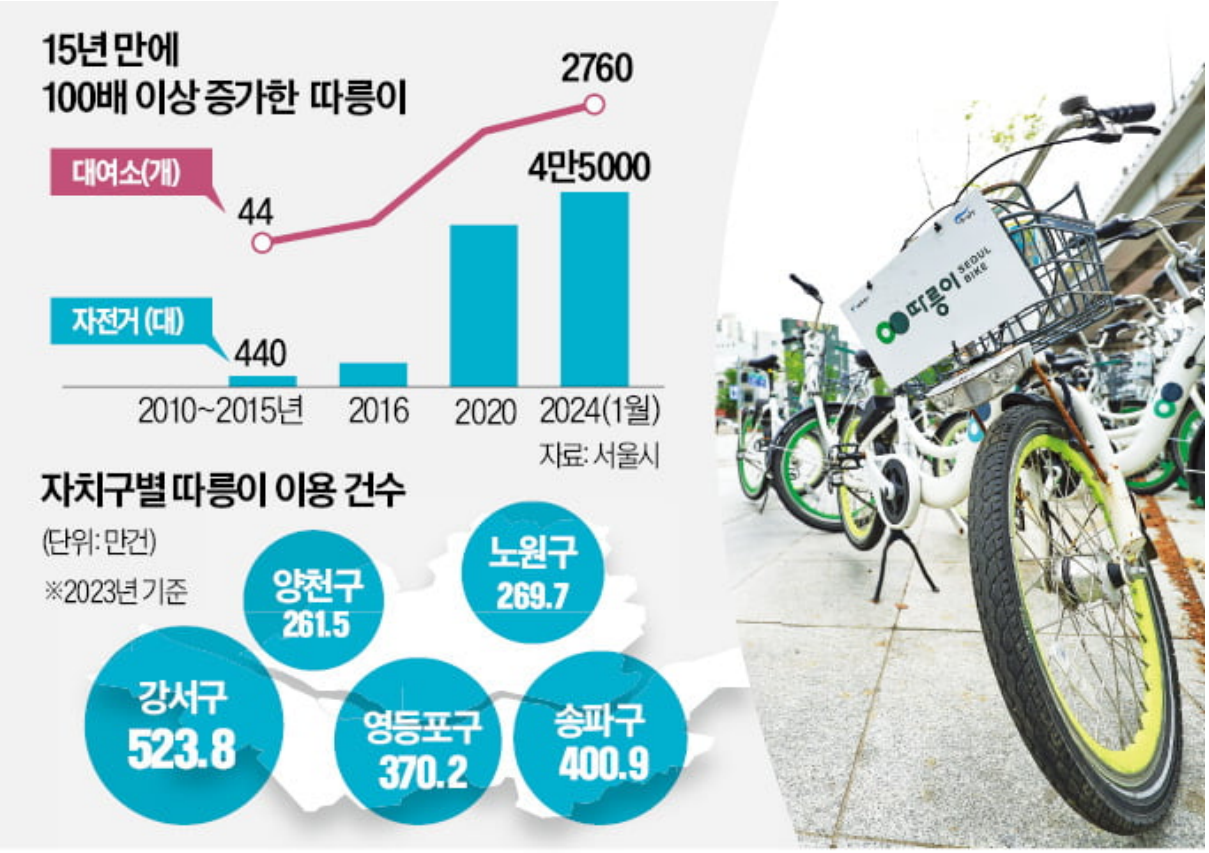
자전거 대여 시스템 데이터 분석

백예나

목차

1 서론

서울시 공공자전거 '따릉이'는 2015년 9월 서비스를 시작한 이후 꾸준히 이용자가 증가하였으며, 2024년 6월 기준 누적 대여 건수 2억 건을 돌파했다. '따릉이'는 시민들이 일상적으로 사용하는 이동 수단으로 자리 잡았고, 출퇴근이나 여가 등 다양한 목적으로 활용되고 있다.



