2025-06-16

主要工作内容：（模型训练及测验）

- 方案1：Transformer + XGBoosts
- 方案2：随机森林 + 突变检测
- 方案3：随机森林改进
  结果展示：
- 方案1：mse = 247.0701
- 方案2：mse = 16.824👑（目前最好成绩）
- 方案3：mse= 21.1019
  最优结果可视化：

更详细的过程：（注：需要下载pdf后查看，内容较大，网址没法渲染）

明日计划

- IMF分解预测
- 研究一下时序预测库darts

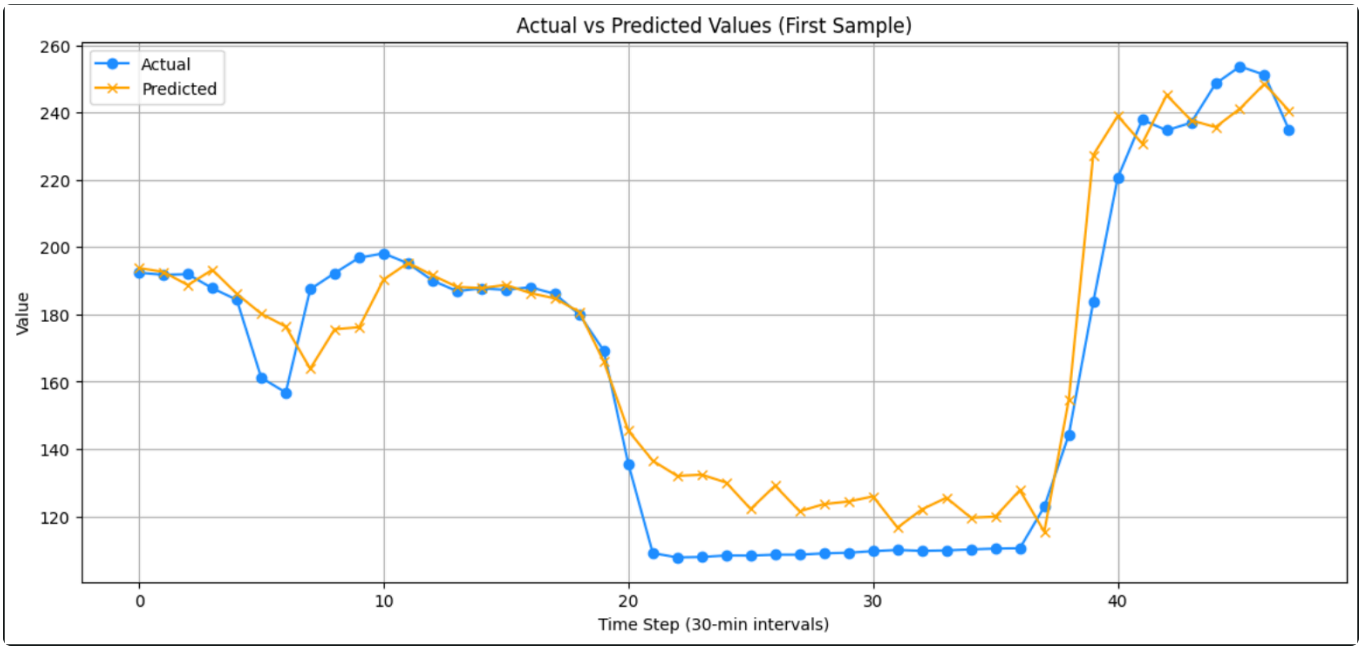
Transformer + XGBoosts

方法：
1.训练Transformer
2.用Transformer提取编码特征 + 序列特征作为XGBoost的输入
3.训练48个XGboost模型（分别预测每个点）

输入：

- Transformer: 一周的输入（48 * 7）
- XGBoost：Transformer编码特征 + 原序列中提取的特征
  输出：
- Transformer：一天（48）
- XGBoost：1点（* 48个模型）

总的效果:
MSE: 535.0076
MAE: 17.8130

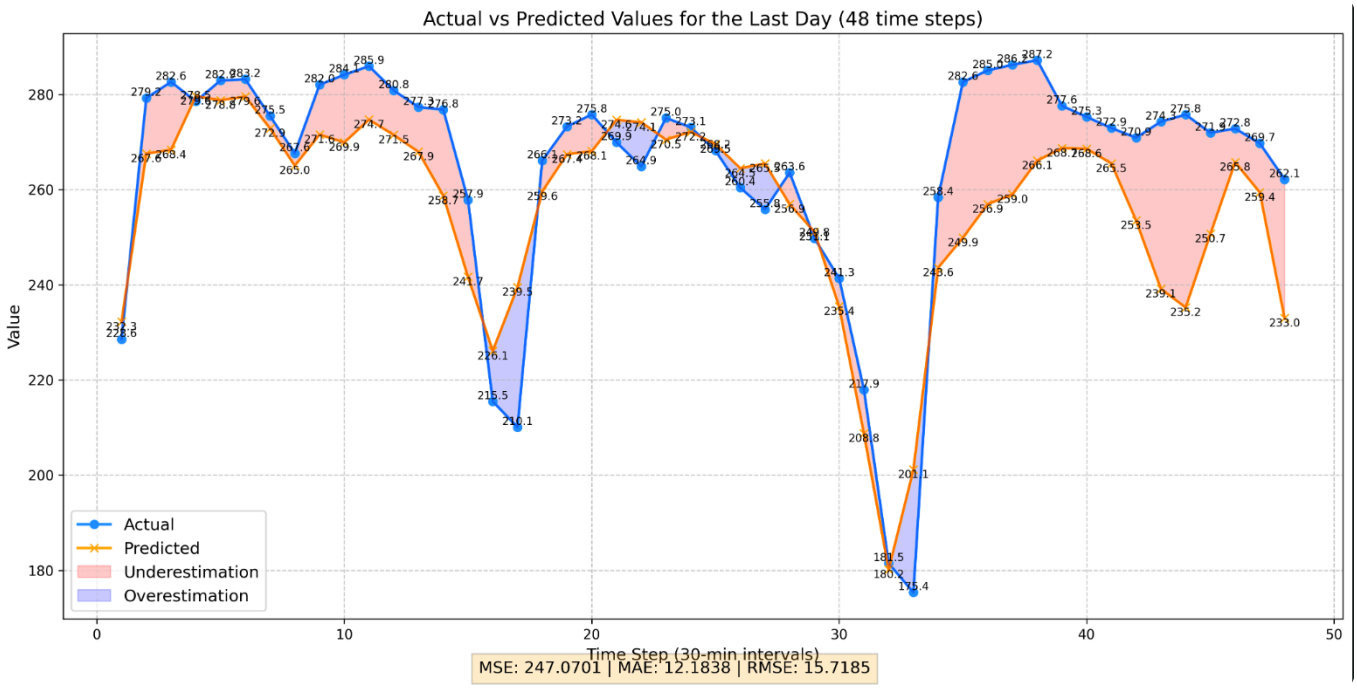


单个模型的效果:

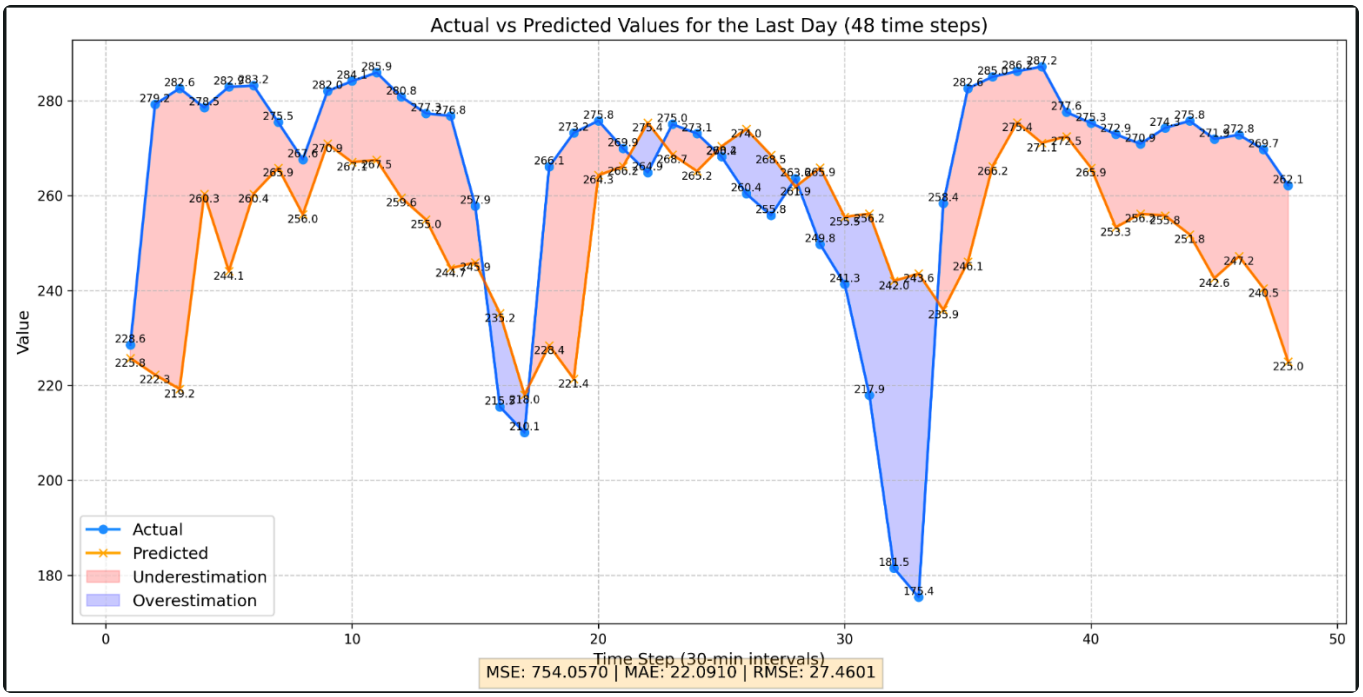
Step 1: MSE=74.2448, MAE=4.8423
Step 2: MSE=279.9203, MAE=11.9470
Step 3: MSE=591.5786, MAE=18.5707
Step 4: MSE=446.2236, MAE=16.1143
Step 5: MSE=629.2975, MAE=19.1668
Step 6: MSE=432.9554, MAE=15.5263
Step 7: MSE=476.9355, MAE=17.2186
Step 8: MSE=354.1018, MAE=14.7172
Step 9: MSE=606.9642, MAE=19.2594
Step 10: MSE=462.1265, MAE=16.8371
Step 11: MSE=593.9020, MAE=18.5792
Step 12: MSE=479.4706, MAE=17.0883
Step 13: MSE=642.3415, MAE=20.0045
Step 14: MSE=640.2535, MAE=19.7626
Step 15: MSE=583.4870, MAE=18.4363
Step 16: MSE=461.9905, MAE=15.6495
Step 17: MSE=431.4224, MAE=15.9024
Step 18: MSE=475.6483, MAE=17.0288
Step 19: MSE=624.3095, MAE=18.8540
Step 20: MSE=611.7211, MAE=18.6079
Step 21: MSE=711.6444, MAE=20.4869
Step 22: MSE=509.4037, MAE=17.3298
Step 23: MSE=497.9035, MAE=18.2152
Step 24: MSE=341.4684, MAE=14.3476
Step 25: MSE=360.7887, MAE=13.6220
Step 26: MSE=307.0382, MAE=12.7648

