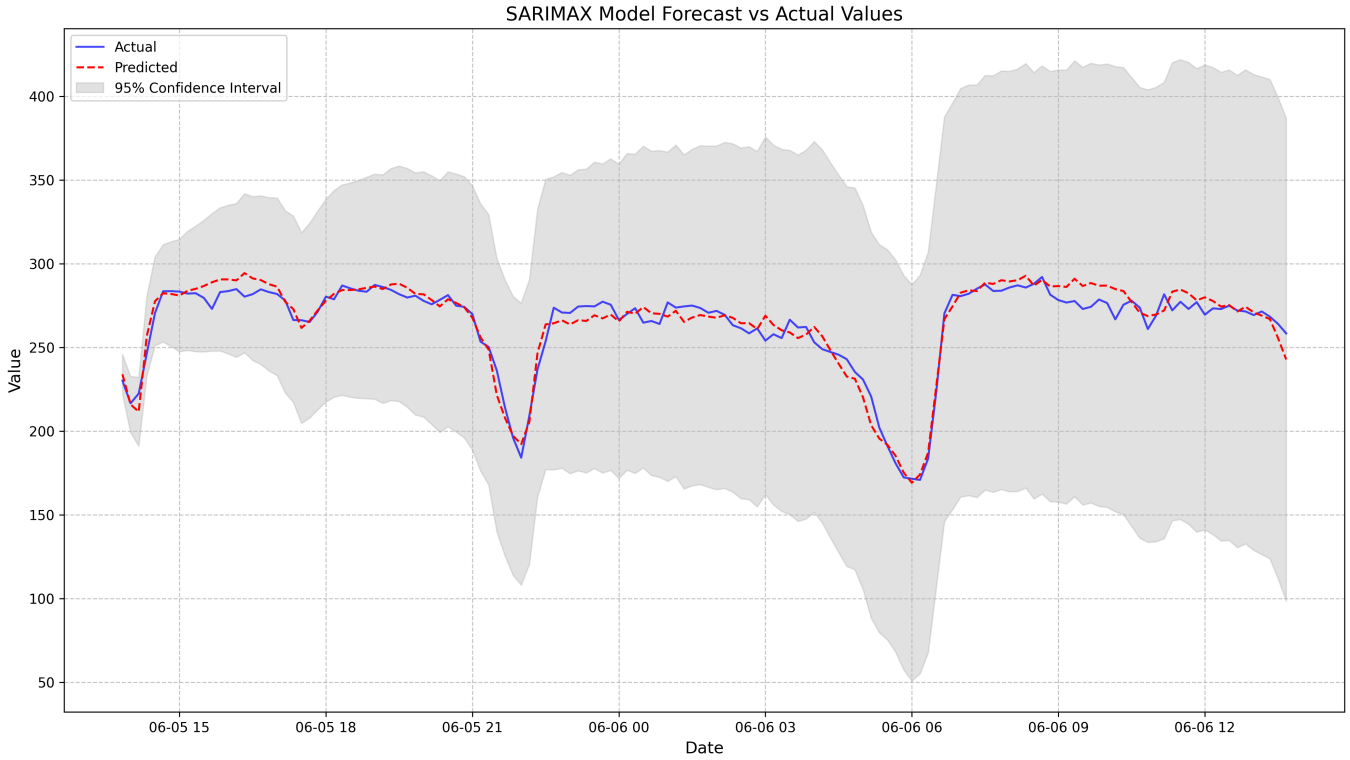
{
"MSE": 7617.262583254203,
"RMSE": 87.27693041837689,

"MAE": 69.41063109342363,
"MAPE": 46.096102710317446,
"Test_size": 144,
"Predictions_size": 144
}
大误差主要是来自6月8日15点之后，后面的实际产气量一直维持在136~137左右