1.1 프로젝트 배경 및 목적

하지만 막상 '따릉이'를 이용하려 했을 때, 대여소에 자전거가 없거나 원하는 장소에 배치되지 않아 이용하지 못한 경험을 한 시민도 적지 않을 것이다. 자전거를 필요한 시간과 장소에 적절히 배치하는 것은 운영 효율성과 사용자 만족도를 높이는 핵심 과제이다. 수요를 정확히 예측한다면, 자전거 부족이나 과잉 배치 같은 문제를 사전에 방지 할 수 있을 것이다.

본 프로젝트에서는 과거의 데이터를 기반으로 자전거 수요를 예측하는 모델을 구축하고자 한다. 머신러닝을 활용해 다양한 변수 간의 관계를 분석하고 RMSLE(Root Mean Squared Logarithmic Error)를 최소화하는 모델을 개발함으로써, 보다 정밀한 수요 예측이 가능하도록 하는 것을 목표로 한다.

1.2 문제 정의

현재 '따릉이' 서비스는 대여소별 시간대별 자전거 수요 변동이 크지만, 이를 실시간으로 반영한 배치가 어렵다. 그렇기 때문에 특정 시간대나 장소에서 자전거가 부족해 시민들이 원하는 서비스를 받지 못하거나, 반대로 불필요한 과잉 배치로 운영 비용이 증가하는 문제가 발생하고 있다. 따라서, 자전거 수요 예측의 정확도를 높여, 수요에 맞는 적절한 배치 전략 수립이 필요하다.

2 데이터 소개 및 전처리

2.1 데이터 구성

본 프로젝트에서 사용한 데이터는 프로젝트 수행을 위해 제공된 데이터로, 학습용(train.csv)과 평가용(test.csv) 데이터로 구분되어 있다. 해당 데이터는 2011년 1월1일00시부터 2012년 12월 19일 23시까지의 자전거 대여 기록을 포함한다.

학습용 데이터에는 casual, registered, count가 포함이 되어 있으며, count는 예측 대상인 종속 변수이다. 평가용 이 세칼럼은 포함되어 있지 않다.

주요 변수는 다음과 같다.