Step 27: MSE=371.2136, MAE=14.1283
Step 28: MSE=299.3625, MAE=12.5451
Step 29: MSE=438.4935, MAE=16.0936
Step 30: MSE=444.1522, MAE=16.5440
Step 31: MSE=501.5207, MAE=18.0142
Step 32: MSE=292.2561, MAE=12.9397
Step 33: MSE=385.7864, MAE=15.7438
Step 34: MSE=388.1755, MAE=15.0495
Step 35: MSE=590.8812, MAE=19.3559
Step 36: MSE=551.7031, MAE=20.2651
Step 37: MSE=687.6052, MAE=23.1294
Step 38: MSE=342.8858, MAE=16.4413
Step 39: MSE=429.3191, MAE=17.4981
Step 40: MSE=323.4023, MAE=14.3016
Step 41: MSE=503.8669, MAE=18.3672
Step 42: MSE=663.1172, MAE=22.4538
Step 43: MSE=1450.5941, MAE=34.2233
Step 44: MSE=1524.3716, MAE=35.7885
Step 45: MSE=1144.5945, MAE=29.9175
Step 46: MSE=773.0673, MAE=20.9544
Step 47: MSE=530.9203, MAE=16.4336
Step 48: MSE=415.9357, MAE=13.9590

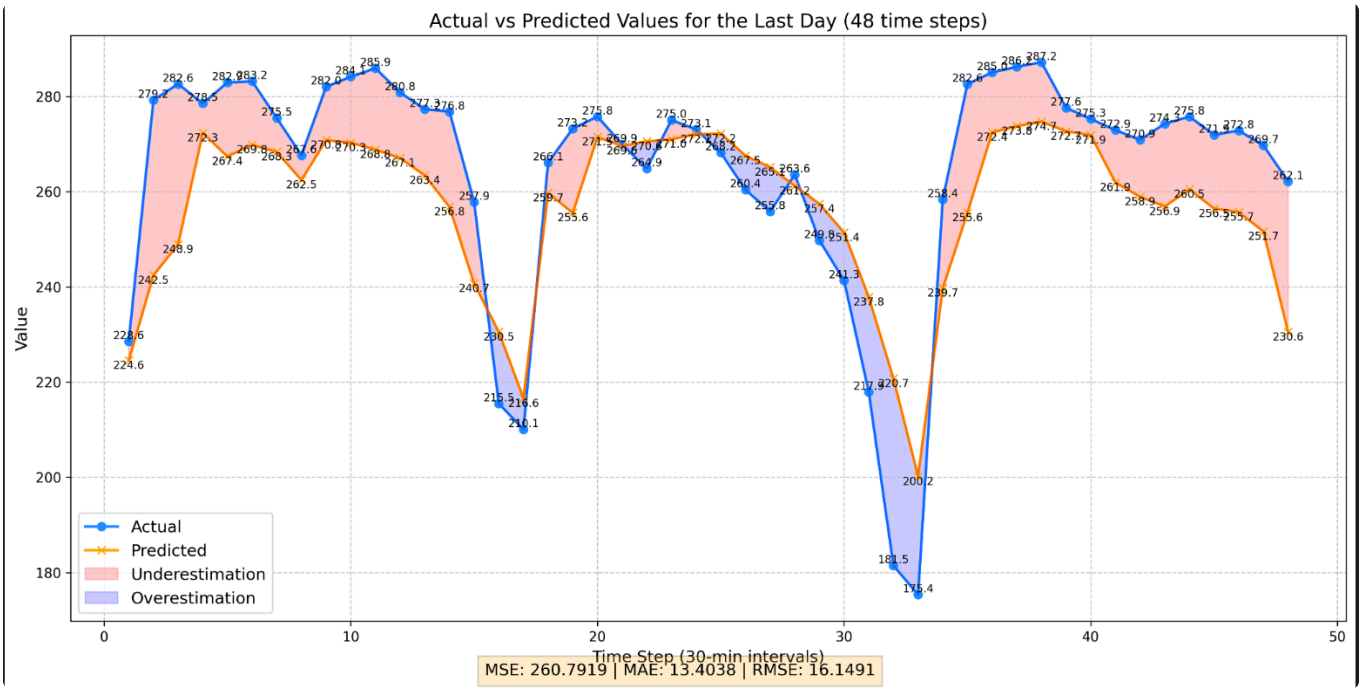
Performance on Last Day (48 time steps):
MSE: 247.0701
MAE: 12.1838
RMSE: 15.7185



MSE: 754



MSE: 260.7919



随机森林 + 突变检测

#二阶段

方法：随机森林 + 突变检测

模型：随机森林

数据集：

- 范围：2025-05-02 06:00:00 到 2025-06-06 13:30:00
- 长度：1696 （30Min采样）
- 特征数：22
- 使用的特征：['hour', 'weekday', 'day', 'residual', 'day_part', 'residual_lag1', 'SARIMA_lag1', 'residual_lag2', 'SARIMA_lag2', 'residual_lag3', 'SARIMA_lag3', 'residual_lag24', 'SARIMA_lag24', 'residual_lag48', 'SARIMA_lag48', 'rolling_8h_mean', 'day_of_week', 'day_of_month', 'month']

策略：

- 1.先采用随机森林训练
- 2.识别突变点
  - 突变点：
    - 通过残差分布，计算出置信区间
    - 超出置信区间的点，被称为突变点
      - anomaly_mask = (y < lower_bound) | (y > upper_bound)
- 3.处理突变点并重新训练
  - 方法1：直接剔除突变点
  - 方法2：将突变点编码为特征（采用独热编码）

= 第一阶段：初始模型训练和突变点检测 =
检测到 110 个突变点 (占总数据 6.49%)

初始模型 性能:
MAE: 2.4166
MSE: 16.8244
RMSE: 4.1018
R^2: 0.9946
残差标准差: 4.1010

= 策略一：剔除突变点后重新训练 =
剔除策略: 使用 1586 个样本 (原始样本 1696)

剔除策略模型 性能:
MAE: 4.3133
MSE: 130.3215
RMSE: 11.4158
R^2: 0.9584

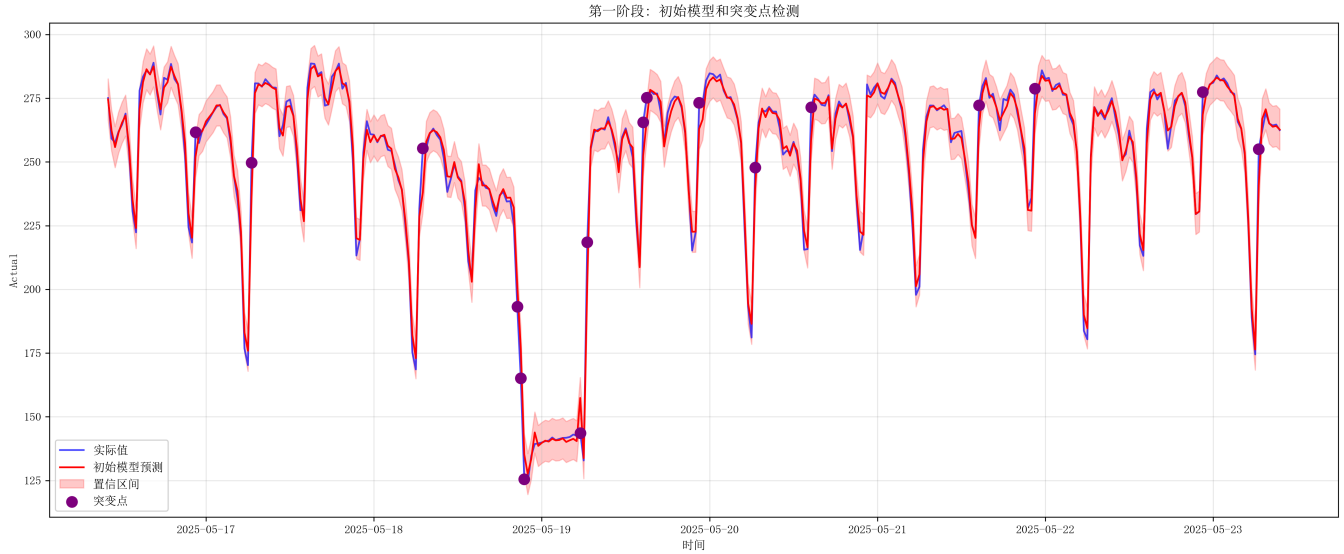
= 策略二：将突变点作为模型参数 =
参数增强策略: 添加 110 个突变点特征

