자전거 수요량 예측 모델링

자전거 대여 시스템의 효율성을 높이기 위한 수요 예측 모델링

AI 3기 이영호

목차

배경 및 목표

1. 자전거 대여 패턴 분석
   - 배치 및 운영 전략 최적화
2. 정확한 수요 예측
   - 시스템 효율성 향상
3. 사용자 만족도 증가
   - RMSLE 최소화 목표

- 분석 범위
  - 2011년~2012년 1시간 단위 자전거 대여 데이터 활용

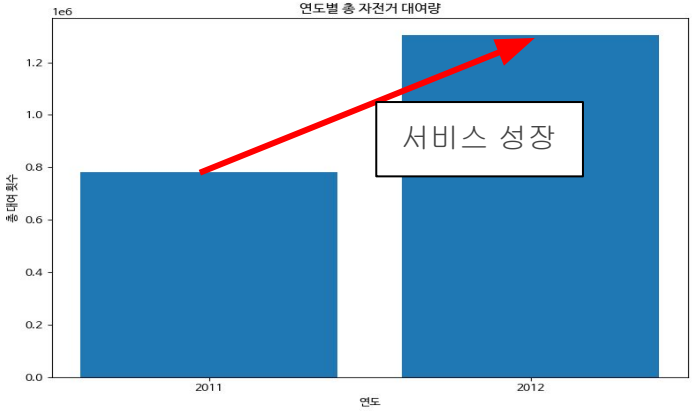
데이터 구성

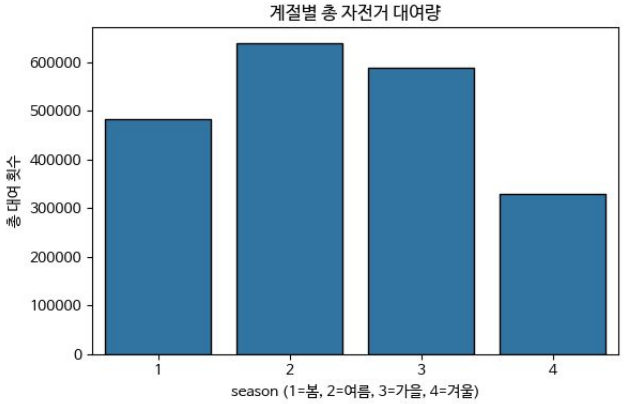
컬럼명	데이터 타입	설명
datetime	datetime	자전거 대여 기록의 날짜 및 시간. 예시: 2011-01-01 00:00:00
season	int	계절 (1: 봄, 2: 여름, 3: 가을, 4: 겨울)
holiday	int	공휴일 여부 (0: 평일, 1: 공휴일)
workingday	int	근무일 여부 (0: 주말/공휴일, 1: 근무일)
weather	int	날씨 상황 (1: 맑음, 2: 구름낌/안개, 3: 약간의 비/눈, 4: 폭우/폭설)
temp	float	실측 온도 (섭씨)
atemp	float	체감 온도 (섭씨)
humidity	int	습도 (%)
windspeed	float	풍속 (m/s)
casual	int	등록되지 않은 사용자의 대여 수
registered	int	등록된 사용자의 대여 수
count	int	총 대여 수 (종속 변수)

