|  |  |
| --- | --- |
| **| 문제의도** | * 과제 1에서 성능이 낮았던 이유는 단일 피처(요일)만 사용했기 때문입니다. 이를 개선하기 위해 0시부터 23시까지 모든 시간대의 교통량을 추가 피처로 사용하여, 모델이 시간대 간의 상관관계를 학습하게 하고, 성능을 개선하는 것이 목표입니다. |
| **| 사전지식** | * pandas를 사용한 데이터 로딩 및 필터링 방법 * scikit-learn의 선형 회귀 모델 사용법 * 학습 데이터와 테스트 데이터를 나누는 방법 |
| **| 문제핵심** | * 추가 피처: 2023년 01월 시트에 있는 0시부터 23시까지의 모든 시간대 교통량을 추가하여 0시 교통량을 예측 * 학습 데이터와 테스트 데이터 분리: train\_test\_split을 사용하여 데이터를 학습용과 테스트용으로 나눔 * 성능 평가: R² 점수를 통해 성능을 평가하고 분석 |

|  |  |
| --- | --- |
| **| 문제해설** | |
| **| 해설** | 1. 피처 추가: 모든 시간대 교통량 정보 도입  과제 1에서는 요일 정보만 사용하여 0시 교통량을 예측했기 때문에 성능이 낮았습니다. 0시부터 23시까지의 모든 시간대 교통량 정보를 추가 피처로 도입함으로써, 모델이 시간대 간의 상관관계를 학습하게 할 수 있습니다. 예를 들어, 23시의 교통량은 0시의 교통량에 영향을 미칠 수 있으며, 이는 모델이 학습할 수 있는 중요한 패턴 중 하나입니다.  2. 모든 시간대 정보를 피처로 설정  0시 교통량을 타깃 값으로 설정하고, 1시부터 23시까지의 시간대 교통량을 모두 피처로 사용하여 모델을 학습시킵니다. 이렇게 하면 모델이 다양한 시간대 데이터를 학습하여 0시 교통량을 더 정확하게 예측할 수 있습니다.  3. 학습 데이터와 테스트 데이터 분할  train\_test\_split 함수를 사용하여 학습용과 테스트용 데이터를 분리합니다. 테스트 데이터로 모델 성능을 평가합니다. 모델 학습 및 평가 LinearRegression 클래스를 사용하여 선형 회귀 모델을 학습시키고, 모델 성능을 R² 점수로 평가합니다.  4. 모델 성능 향상 분석  실행 결과 R² 값이 0.49845696871884915로 나타났습니다. 이는 과제 1의 R² 값(0.28356160505  73254)보다 성능이 크게 향상된 것입니다. 그 이유는, 단일 피처(요일 정보)만 사용했던 과제 1과 달리, 0시부터 23시까지의 모든 시간대 교통량 데이터를 추가로 제공했기 때문에, 모델이 시간대별 패턴을 학습할 수 있었기 때문입니다.  모든 시간대를 추가로 사용한 이유는, 교통량이 시간대 간에 상관관계를 가질 수 있기 때문입니다. 예를 들어, 23시의 교통량이 0시의 교통량에 영향을 미칠 수 있다는 점을 모델이 학습하면서 성능이 향상되었습니다. 다양한 시간대의 데이터를 활용하면, 교통량의 변화 양상이나 패턴을 더 잘 파악할 수 있기 때문에 성능이 상승할 가능성이 큽니다.  결론: 여러 시간대의 데이터를 함께 사용함으로써 모델이 교통량 패턴을 더 잘 학습할 수 있었고, 그 결과로 성능이 개선되었습니다 |