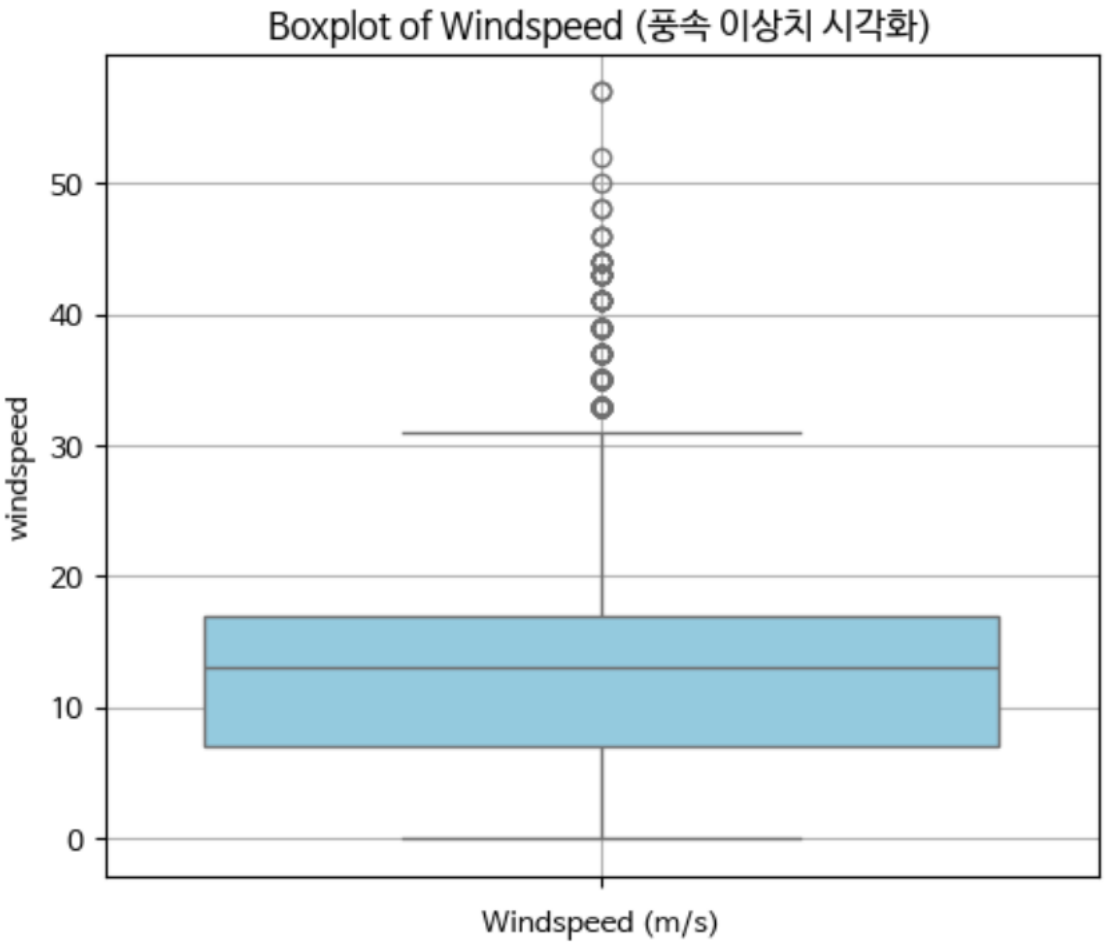
변수명	설명	데이터 타입 값 종류
datetime	대여 기록의 날짜와 시간	datetime
season	계절(1:겨울, 2:봄, 3:여름, 4:겨울)	category(1~4)
holiday	공휴일 여부(0: 평일, 1: 공휴일)	category(0,1)
workingday	근무일 여부(0: 주말/공휴일, 1: 근무일)	category(0,1)
weather	날씨(1: 맑음, 2: 구름/안개, 3: 비/눈, 4: 폭우/폭설)	category(1~4)
temp	실측 온도(섭씨)	float
atemp	체감 온도(섭씨)	float
humidity	습도 (%)	int
windspeed	풍속 (m/s)	float
casual	등록되지 않은 사용자의 대여 수	int
registered	등록된 사용자의 대여 수	int
count	총 대여 수(casual+registered)	int

2.1 결측값 및 이상치 처리

2.1.1 결측값 처리

학습용 및 테스트 데이터 모두에 대해 결측값 여부를 확인한 결과, 모든 변수에 결측값이 존재하지 않음을 확인하였다. 따라서 별도의 결측값 처리는 수행하지 않았다.

2.1.2 이상치 처리



수치형 변수(temp, atemp, humidity,, casual, registered, count)에 대해 박스플롯 및 기초 통계량을 활용해 이상치 여부를 확인하였다. 일부 극단적인 값들이 존재하였으나, 이는 실제 측정된 자전거 대여 패턴의 특수한 경우로 판단되어 이상치로 간주하지 않고 그대로 유지하였다.

또한 season, weather, holiday 등의 범주형 변수에 따라 자전거 대여 수요가 급격히 변화하는 경우가 있지만, 이는 현실적인 수요 변동으로 간주해 이상치로 처리하지 않았다.

풍속(windspeed)의 경우, 유난히 큰 값들이 일부 존재하여 이를 사분위수 기준으로 이상치 여부를 판단하였다.

IQR 결과, 풍속이 31.99를 초과하는 데이터 227건이 이상치로 확인되었고, 전체 데이터의 약 2.09%를 차지했다. 이러한 값들은 실제 상황에서도 드물게 발생할 수 있는 극단적

인 경우이며, 예측 모델에 영향을 줄 수 있어 해당 값들을 제거하고 분석을 진행하였다.