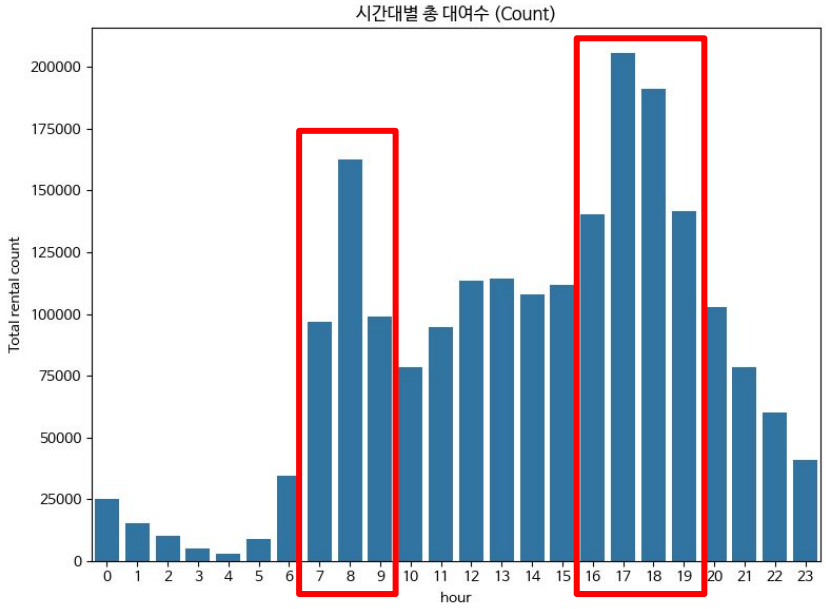
훈련용 데이터는 매월 1~19일
데이터로 종속변수(count)를 포함

평가용 데이터는 매월 20~31일
데이터로 종속변수 없음.