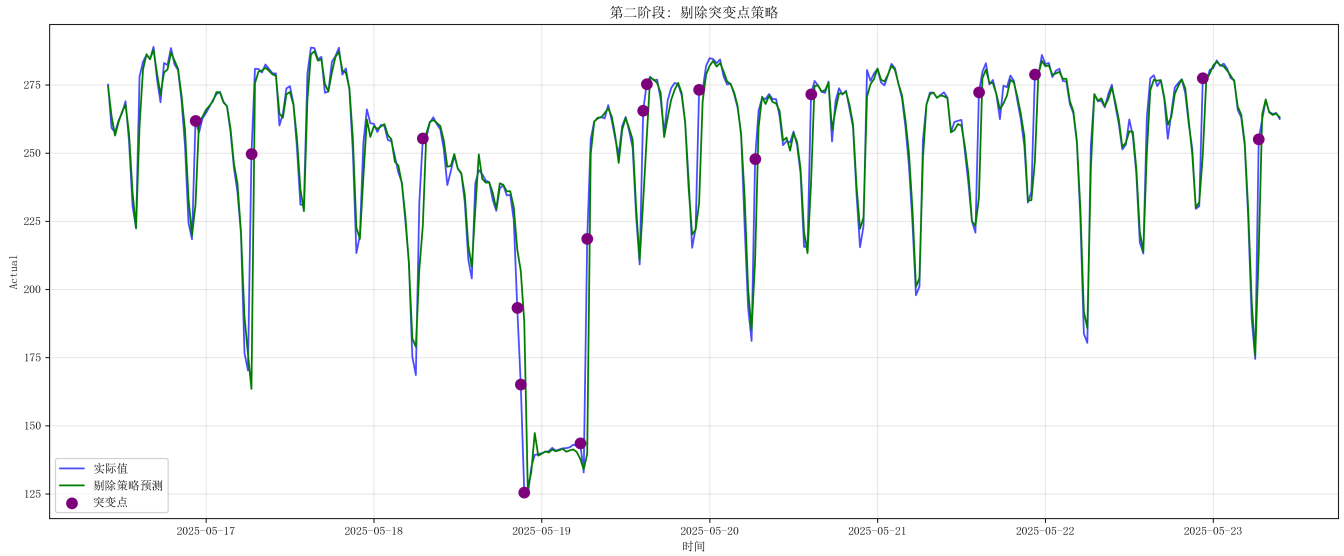
参数增强模型 性能:
MAE: 2.4037
MSE: 16.8862

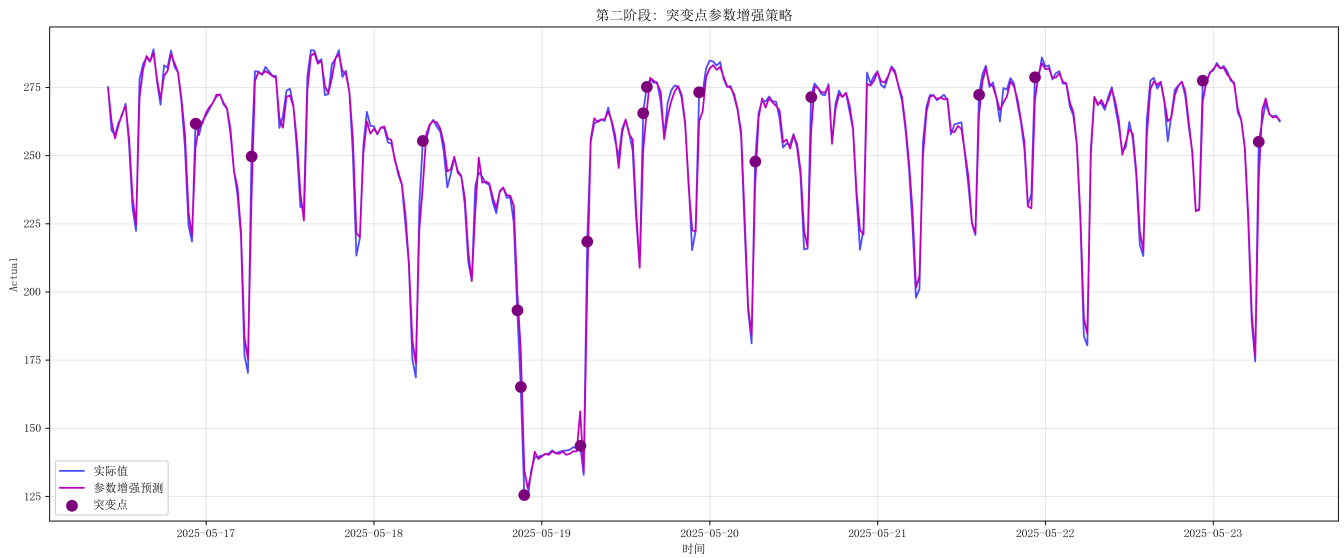
RMSE: 4.1093
R^2: 0.9946

模型性能汇总:
model mae mse rmse r2
0 初始模型 2.416622 16.824443 4.101761 0.994628
1 剔除策略模型 4.313301 130.321480 11.415843 0.958386
2 参数增强模型 2.403729 16.886206 4.109283 0.994608