2.3 파생 변수 생성 및 인코딩

자전거 대여 수요는 시간적 패턴이나 요일, 날씨 등의 외부 요인에 민감하게 반응하므로, 이러한 패턴을 효과적으로 반영하기 위해 다양한 파생 변수를 생성하고 범주형 변수 인코딩을 수행하였다.

먼저, datetime 변수로부터 월(month), 요일(weekday), 시간(hour) 정보를 추출해 새로운 변수로 추가하였다. 이는 자전거 대여량이 요일이나 시간대에 따라 명확한 차이를 보이는 경향이 있었다. 해당 시간 관련 변수들은 주기성을 갖는 특성이 있기에, 사인(Sine)과 코사인(Cosine) 변환을 적용해 모델이 시간의 순환성을 잘 학습할 수 있도록 구성하였다.

또한, 시간대를 좀 더 직관적으로 표현하기 위해 시간(hour)을 기준으로 출근 시간대(6-9시), 주간(9-17시), 퇴근 시간대(17-20시), 야간(그외)의 네 구간으로 나눈 time_of_day 파생 변수를 생성하였다. 출퇴근 시간에 수요가 급증하는 패턴을 고렇나 처리이며, one-hot 인코딩을 통해 모델이 시간대별 특성을 잘 인식할 수 있도록 하였다.

추가적으로, 주말 여부를 나타내는 이진 변수 is_weekend도 생성하여 휴일과의 차이를 반영하였다.

season, weather, holiday, workingday와 같은 범주형 변수는 모두 One-hot 인코딩을 적용하되, 다중 공선성을 방지하기 위해 drop_first=True 옵션을 사용하였다.

또한, 예측 대상인 count 변수는 0 이상의 정수값을 가지며 분포가 비대칭적이었기에, 모델의 안정성과 성능 향상을 위해 로그 변환을 적용한 count_log = log1p(count) 파생 변수를 생성하여 타겟으로 사용하였다. 이는 예측값이 음수로 나오는 현상을 방지하고, 데이터 분포를 정규 분포에 가깝게 만들어 모델이 극단값에 덜 민감하도록 하기 위함이다.

반면, casual과 registered 변수는 count를 구성하는 내부 요소로서 직접적인 예측에 포함시키는 것이 부적절하다고 판단되어 모델 학습에서 제외하였다.

또한, 온도 관련 변수인 temp와 atemp는 높은 상관관계를 가지므로 다중공선성 문제를 방지하기 위해 atemp만 남기고 temp는 제거하였다.

3 모델링 및 평가

3.1 사용모델

본 분석에서는 자전거 대여 수요를 예측하기 위해 회귀 기반 모델인 선형 회귀(Linear Regression)와 릿지 회귀(Ridge Regression)를 사용하였다.

Linear Regression은 변수 간 선형 관계를 기반으로 예측하는 가장 기본적인 모델이며, Ridge Regression은 L2 정규화를 통해 과적합을 방지하고 일반화 성능을 향상시킬 수 있는 모델이다.

두 모델 모두 타깃변수로는 로그 변환된 count_log를 사용하였고, 모델 성능은 검증 데이터에 대한 RMSL와 5-Flod 교차 검증을 통해 평가하였다.

Ridge 모델의 경우, 하이퍼파리미터 alpha에 대해 그리드 서치를 적용해 최적의 값을 찾았다.

Linear Regression과 Ridge Regression 두 모델 모두 로그 변환된 타깃 변수(count_log)를 예측 대상으로 사용하였으며, 학습 및 교차 검증 과정을 통해 성능을 비교하였다.