EDA - 시간에 따른 대여량 변화 추이

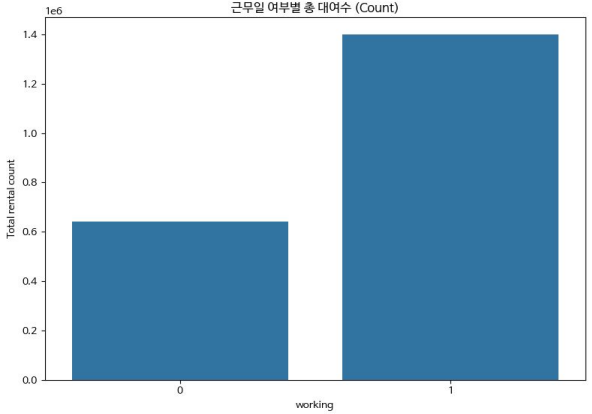


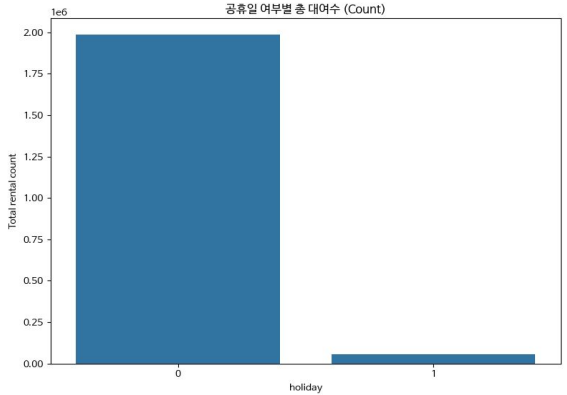


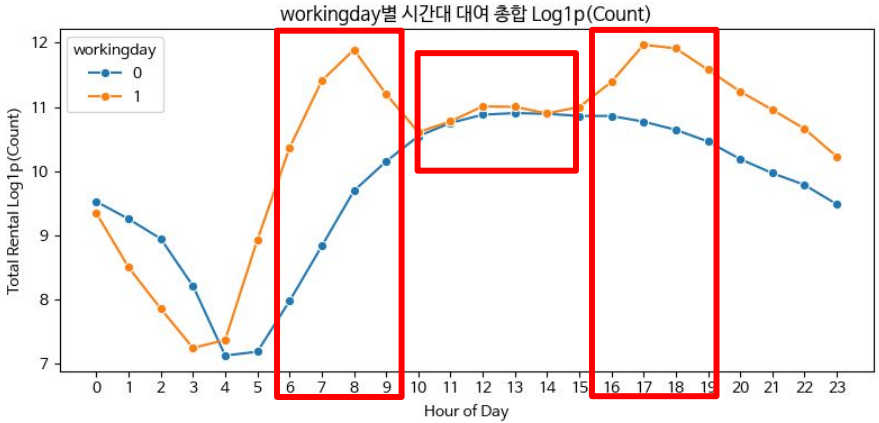


출퇴근 시간대 피크

EDA - 근무일 / 공휴일 별 대여량 분포



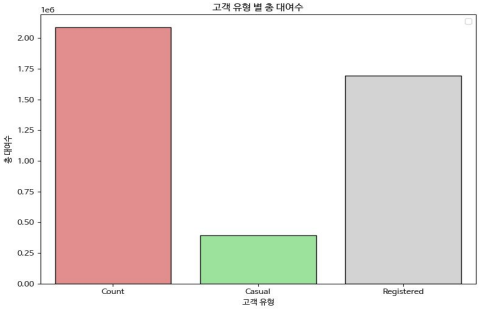


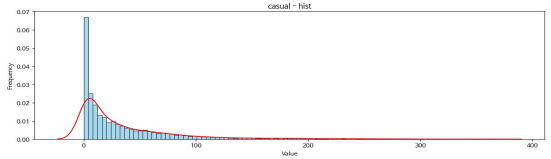


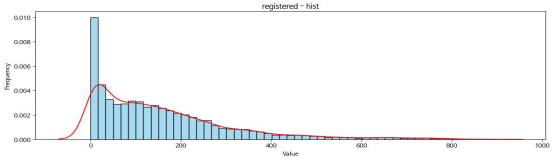
07~09시 / 16~19시 사용량 증가
11~13시 사용량 증가

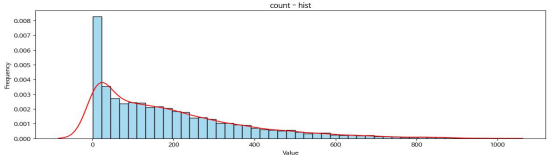
→ 근무일엔 출/퇴근 시간대에 많은 이용
