RFECV 결과를 분석해보니 **예상과 다른 흥미로운 패턴**이 나타났습니다. 자세히 분석해드리겠습니다.

**📊 RFECV 결과 분석**

**1. 선택된 6개 특성**

선택된 특성:

1. TIME (시계열 인덱스)

2. age (개체 나이)

3. expec\_mdate\_diff (예상 분만일까지 남은 일수)

4. ex\_year (예상 분만 연도)

5. lin\_year (수정 연도)

6. lin\_mon (수정 월)

RMSE: 4.12일 (예측 오차)

**2. 놀라운 발견: 행동 데이터가 제외됨! 🤔**

**제외된 주요 특성들:**

* ❌ mact, mact4 (활동량)
* ❌ mrum, mrum4 (반추 시간)
* ❌ stdact, stdrum (행동 변동성)
* ❌ m4\_act\_diff, m4\_rum\_diff (행동 변화량)

**3. 왜 이런 결과가 나왔을까?**

**가능성 1: 데이터 누수 (Data Leakage)**

# expec\_mdate\_diff와 bun\_mdate\_diff의 관계

expec\_mdate\_diff = 예상분만일 - 측정일

bun\_mdate\_diff = 실제분만일 - 측정일

# 두 변수는 매우 높은 상관관계를 가짐

# 예상 분만일이 정확하다면 거의 동일한 값

**가능성 2: 시간 정보의 과도한 영향**

* TIME: 시계열 순서
* ex\_year, lin\_year, lin\_mon: 날짜 정보
* 이들이 실제 분만일과 강한 패턴을 보일 수 있음

**4. 문제점 진단## 🚨 핵심 문제: 데이터 누수 (Data Leakage)**

**1. 가장 큰 문제: expec\_mdate\_diff 선택**

# 이 두 변수의 관계

expec\_mdate\_diff = 예상분만일(고정) - 측정일(변함)

bun\_mdate\_diff = 실제분만일(고정) - 측정일(변함)

# 예시

측정일: 2023-02-11

예상분만일: 2023-02-18 (수정일 + 283일)

실제분만일: 2023-02-26

expec\_mdate\_diff = -7일

bun\_mdate\_diff = -15일

**문제**: expec\_mdate\_diff는 타겟과 거의 동일한 정보를 담고 있어, 모델이 이 변수에만 의존하게 됨

**2. 실제로 유의미한 변수들 (하지만 제외됨)## 📌 결론 및 권장사항**

**1. 현재 RFECV 결과의 문제점**

* expec\_mdate\_diff가 선택됨 → **데이터 누수**
* 행동 데이터가 모두 제외됨 → **생물학적 신호 무시**
* 날짜 변수들이 과도하게 선택됨 → **과적합 위험**

**2. 올바른 접근 방법**

**Step 1: 변수 제외**

# 반드시 제외해야 할 변수

exclude = ['expec\_mdate\_diff', 'TF', 'ex\_year', 'ex\_mon',

'lin\_year', 'lin\_mon', 'TIME']

**Step 2: 핵심 변수 포함**

# 반드시 포함해야 할 변수

include = ['mact', 'mrum', 'stdact', 'stdrum',

'm4\_act\_diff', 'm4\_rum\_diff', 'age', 'par']

**3. 실제로 유의미한 변수 (생물학적 관점)**

| **순위** | **변수명** | **의미** | **중요도** |
| --- | --- | --- | --- |
| 1 | mact | 현재 활동량 | 매우 높음 |
| 2 | mrum | 현재 반추 시간 | 매우 높음 |
| 3 | stdact | 활동량 변동성 | 높음 |
| 4 | m4\_act\_diff | 4일 전 대비 활동 변화 | 높음 |
| 5 | age | 개체 나이 | 중간 |
| 6 | par | 산차 | 중간 |

**4. 개선된 모델 예상 성능**

* 현재: RMSE 4.12일 (데이터 누수 포함)
* 개선 후: RMSE 5-6일 (실제 예측력)
* 하지만 더 **신뢰할 수 있고 일반화 가능한** 모델

**5. 다시 실행하기**

# 올바른 RFECV 실행 코드

# 1. expec\_mdate\_diff 제외

feature\_cols = [col for col in df.columns

if col not in exclude\_list]

# 2. 생물학적 특성만 사용

biological\_features = ['mact', 'mrum', 'stdact', 'stdrum',

'm4\_act\_diff', 'm4\_rum\_diff', 'age', 'par',

'mact4', 'mrum4', 'mstdact4', 'mstdrum4']

X = df[biological\_features]

y = -df[df['bun\_mdate\_diff'] < 0]['bun\_mdate\_diff'] # 양수 변환

**핵심**: 현재 결과는 통계적으로는 좋아 보이지만, 실제 분만 예측에는 도움이 되지 않습니다. 행동 데이터를 포함한 생물학적으로 의미 있는 변수로 다시 분석하시기를 강력히 권장합니다!