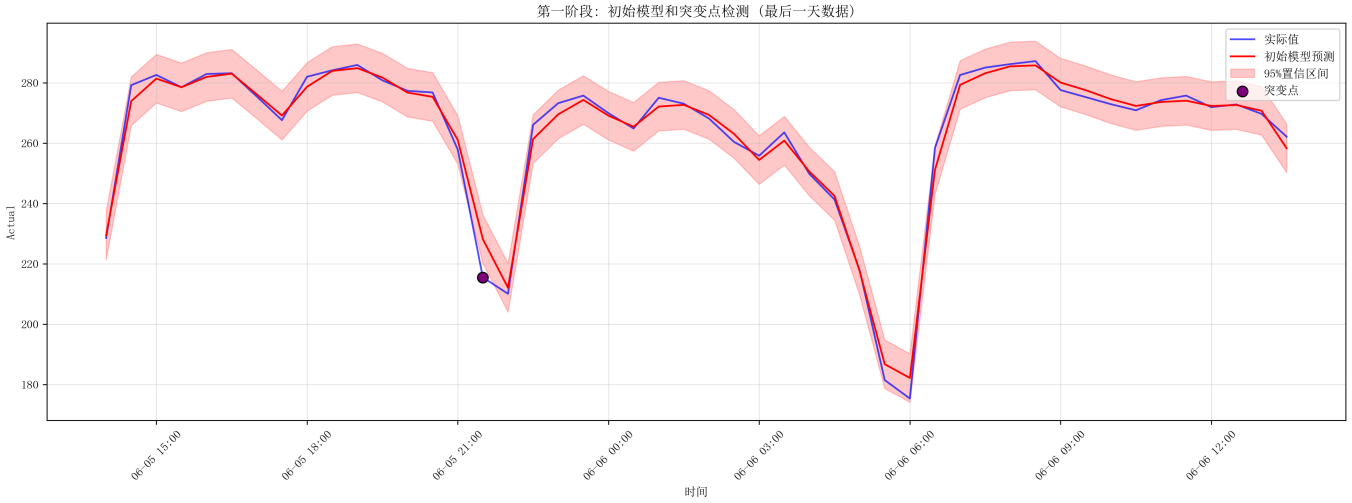
一周的预测效果（5-17 ~ 5-23）

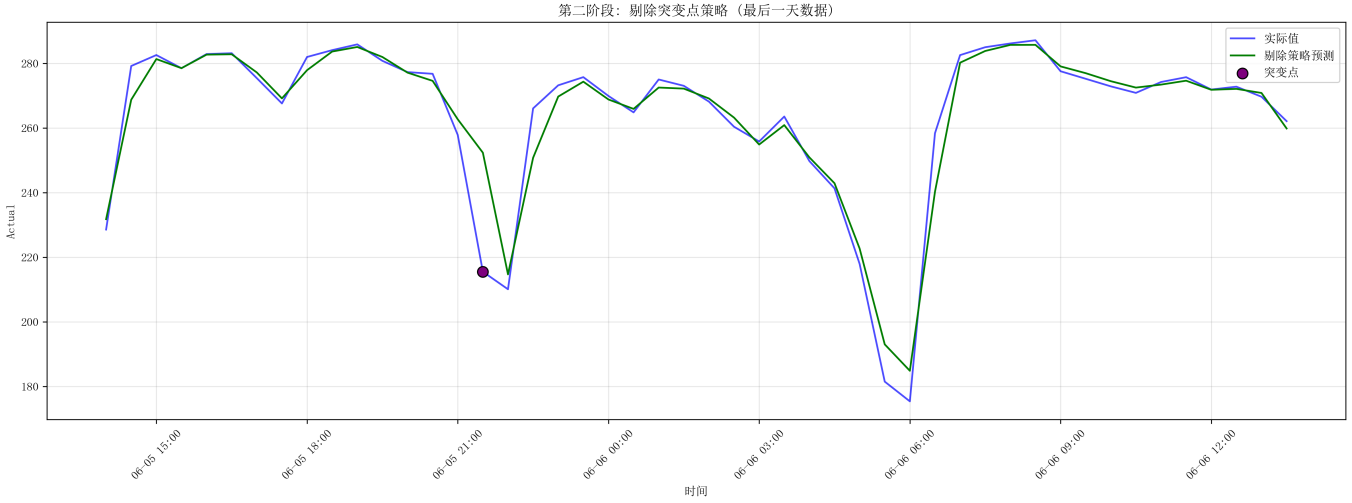

第一阶段：初始模型和突变点检测
Actual
时间
实际值
初始模型预测
置信区间
突变点
2025-05-17
2025-05-18
2025-05-19
2025-05-20
2025-05-21
2025-05-22
2025-05-23

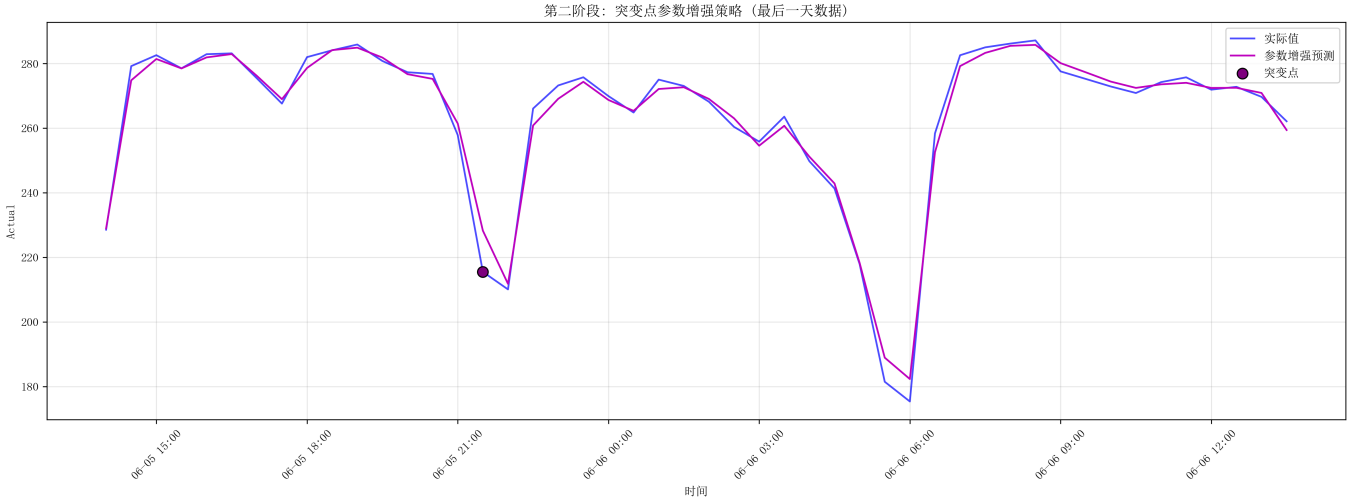

第二阶段：剔除突变点策略
Actual
时间
实际值
剔除策略预测
突变点
2025-05-17
2025-05-18
2025-05-19
2025-05-20
2025-05-21
2025-05-22
2025-05-23


第二阶段：突变点参数增强策略
Actual
时间
实际值
参数增强预测
突变点
2025-05-17
2025-05-18
2025-05-19
2025-05-20
2025-05-21
2025-05-22
2025-05-23