→ 휴일엔 낮시간에 많은 이용

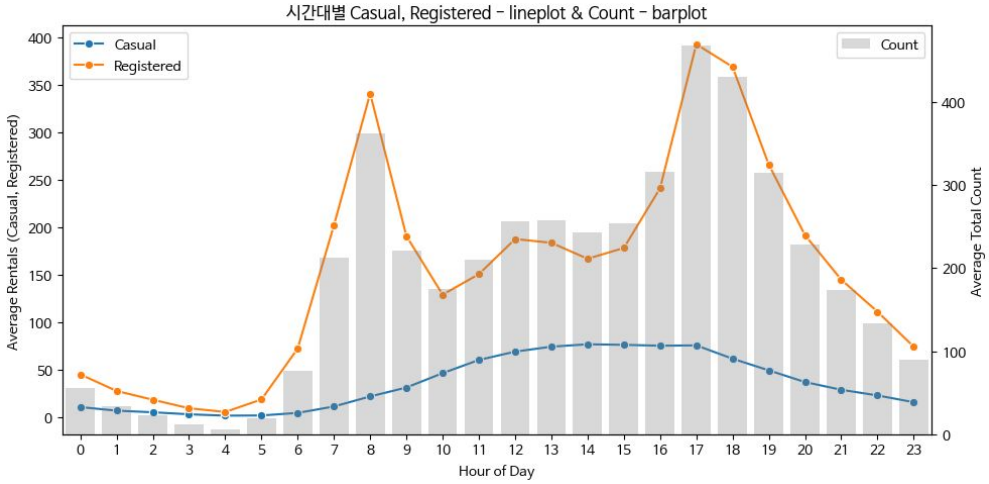
EDA - 고객 유형 별 시간대별 대여량 추이











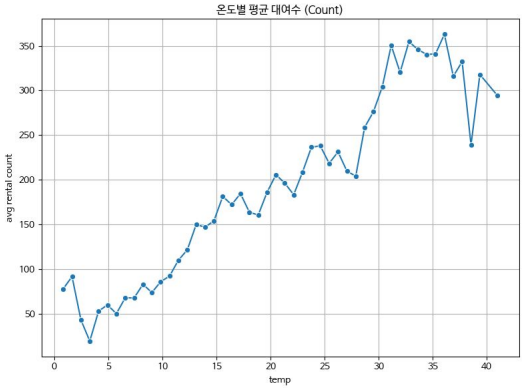
회원 고객이 다수

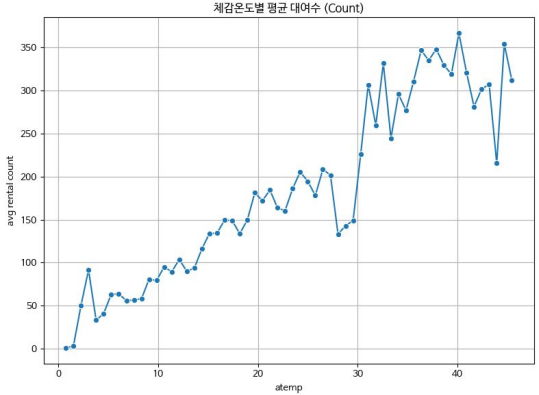
고객 유형

회원고객 : 출퇴근시간대 많은 이용, 낮시간도 이용

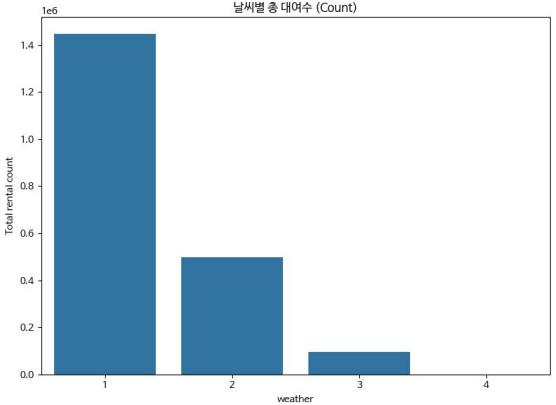
비회원고객 : 낮시간 많은 이용

EDA - 온도 / 날씨 별 대여량 추이

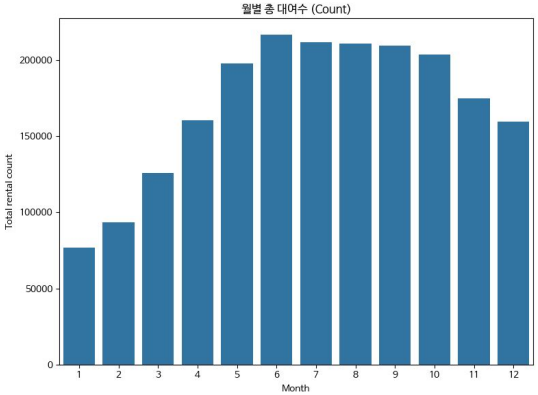




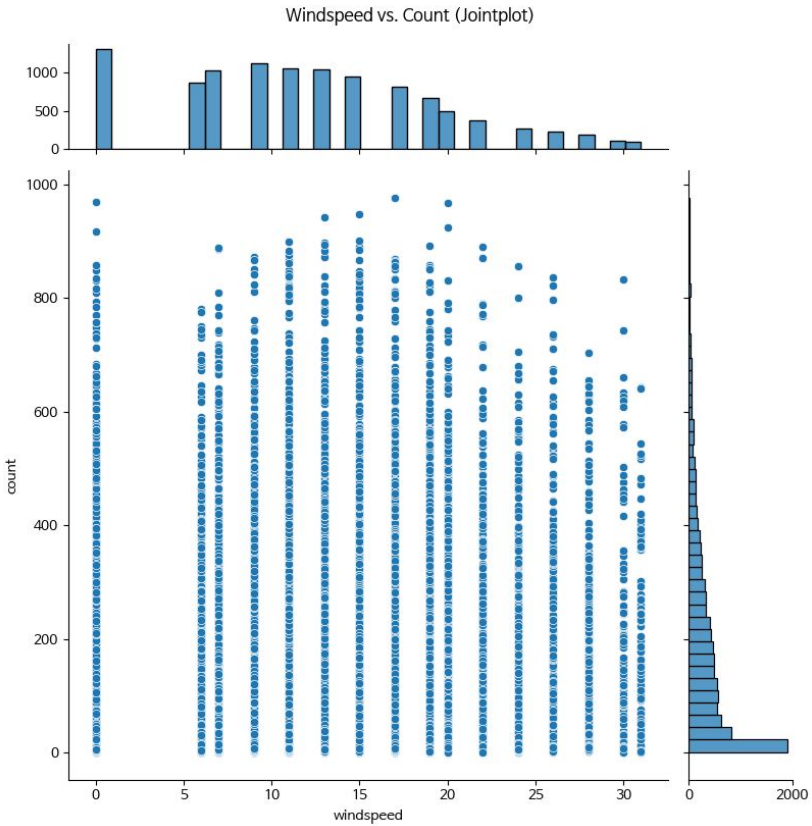
날씨가 맑을수록 대여량이 많다.
쾌적한 온도에서 대여량이 많다.