	A	B	C	D	E
1	date	actual_value	prediction	lower_bound	upper_bound
91	2025-06-08 14:50:00	137.074	281.6375934	152.3938006	410.8813862
92	2025-06-08 15:00:00	136.838	283.0248961	152.9847691	413.0650231
93	2025-06-08 15:10:00	136.577	286.8968403	156.0634726	417.7302081
94	2025-06-08 15:20:00	136.348	283.0318256	151.4082546	414.6553966
95	2025-06-08 15:30:00	136.787	284.7523558	152.3415646	417.163147
96	2025-06-08 15:40:00	136.66	283.2084188	150.0133377	416.4034999
97	2025-06-08 15:50:00	136.234	285.035981	151.0594887	419.0124732
98	2025-06-08 16:00:00	136.307	283.7308603	148.9757856	418.485935
99	2025-06-08 16:10:00	136.72	282.715512	147.1846348	418.2463892
100	2025-06-08 16:20:00	136.907	283.8505056	147.5465583	420.1544529
101	2025-06-08 16:30:00	136.458	281.5749086	144.5005775	418.6492396
102	2025-06-08 16:40:00	136.29	282.7311802	144.8891065	420.5732539
103	2025-06-08 16:50:00	135.957	284.6434984	146.0362795	423.2507174
104	2025-06-08 17:00:00	137.795	278.4495239	139.0797141	417.8193336
105	2025-06-08 17:10:00	136.088	269.8492361	129.7193484	409.9791237
106	2025-06-08 17:20:00	137.042	262.6679606	121.7804671	403.555454
107	2025-06-08 17:30:00	136.822	255.6090872	113.9664205	397.2517539
108	2025-06-08 17:40:00	136.327	261.2749077	118.8794615	403.6703538
109	2025-06-08 17:50:00	137.124	266.1728997	123.02703	409.3187693
110	2025-06-08 18:00:00	136.442	268.1641065	124.2701325	412.0580804
111	2025-06-08 18:10:00	136.48	271.4179335	126.7781384	416.0577285
112	2025-06-08 18:20:00	137.146	268.2533744	122.8700063	413.6367426
113	2025-06-08 18:30:00	136.409	275.5284461	129.4037186	421.6531736
114	2025-06-08 18:40:00	135.737	277.5354569	130.6715502	424.3993636
115	2025-06-08 18:50:00	137.084	280.3494811	132.7485427	427.9504195
116	2025-06-08 19:00:00	136.913	284.6200435	136.2841888	432.9558983

采用完整数据训练模型

成功处理数据! 数据范围: 2025-05-01 06:20:00 到 2025-06-06 13:40:00
总计 5229 个10分钟间隔的数据点
每天有 144 个数据点
训练集大小: 5085 个点 (约 35.3 天)
测试集大小: 144 个点 (1天)
程序运行总时间: 1349.34 秒



{
"MSE": 44.70846263159985,
"RMSE": 6.686438710674005,
"MAE": 5.330650276265857,
"MAPE": 2.049238354365466,
"Test_size": 144,
"Predictions_size": 144
}

问题：

1.训练好的模型太大

对比，只采用一周的数据的模型大小

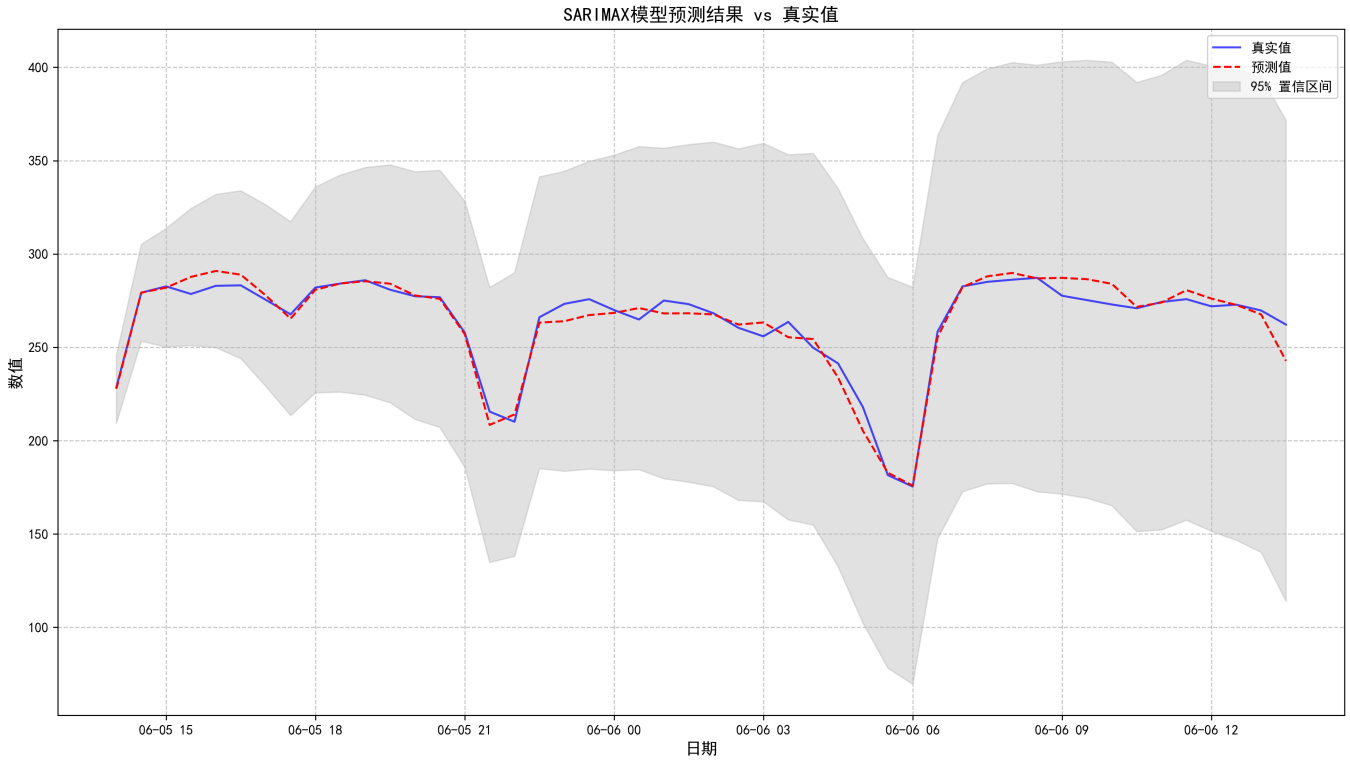
2.SARIMIA是严格按照时间连续预测的。
训练的最后日期，就是模型的预测的起点。
每次实时预测，都需要去额外增加一段预测时间