一天的预测效果

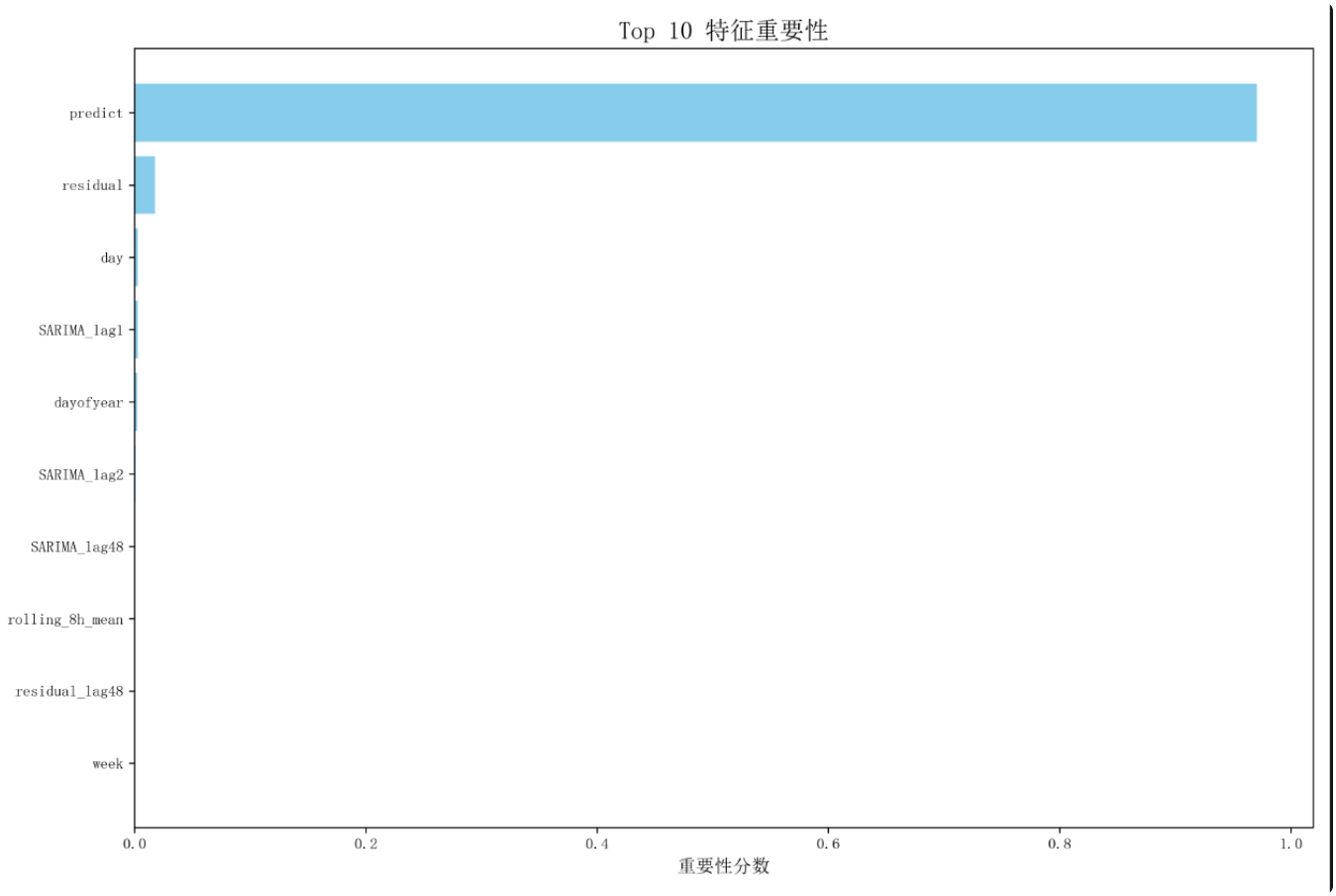




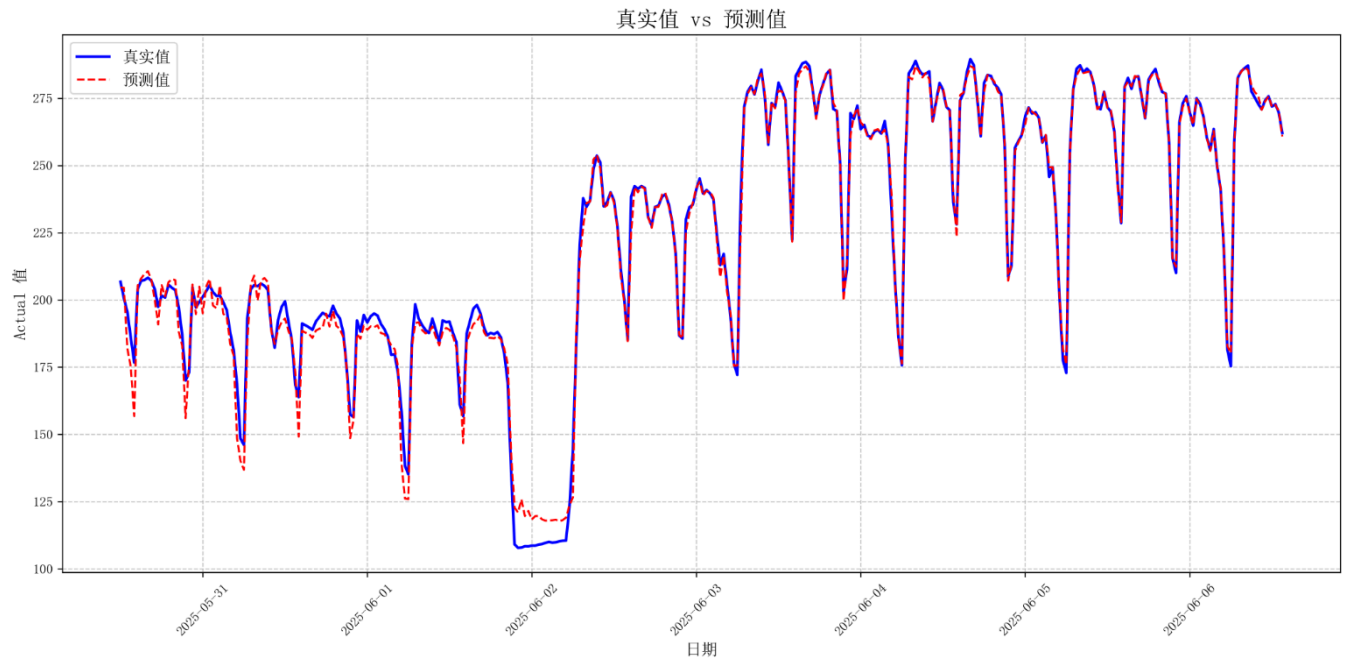


纯随机森林优化

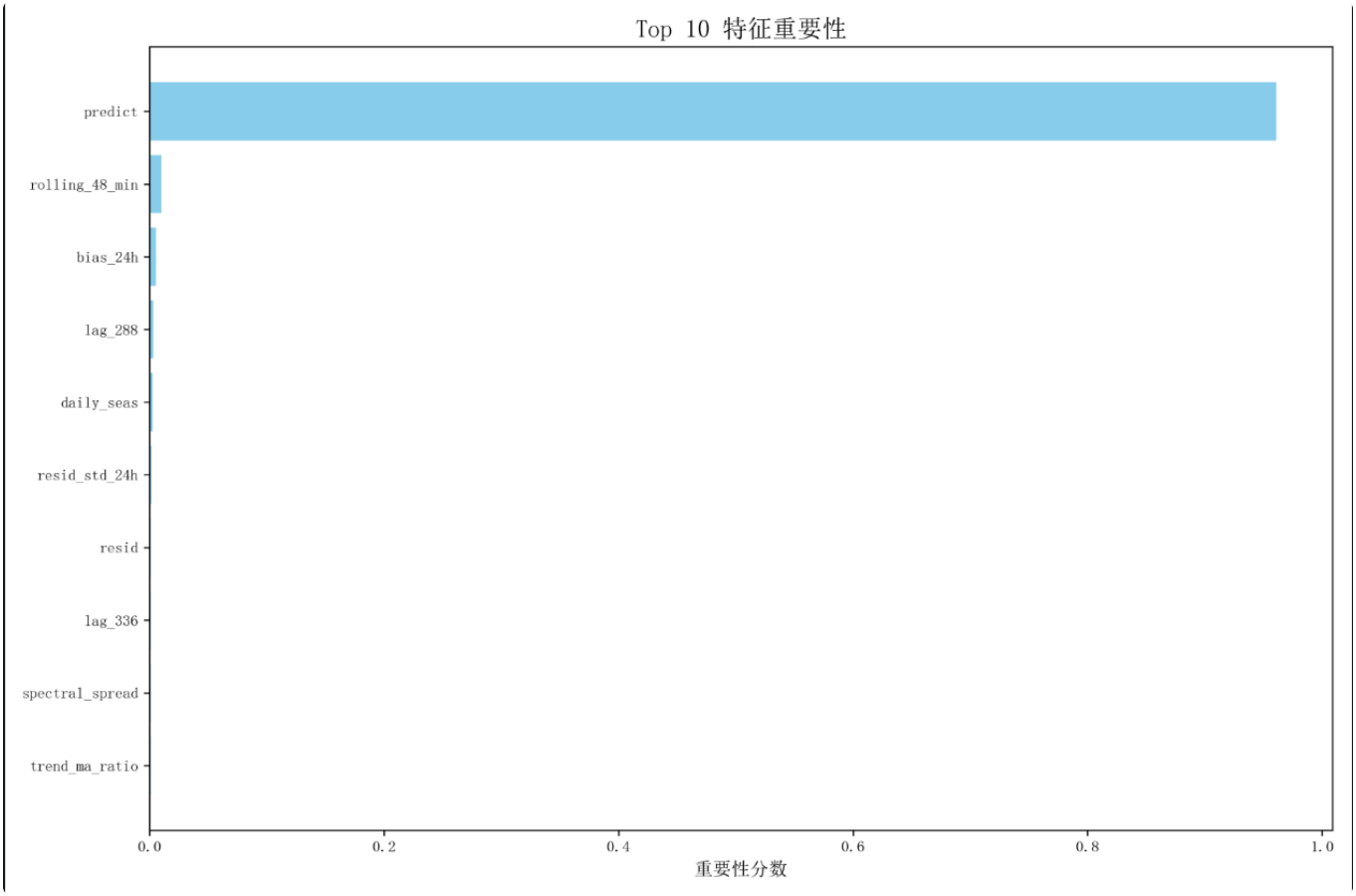
原始特征



评估指标:
MSE: 21.1019
RMSE: 4.5937
MAE: 2.7904
R^2: 0.9908



加入频域和趋势特征



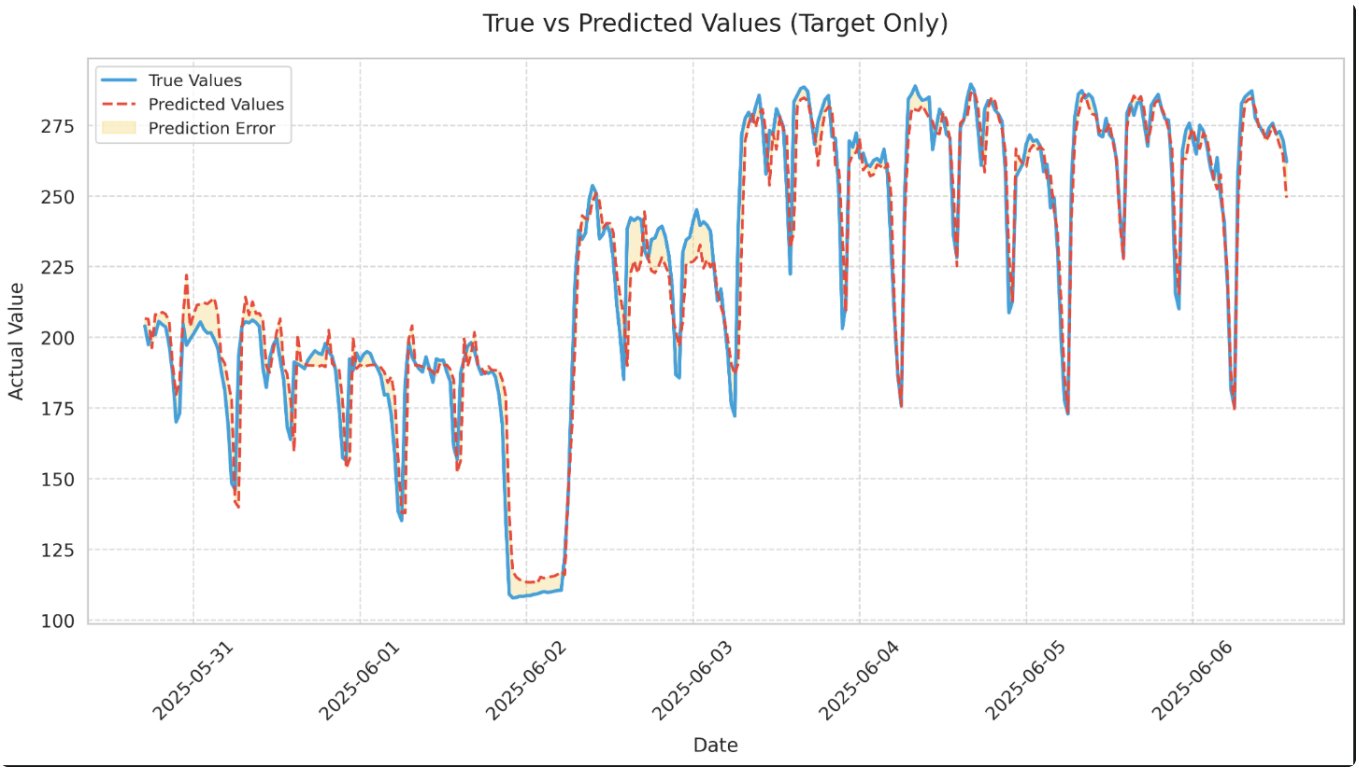
MSE: 124.6157
RMSE: 11.1631
MAE: 8.0628
R^2: 0.9499