(1: 맑음, 2: 구름낌/안개, 3: 약간의 비/눈, 4: 폭우/폭설)



EDA - 풍속별 대여량 추이

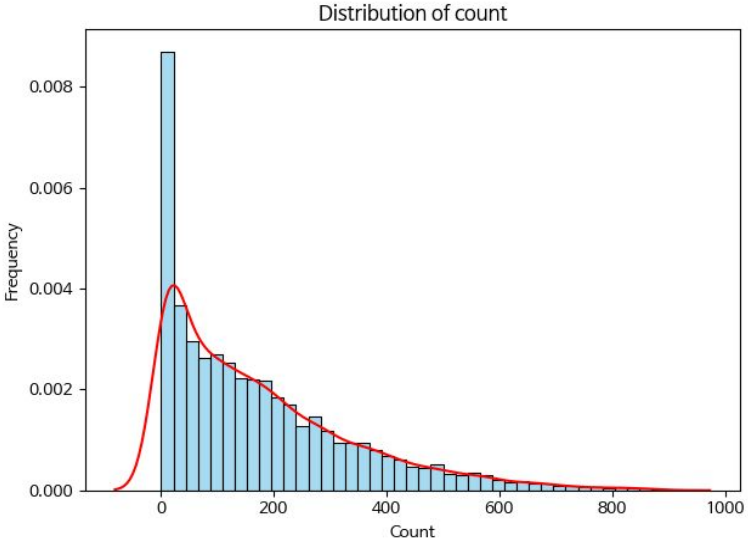


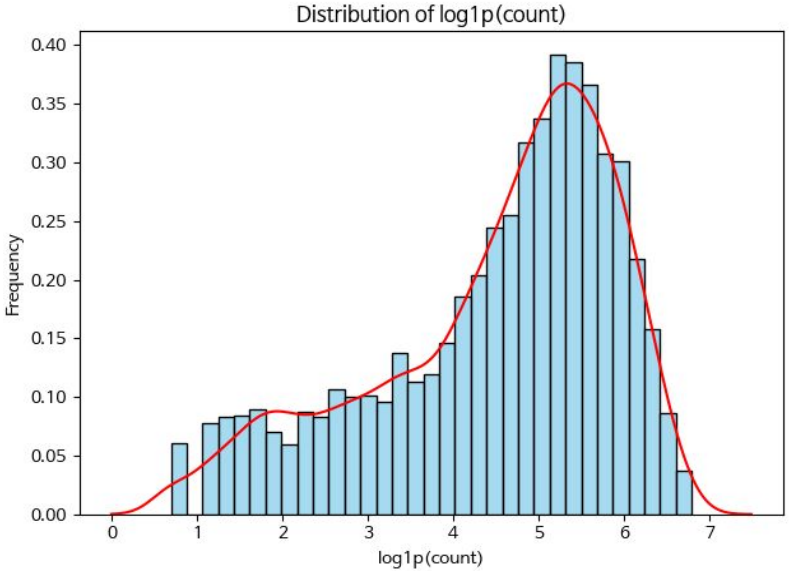
현재의 보퍼트 풍력 계급 [편집]