- 用少量数据边训练边预测。实时更新模型
- 与预测起点差距越大，越不准确

30分钟频率

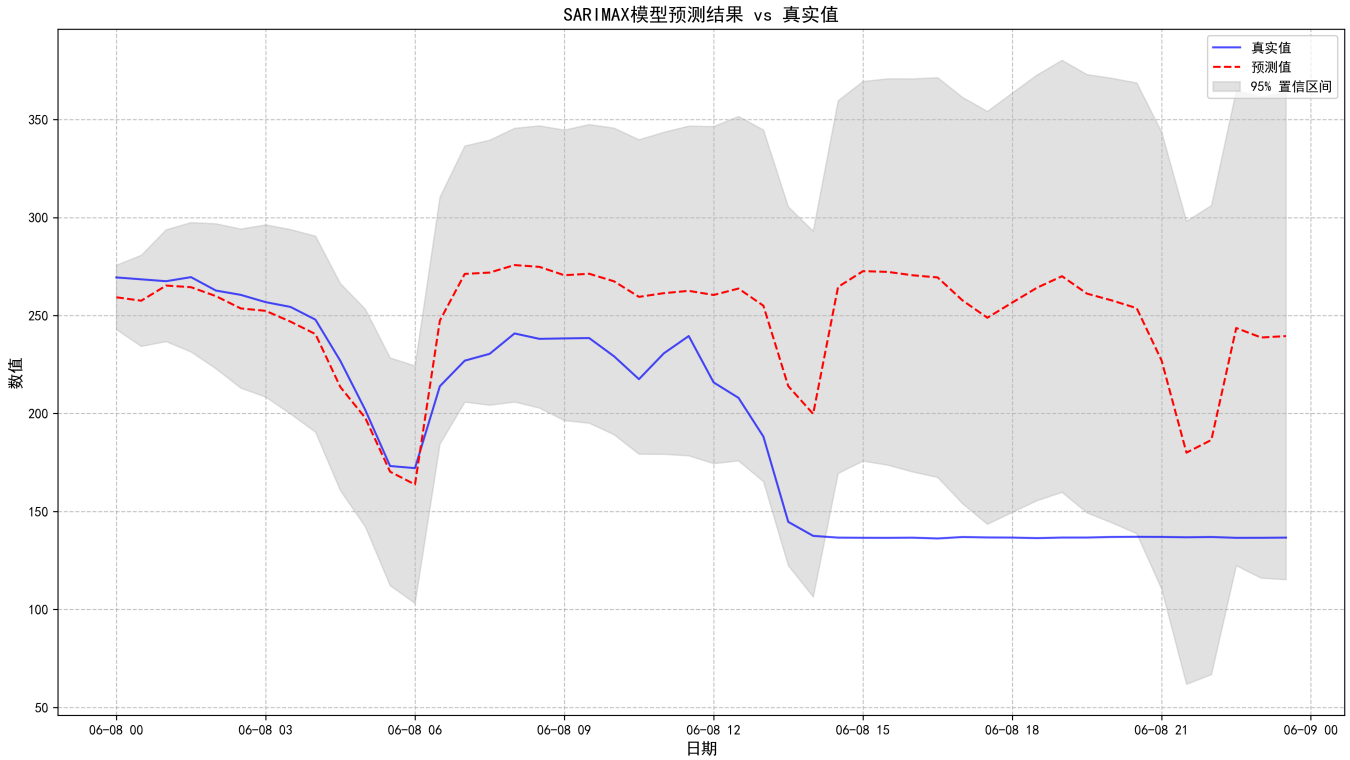
完整数据集

成功处理数据! 数据范围: 2025-05-01 06:00:00 到 2025-06-06 13:30:00
总计 1744 个30分钟间隔的数据点
每天有 48 个数据点
训练集大小: 1696 个点 (约 35.3 天)
测试集大小: 48 个点 (1天)



{
"MSE": 35.66697550731643,
"RMSE": 5.972183479039842,
"MAE": 4.287040606923704,
"MAPE": 1.6422702400644373,
"Test_size": 48,
"Predictions_size": 48
}

一周数据集



{
"MSE": 7617.262583254203,
"RMSE": 87.27693041837689,
"MAE": 69.41063109342363,
"MAPE": 46.096102710317446,
"Test_size": 144,
"Predictions_size": 144
}