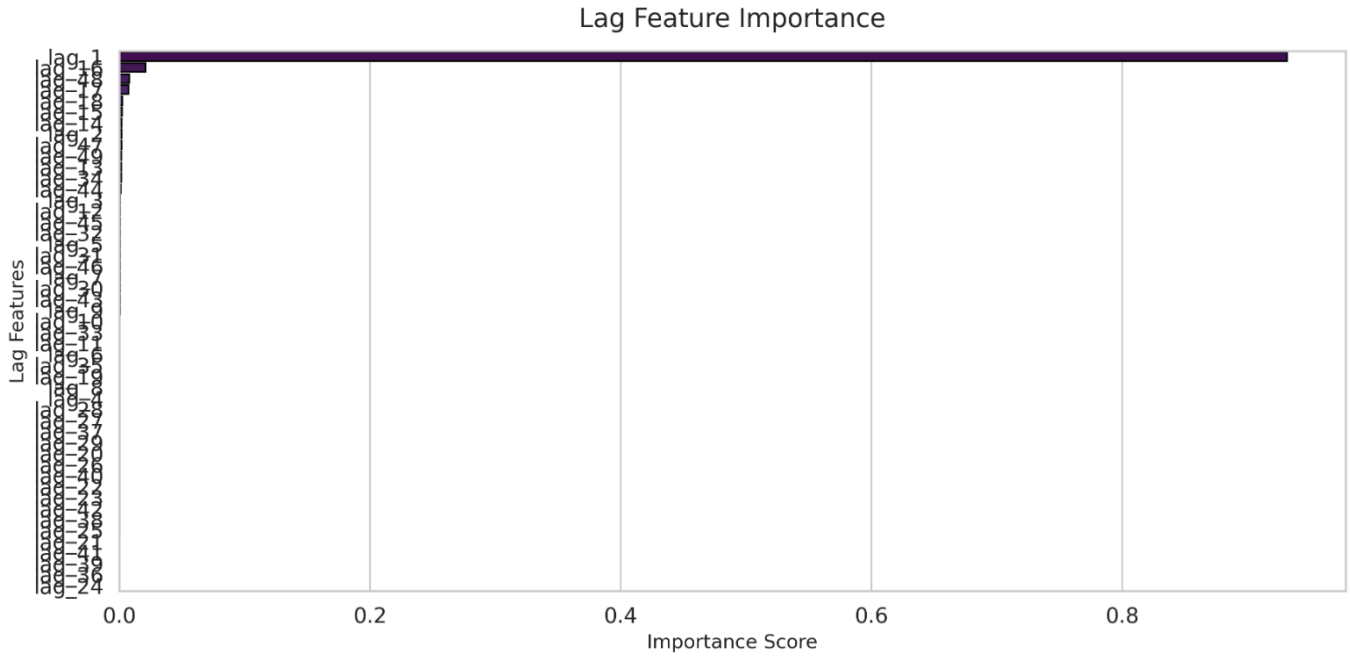
仅使用排名前2的两个特征

MSE: 103.2793
RMSE: 10.1626
MAE: 7.3237
R^2: 0.9548

一阶段使用随机森林做预测

最佳参数:
{'max_depth': 15, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 300}





评估指标:
MSE: 132.3881
RMSE: 11.5060
MAE: 7.7979
R^2: 0.9432

仅用排分高的3个滞后特征
lag_1,lag_16,lag_48
评估指标:

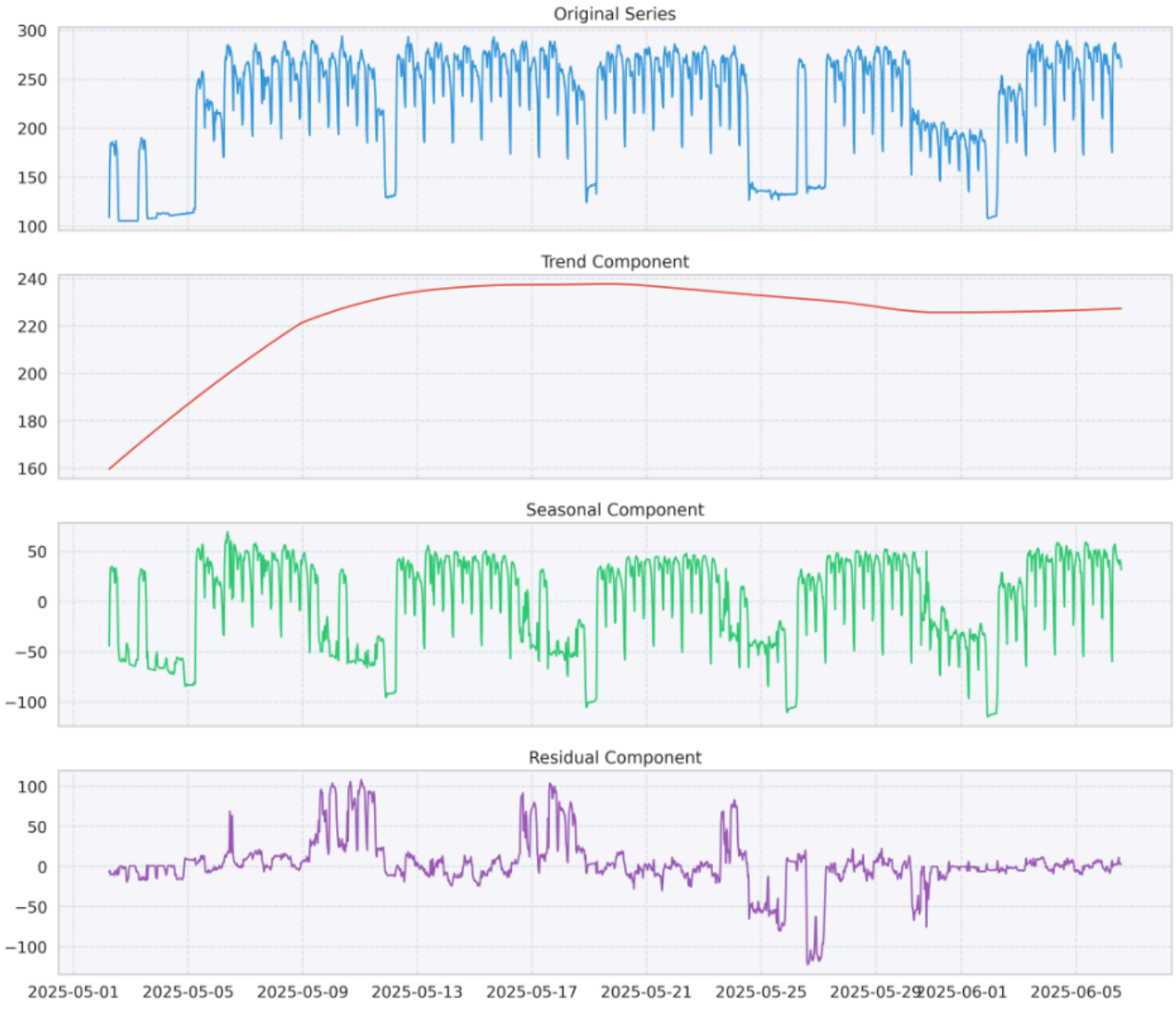
MSE: 194.0482
RMSE: 13.9301
MAE: 9.5575
R^2: 0.9168

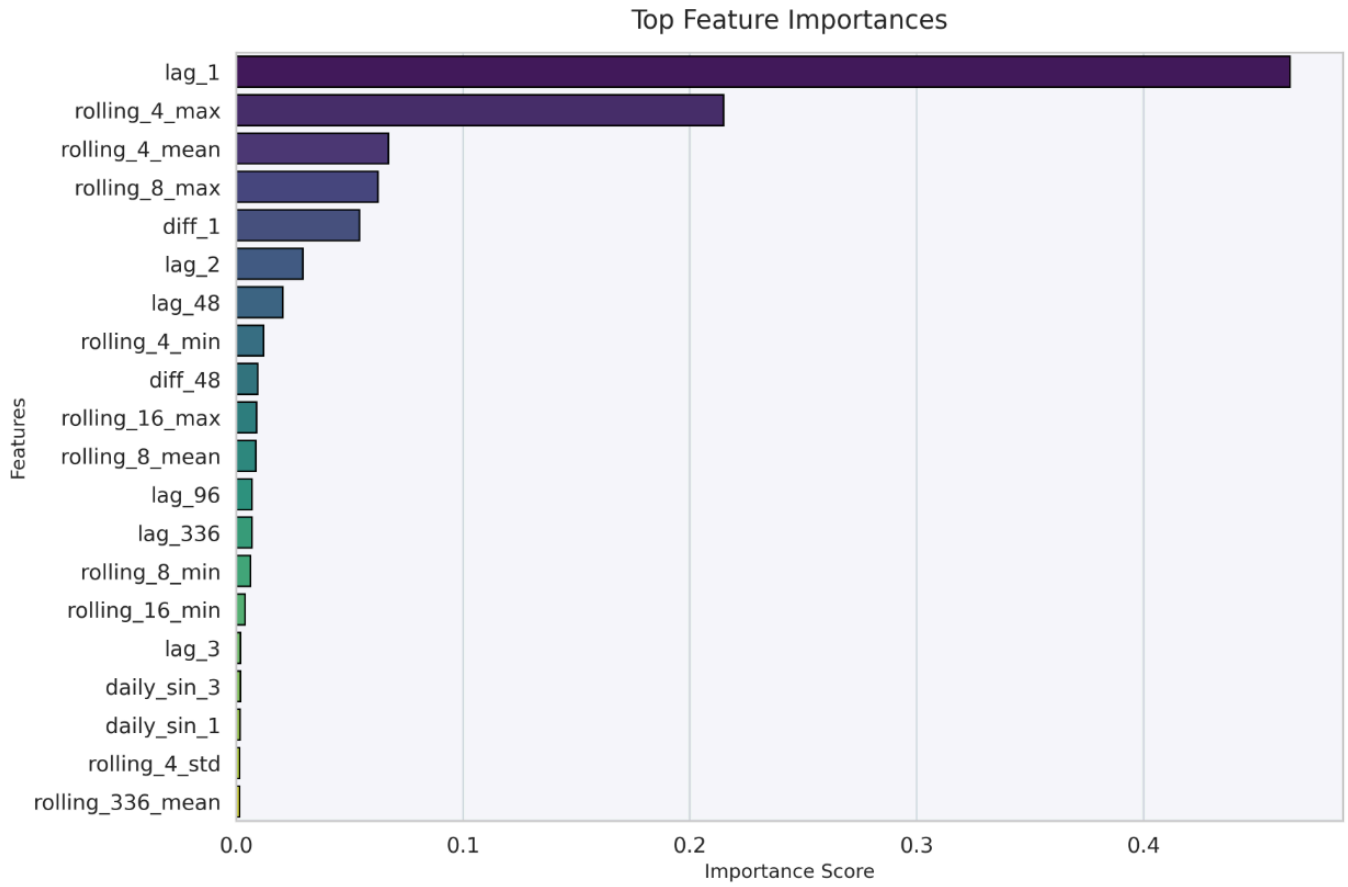
Random Forest w/ Cond. Deseasonalize & Detrending