계급	풍속				명칭	파고		육상 상태	해상 상태	사진
	m/s	km/h	kt	mph		m	ft			
0	<0.3	<1	<1	<1	고요	0	0	연기가 수직으로 올라간다.	해면이 거울 같이 반사될 정도로 고요하다.	
1	0.3~1.5	1~5	1~2	1~3	실바람	0.1	0.33	풍향은 연기가 날아가는 것으로 알 수 있으나, 풍향계는 잘 움직이지 않는다.	물결이 생선 비늘 같이 작고, 물거품이 없다.	
2	1.5~3.3	6~11	4~6	3~7	남실바람	0.2	0.66	바람이 피부에 느껴진다. 나뭇잎이 흔들리며, 풍향계가 움직이기 시작한다.	물결이 작고, 파도의 마루 부분이 부서지지 않고 모양이 뚜렷하다.	
3	3.3~5.5	12~19	7~10	8~12	산들바람	0.6	2	나뭇잎과 작은 가지가 끊임없이 흔들리고, 깃발이 가볍게 날린다.	물결이 커지고, 파도의 마루가 부서져서 물거품이 생겨 흰 파도가 간간이 보인다.	
4	5.5~8.0	20~28	11~15	13~17	건들바람	1	3.3	먼지가 일고 종이조각이 날리며, 작은 가지가 흔들린다.	파도가 잘고, 파장이 길어지며 흰 파도가 많이 보이기 시작한다.	
5	8.0~10.8	29~38	16~20	18~24	흔들바람	2	6.6	잎이 무성한 작은 나무 전체가 흔들리고, 호수에 물결이 일어난다.	파도가 조금 높아지고, 물거품이 생기기 시작한다.	
6	10.8~13.9	39~49	21~26	25~30	된바람	3	9.9	큰 나뭇가지가 흔들리고, 전선이 울리며 우산을 사용하기 어렵다.	파도가 높아지기 시작하고, 물거품이 광범위해지며 물보라가 생긴다.	
7	13.9~17.2	50~61	27~33	31~38	센바람	4	13.1	나무 전체가 흔들리며, 바람을 안고서 걷기 곤란하다.	파도가 높아지고, 파도가 서로 부서져서 물거품이 생겨 줄을 이루며 바람에 의해 날린다.	
8	17.2~20.7	62~74	34~40	39~46	큰바람	5.5	18	작은 나뭇가지가 꺾이며, 바람을 안고서 걸을 수 없다.	파도가 제법 높고, 파장이 더 길고 마루의 끝이 거꾸로 뒤다. 물거품이 강풍에 날린다.	
9	20.7~24.5	75~88	41~47	47~54	큰센바람	7	23	큰 나뭇가지가 꺾이고, 가옥에 다소 피해가 생긴다. 굴뚝이 넘어지고 기와가 벗겨진다.	파도가 높고, 물거품이 바람에 따라 짙은 줄무늬를 띤다. 마루가 흩어져 말리고 물보라 때문에 시정이 나빠진다.	
10	24.5~28.4	89~102	48~55	55~63	노대바람	9	29.5	나무가 뿌리째 뽑히고, 가옥에 큰 피해가 일어난다. 내륙 지방에서는 보기 드문 현상이다.	파도가 앞으로 긴 마루로 되어 몹시 높고, 물거품이 큰 덩어리가 되어 강풍에 날린다. 파도가 심하게 부서지고 시정이 나쁘다.	
11	28.4~32.6	103~117	56~63	64~72	왕바람	11.5	37.7	광범위한 피해가 생긴다.	파도가 대단히 높고, 주위의 배는 파도에 가려 볼 수 없고 길게 줄지은 물거품들이 바다를 덮는다. 시정이 극히 나쁘다.	
12	≥32.6	≥118	≥64	≥73	싹쓸바람	≥14	≥46	매우 광범위한 피해가 생긴다.	파도가 매우 높고, 바다는 물거품과 물보라로 가득 차 바로 한치 앞도 분간하기 어려울 정도다.	

풍속이 낮을수록 대여 건수가 높음.
풍속이 높아질수록 대여 건수가 감소.
→ 바람이 강할수록 자전거를 덜 빌린다는
뚜렷한 음의 상관관계

예측 target (count)





원본 count

- 왼쪽으로 몰려있는 오른쪽 치우침(right skewed)
  대부분의 시간대/상황엔 대여량이 적은 편이지만,
  일부 시간대엔 매우 높은 대여량이 발생
- 일부 모델의 부정적인 영향을 줄 수 있을

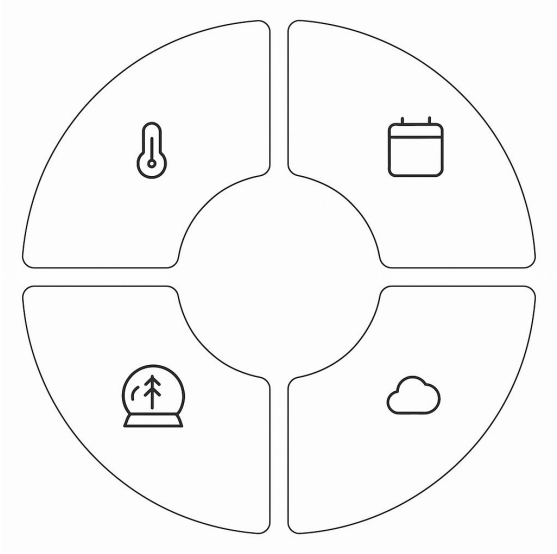
Log1p count

- 비대칭(skewed) 분포 완화
- 분산 안정화 : 작은값↔큰값 분산 차이 감소
- Outlier 영향 완충
- 예측 성능과 일반화 능력 개선 가능성 증가