3.2 평가지표

예측 결과는 학습 시 로그 변환된 count_log를 사용하였기 때문에, 예측값은 np.expm1()을 통해 원래 스케일의 count 값으로 복원한 후 평가하였다. 평가지표로는 RMSLE(Root Mean Squared Logarithmic Error)를 사용하였는데, 이는 실제값과 예측값이 모두 양수이고, 상대적인 오차를 중요시하는 문제에서 적합하다. 특히 RMSLE는 극단적인 값보다 전체적인 비율 차이에 민감해, 본 과제와 같은 자전거 수요 예측에 효과적인 지표라 생각한다.

3.3 모델별 성능 비교

자전거 대여 수요 예측을 위해 선형 회귀(Linear Regression)와 릿지 회귀(Ridge Regression)를 적용하고, 로그 변환된 타깃 변수 count_log를 예측한 후, 예측 값을 np.expm1()로 복원해 RMSLE로 성능을 평가하였다.

두 모델 모두 5-Fold 교차 검증을 통해 안정적인 평가를 수행하였으며, 각 모델의 성능은 다음과 같다

모델	RMSLE	5-FOLD RMSLE	5-Fold 평균 RMSLE
Linear Regression	0.7553	[0.7553, 0.7617, 0.7517, 0.7569, 0.7838]	0.7619
Ridge Regression	0.7553 최적 alpha: 1	[0.7553, 0.7618, 0.7517, 0.7569, 0.7839]	0.7619

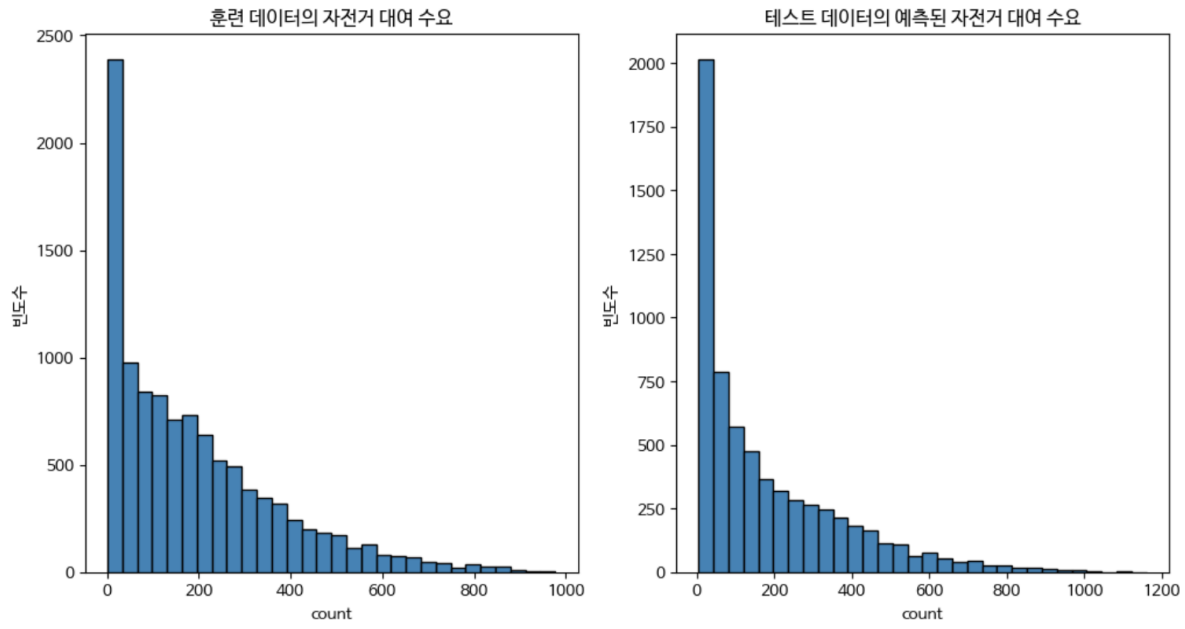
두 모델 모두 유사한 성능을 보였으며, Ridge 회귀는 L2 정규화를 통해 약간 더 안정적인 결과를 보여주었다. 따라서 단순한 선형 회귀로도 충분한 예측 성능을 확보할 수 있었으며, Ridge 회귀는 과적합을 방지하고자 할 때 보다 안정적인 대안이 될 수 있다.