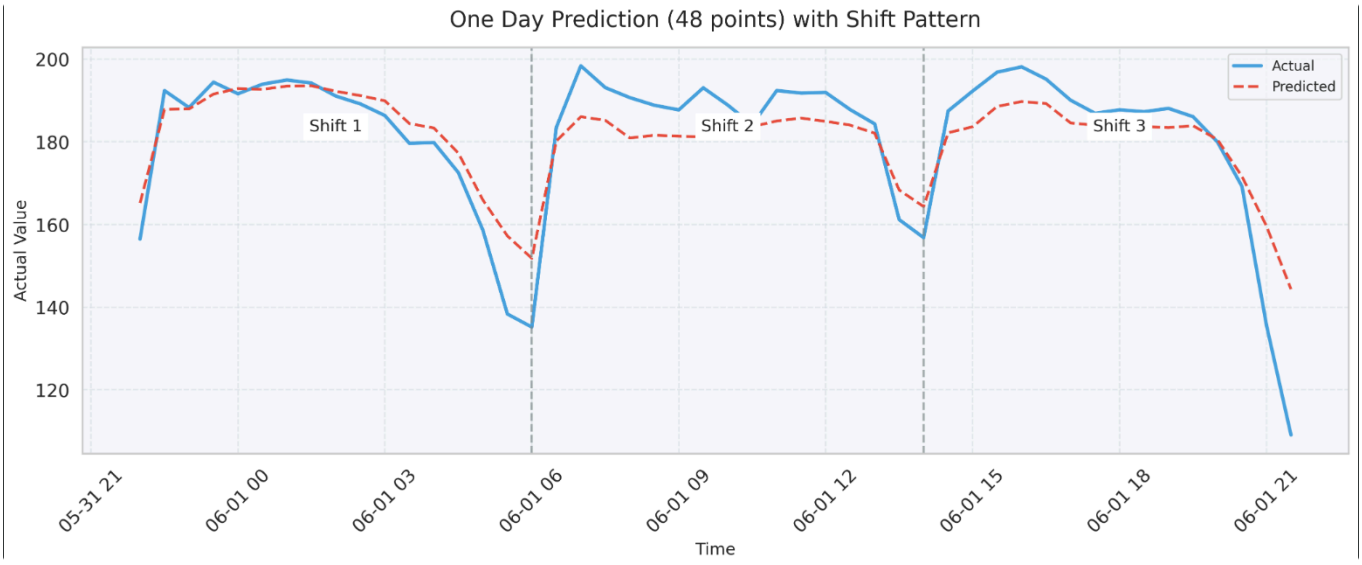
（一阶段）
方法：（时序分解预测法）

1. 首先，检测时间序列的季节性，并进行去季节化STL分解（如果存在显著季节性）。
2. 然后，去除时间序列的趋势线性回归（可能是线性或非线性趋势）。
3. 在去季节化和去趋势后的残差序列上建立随机森林模型（使用滞后特征）。
4. 预测后，再将季节性和趋势成分加回到预测结果中

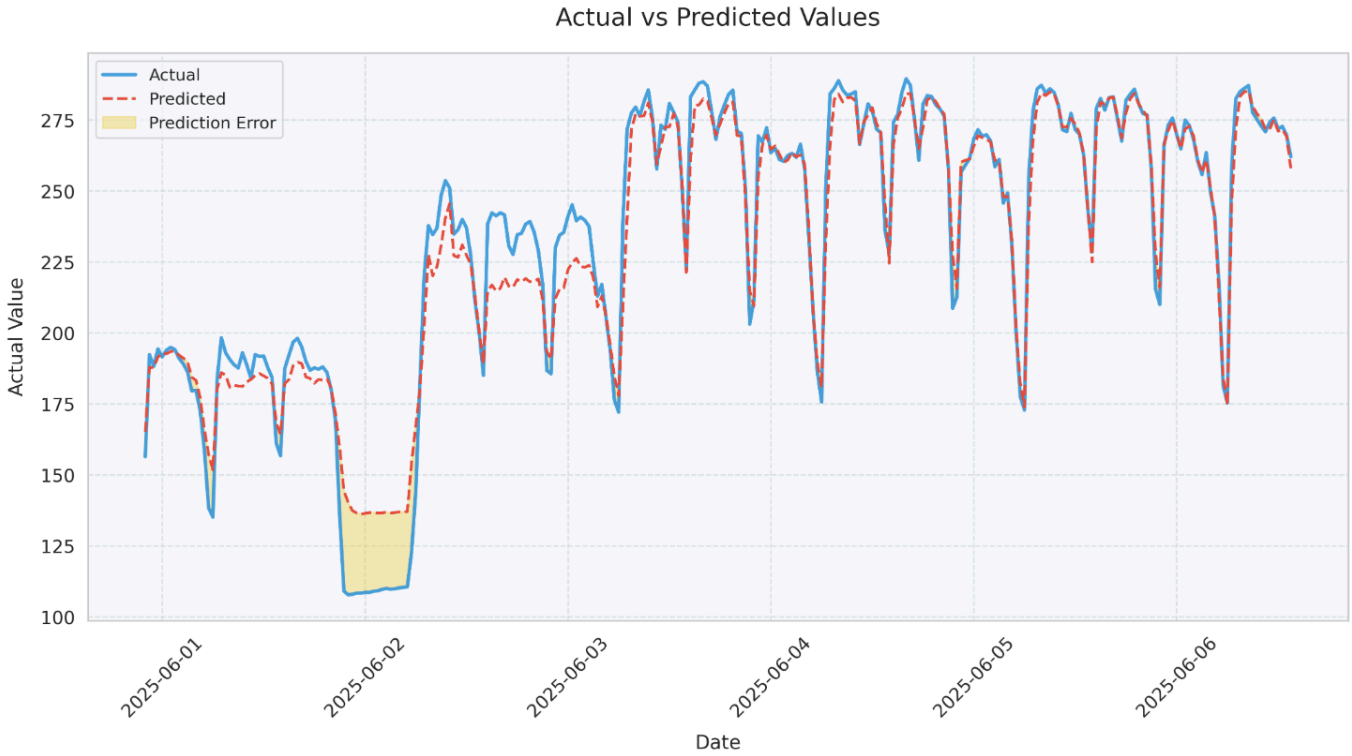
STL分解







Test set metrics:
MSE: 120.9230
RMSE: 10.9965
MAE: 7.1077
R^2: 0.9513



#最优超参数

二阶段

Best parameters found:

model**bootstrap: False**

modelmax_depth: 10

model**max_features: 0.42066805426927745**

modelmin_samples_leaf: 1

model**min_samples_split: 2**

modeln_estimators: 170

Best CV score (negative MSE): -110.7858

Test set metrics:

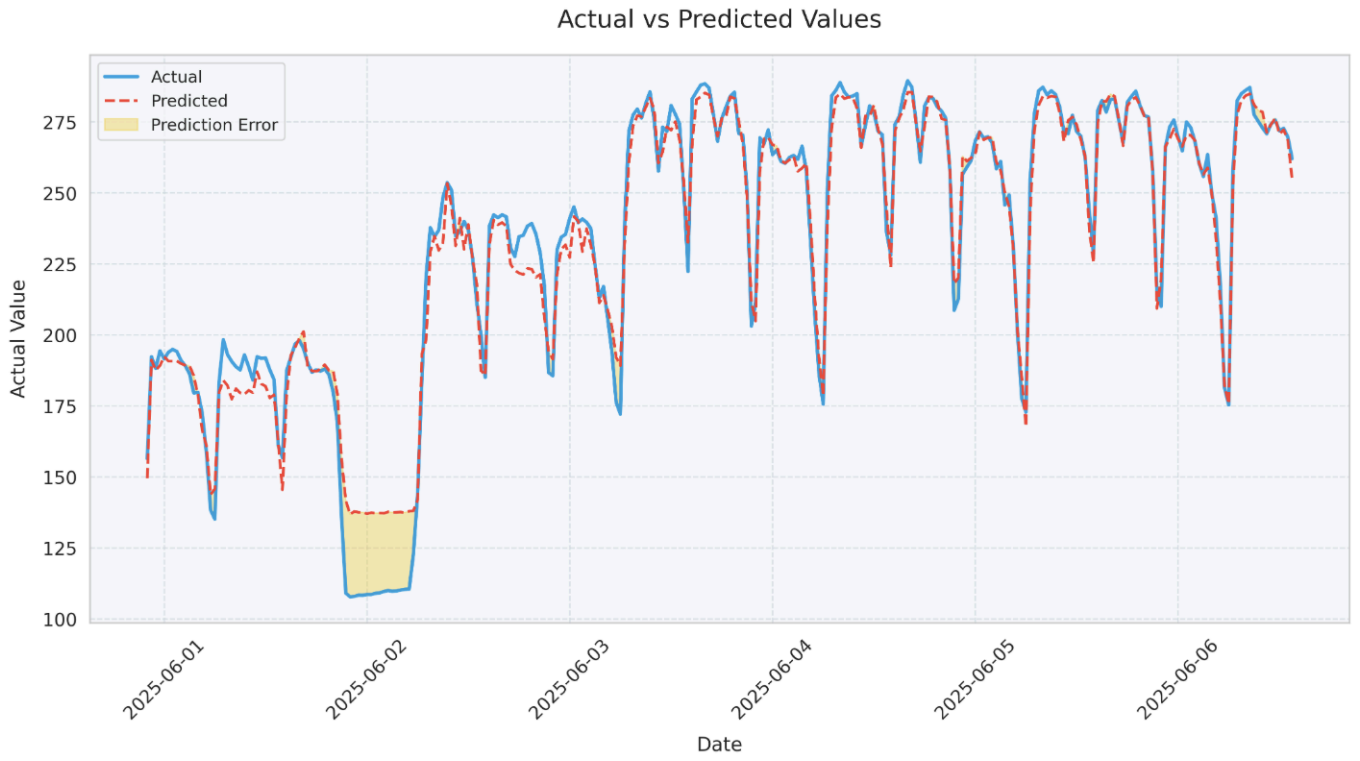
MSE: 84.1480

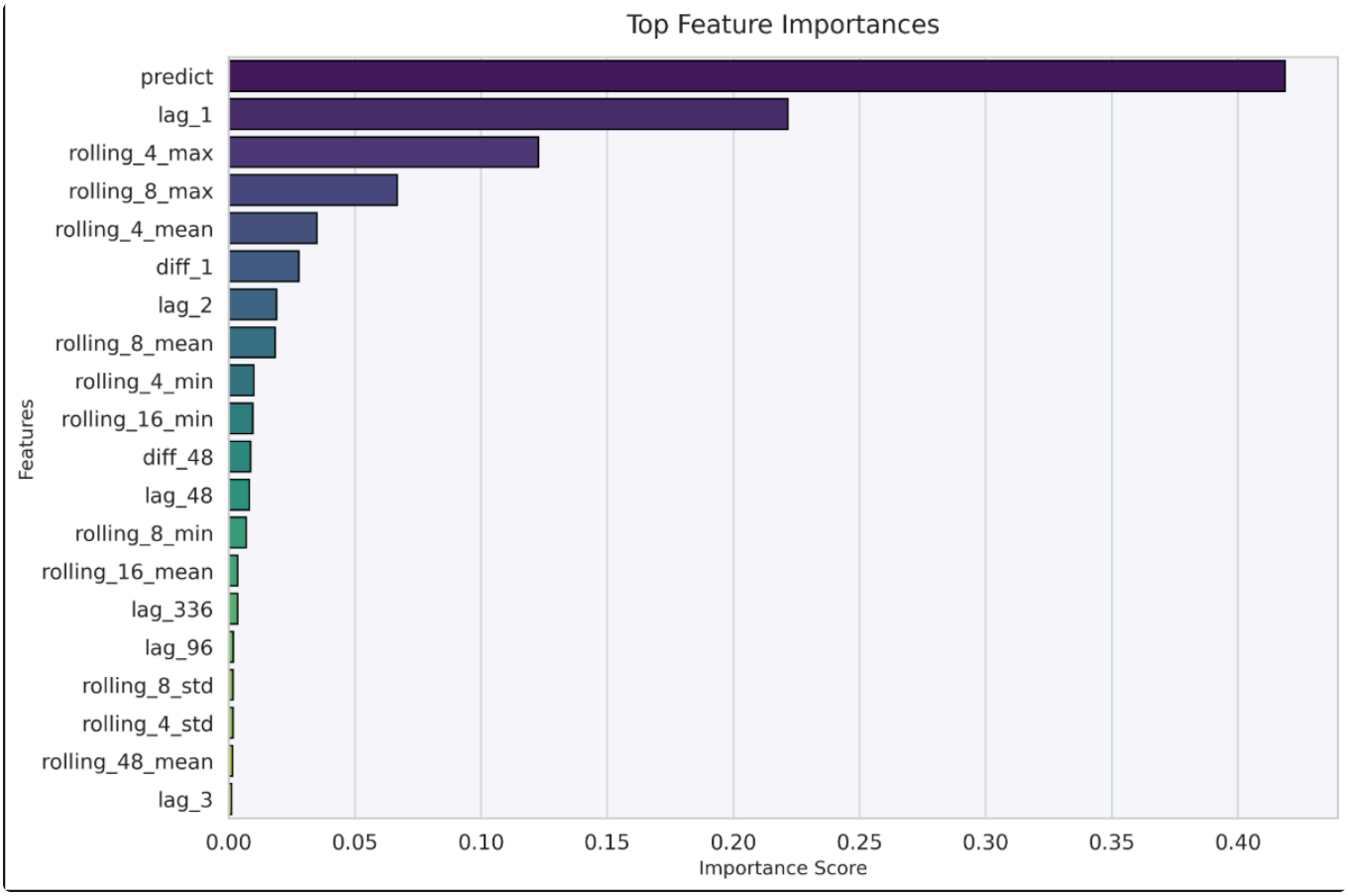
RMSE: 9.1732

MAE: 5.9124

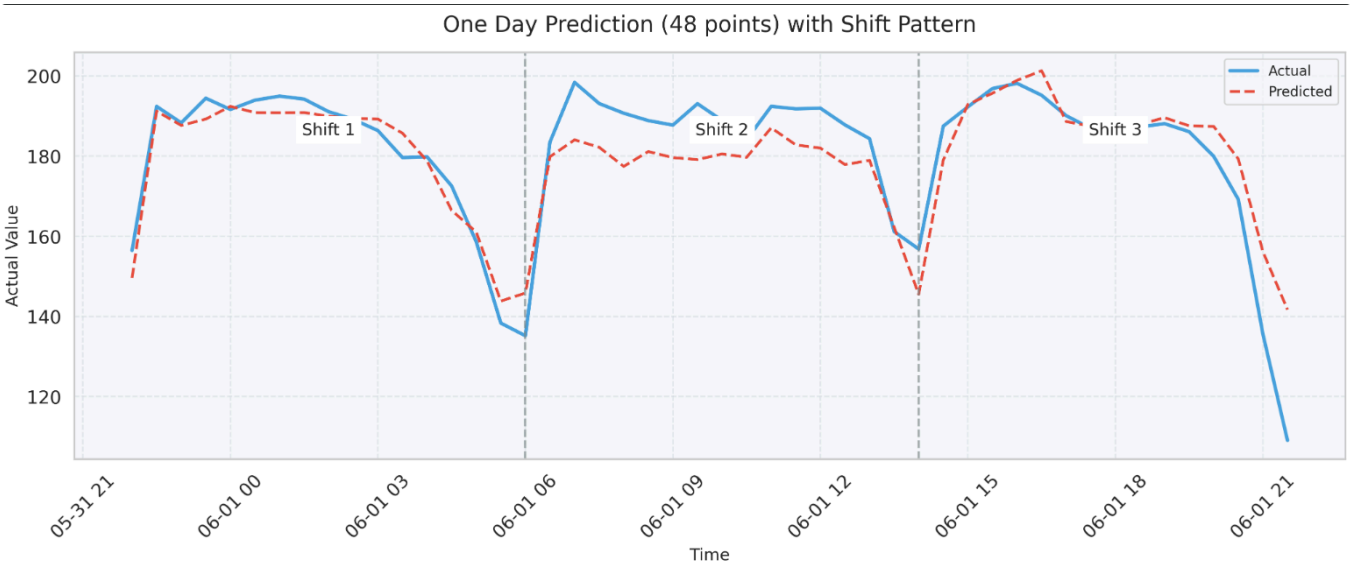
R^2: 0.9661

测试集预测效果





一天内的预测效果


One Day Prediction (48 points) with Shift Pattern
Actual
Predicted
Shift 1
Shift 2
Shift 3
Actual Value
200
180
160
140
120
05-31 21
06-01 00
06-01 03
06-01 06
06-01 09
06-01 12
06-01 15
06-01 18
06-01 21
Time