상관관계 인사이트

온도
적정 온도에서 이용률 증가

계절
봄/ 여름 / 가을 이용률 다수



요일 / 시간
출퇴근 시간과 주말 패턴 상이

날씨
날씨가 맑을수록 이용률 증가

- 연도별 성장세
- 계절, 평일/주말, 시간대 패턴이 두드러짐
- 환경적 요소(날씨, 온도)에 영향을 많이 받음

모델링 준비

- Feature Engineering
  - 'datetime' → 'year', 'month', 'day', 'hour'
  - 'windspeed' → 'bf_0' ~ 'bf_12' (※. 보퍼트 풍력 계급으로 나눔)
  - 'season' 재 매칭 (3~5월 : 봄[1], 6~8월 : 여름[2], 9~11월 : 가을[3], 12~2월 : 겨울[4])
- 범주형 변수
  - 'season', 'holiday', 'workingday', 'weather', 'year', 'month', 'hour'
  - One-Hot Encoding 진행
- 수치형 변수
  - 'temp', 'atemp', 'humidity'
  - 정규화 (Z-score)
- 독립변수
  - 'count', 'datetime', 'casual', 'registered', 'windspeed' 제외
- 종속변수
  - 'count' log1p 적용
- 데이터 분할
  - train 80%, test 20%

모델 선택

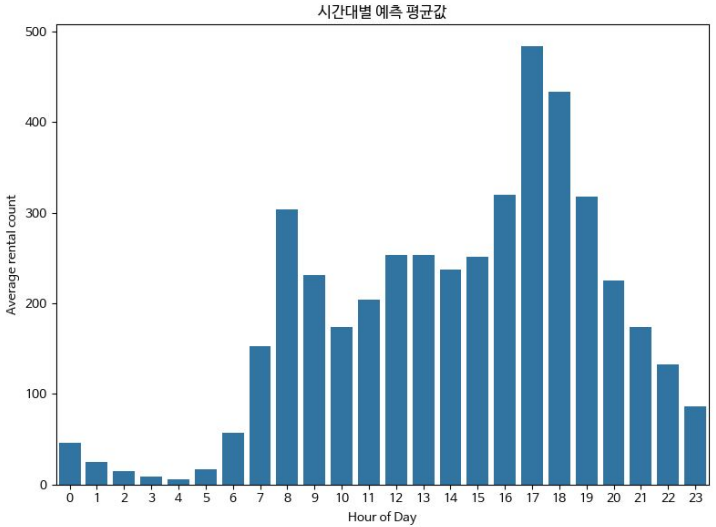
- 선형 회귀
  - 가장 기본적인 회귀모델
  - 피처들과 타겟 변수 간의 선형 관계를 가정
  - 해석이 용이, 베이스라인 성능 파악에 좋음
- 릿지 회귀
  - 선형회귀에 L2 규제를 추가
  - 피처들의 가중치가 너무 커지는 것을 방지
  - 다중공선성이 있을 때 안정적
- 라쏘 회귀
  - 선형회귀에 L1 규제를 추가
  - 불필요한 피처들의 가중치를 0으로 만듦

모델 성능 검증 및 안정화

- K-fold
  - 모델의 일반화 성능 검증 및 안정화
  -
- GridSearchCV
  - 하이퍼파라미터를 최적화하여 모델의 예측 정확도 향상

최종 모델 예측 성능

- 최종 모델
  - Ridge
- 최종 모델 예측 성능
  - 핵심 지표 : RMSLE
  - 달성 RMSLE 값 : 0.57



결론 및 제언

- EDA를 통해 주요 피처의 패턴이 모델의 성능에 영향
  - 외부적 요소 (계절, 날씨, 온도, 습도, 풍속)
    - 쾌적한 상태에서 자전거를 이용하려는 성향이 있음
  - 시간적 요소 (주중/주말, 출/퇴근)
    - 등록회원은 주로 주중에는 출/퇴근시간대와 주말엔 낮시간대에 이용
    - 비등록회원은 낮시간대에 이용
- 제언
  - 운영 측면
    - 출퇴근 시간대(평일 07~09시, 17~19시) 주말 낮(10~15시)에 자전거가 많은 대여소에서 대여하거나, 적게 있는 대여소에서 반납하면 포인트 적립 → 자전거를 운반하는 비용 감소
    - 계절, 날씨 등을 고려해 운영량 조절 및 전략 필요
    - GPS 등 위치 정보 취득
      - 상업, 주거 지역, 교통 환승 구역 등에서 대여 수요를 파악 가능 → 더욱 다양한 수요 예측 가능
    - 관리자 웹, 이용자 앱 활용하여 UX 적극 활용

Q&A

감사합니다