3.4 테스트 데이터 예측 결과

최종적으로 학습된 Ridge 모델을 사용해 테스트 데이터에 대한 자전거 대여 수요를 예측하였다.

테스트 데이터는 학습 데이터와 동일한 방식으로 전처리를 수행한 후, 예측 결과는 로그 변환된 스케일로 출력되므로 np.exp1() 함수를 통해 원래 스케일의 count 값으로 복원하였다.

예측 결과가 실제로 타당한 분포를 보이는지 확인하기 위해, 예측된 대여수(count)에 대한 히스토그램을 시각화 하였다. 아래 그림은 훈련 데이터의 실제 자전거 대여 수요 분포(좌측)와 테스트 데이터에 대해 예측한 수요 분포(우측)을 비교한 것이다.



두 분포는 전제적으로 유사한 형태를 보여, 이를 통해 모델이 테스트 데이터에서도 전반적으로 합리적인 예측을 수행했음을 시각적으로 확인할 수 있었다. 이러한 확인 과정은 모델이 학습 데이터에만 과도하게 적합 된 것이 아니라, 일반적인 수요 패턴을 잘 학습했다는 점에서 의미 있는 확인 과정이라 할 수 있다.

4 결론 및 향후 방향

4.1 주요 인사이트 정리

자전거 대여 시스템의 수요 예측을 통해 운영 효율성을 높이고 사용자 만족도를 향상 시키기 위한 분석을 수행하였다. 여러 가설을 바탕으로 데이터를 탐색한 결과, 다음과 같은 주요 인사이트를 도출할 수 있었다.

4.1.1 출퇴근 시간대 집중 대여

정기 이용자인 등록자는 평일 출근(7~8)과 퇴근(17~20)에 대여 수가 뚜렷하게 증가하는 패턴을 보였다. 이는 등록자가 주로 직장인이나 학생으로 출퇴근 목적으로 자전거를 이용함을 시사한다. 반면 주말에는 등록자들의 대여가 비교적 분산되어 여가용으로 활용되는 경향이 관찰되었다.

비정기 이용자인 비등록자의 경우, 평일 낮 시간대에 점진적으로 대여가 증가하며, 주말에는 전 시간대에 걸쳐 높은 대여량을 나타냈다. 이는 비등록자가 주로 관광객이나 일시적 이용자로서 날씨가 좋고 활동하기 편한 봄과 여름철 주말에 여가 목적으로 자전거를 이용하는 경향이 크다는 것을 의미한다.

4.1.2 여름철 대여량 증가

계절별로는 여름철에 대여량이 가장 높았고 겨울철에 가장 낮았다. 정기 이용자는 계절 변화에 비교적 둔감한 반면, 비등록자는 봄과 여름에 대여가 집중되는 특성을 보였다. 이는 기온과 날씨에 따른 여가 활동 증가와 밀접한 관련이 있는 것으로 판단된다.

4.1.3 날씨가 맑고 쾌적할 때 대여량

날씨가 맑고 쾌적할 때 대여량이 증가하며, 비, 눈, 폭우/폭설 등의 악천후에는 대여가 급감하는 것으로 나타났다. 특히 비등록자의 대여량 감소폭이 더 크게 나타나 날씨에 민감한 여가 이용자임을 확인하였다.

분석 결과를 바탕으로 자전거 대여 시스템 운영의 효율성과 사용자 만족도를 높이기 위한 다음과 같은 전략을 제안한다.

4.2 자전거 대여 효율화를 위한 운영 및 마케팅 전략

4.2.1 출퇴근 시간대 집중 배치 전략

평일 출퇴근 시간대에는 주요 대여소에 정기 이용자를 위한 자전거를 집중 배치하여 피크 수요를 효과적으로 충족시켜야 한다. 반면, 주말과 낮 시간대에는 관광지 및 공원 등 여가 이용이 많은 지역에 정기 및 비정기 이용자 모두를 위한 자전거를 확대 배치하는 것이 필요하다.

4.2.2 기상 및 계절에 따른 탄력적 운영

기상 상황과 계절 변화를 고려한 탄력적인 운영이 중요하다. 맑고 쾌적한 날씨에는 자전거 공급량을 확대하고 프로모션을 진행하여 수요를 적극 유도하되, 악천후 시에는 운영 시간 조정과 안전 안내를 강화해 이용자의 안전을 확보해야 한다. 또한 겨울철에는 대여량이 감소하는 점을 고려해 유지보수 중심의 운영 계획을 수립하는 게 좋을 것 같다.

4.2.3 기상 및 계절에 따른 탄력적 운영

실시간 기상 데이터와 과거 대여 패턴을 결합한 수요 예측 모델을 활용해 자전거 재고와 배치 계획을 최적화하고, 출퇴근 시간대 피크를 대비한 자전거 회수와 재배치 자동화 시스템 도입이 필요할 것 같다.

4.2.4 이용자 유형별 맞춤형 마케팅 전략

정기 이용자와 비정기 이용자 각각의 특성을 반영한 맞춤형 마케팅 전략이 필요하다. 정기 이용자에게는 구독 할인 등의 혜택을 제공하고, 비등록자 대상으로는 계절별 이벤트 또는 추천 경로 안내 등을 통해 이용 편의를 향상시키는 방안을 마련해야 한다.

이와 같은 전략적 접근을 통해 자전거 대여 시스템의 효율성을 극대화하고, 다양한 이용자들의 만족도를 높일 수 있을 것으로 기대한다.

4.3 개선점 및 향후 보완 아이디어

이번 분석은 자전거 대여 데이터를 바탕으로 계절, 시간대, 날씨 등의 주요 변수와 대여량 간의 관계를 탐색하고 수요 예측 모델을 구축하는 데 중점을 두었다.

4.3.1 선형 모델의 구조적 한계

본 프로젝트에서는 선형 회귀 계열 모델인 Linear 및 Ridge 회귀 모델을 사용하였다. 해당 모델들의 RMSLE 점수는 약 0.75 수준으로, 다양한 파생변수를 추가하여도 성능 개선에 한계가 있었다.

이는 온도와 체감온도, 출퇴근 시간과 평일 여부 등은 단순한 선형 회귀식으로 설명되기 어렵기에, 자전거 대여량과 독립 변수들 사이에 비선형 관계 및 변수 간 상호작용이 존재함을 시사한다.

4.3.2 비선형 모델 도입 필요

선형 모델의 한계를 극복하기 위해, Random Forest와 같은 비선형 앙상블 모델을 도입하는 것이 필요하다.

Random Forest는 변수 간의 비선형 관계 및 상호작용을 자동으로 학습하며, 이상치 및 다중공선성에 강하다. 또한 파생변수의 효과를 더 잘 반영할 수 있어, Random Forest가 적합할 것으로 보인다.

향후 Random Forest 기반의 모델을 적용하고, 추가로 하이퍼파라미터 튜닝과 변수 중요도 분석을 진행하면 예측 성능을 더욱 향상시킬 수 있을 것이다.

5 결론

이번 분석에서는 자전거 대여 데이터를 기반으로 수요 예측 모델을 구축하였다. 선형 회귀 계열 모델을 적용한 결과, 일정 수준의 설명력은 있었지만 비선형적 관계를 충분히 반영하지 못하는 한계를 보였다.

이에 따라, 향후 분석에서는 Random Forest와 같은 비선형 모델을 도입해 변수 간의 복잡한 상호작용을 보다 효과적으로 반영하고, 예측 정확도를 향상시키는 방향으로 발전시킬 수 있을 것으로 기대된다.