Bike Sharing Demand

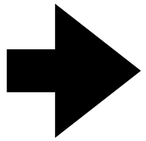
머신러닝 예측모델을 통한 수요 예측 및 인사이트 도출

AI-05 오승원

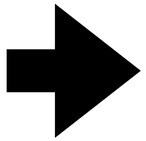
프로젝트 목표

2011, 2012년 자전거 대여 실적을 통해 수요량을 예측할 수 있는 머신러닝 모델을 만들고 정확도가 높은 수요 예측을 통해 경영 전략(마케팅 및 운영관리)를 세우고자 함.

Data 분석
- 시간, 기후 등을 통해 분석 진행



머신러닝 모델 개발
- 선형회귀, 랜덤포레스트, XGBoost



인사이트 도출
- 수요 예측 및 전략 수립

1. 선정 이유 및 문제 정의

1) 데이터 선정 이유
- 서울을 비롯한 각 지방자치 정부 조직에서 따릉이와 같은 공유 자전거 사업을 진행하고 있으며 그 확장성이 크며, 많은 시민들이 공유 자전거를 이용하는 모습을 많이 봄.
- 정당 대표가 공유 자전거를 타고 출근하는 모습이 화제가 됨.
- 개인적으로 자전거는 여가 및 취미 생활로 생각하고 있었는데 과연 출퇴근용(대중 교통 대체재)으로 사용되는지 의문이 들었음.
- 외국 사례를 통해 궁금증을 해결하고 싶었음.
- 또한, 해당 데이터가 잘 정리되어 있어 머신러닝을 학습하는데 있어 유용한 데이터라고 생각하여 선정함.

2) 문제정의

- 유형 분석: 이 데이터를 통해 예측하고자 하는 것은 자전거 대여수임. 따라서 연속형 변수에 해당하며 회귀 문제로 접근이 필요함.

2. 가설 및 평가지표, 베이스라인 선택

1) 가설
- 자전거는 대중 교통을 대체하는 교통 수단이기 보다는 여가 및 취미활동으로 인한 수요가 높다.
- 기후는 대여 수요를 결정하는데 중요한 요소일 것이다.

2) 예측하고자 하는 것(Target)
- 자전거 대여 수(수요 예측)

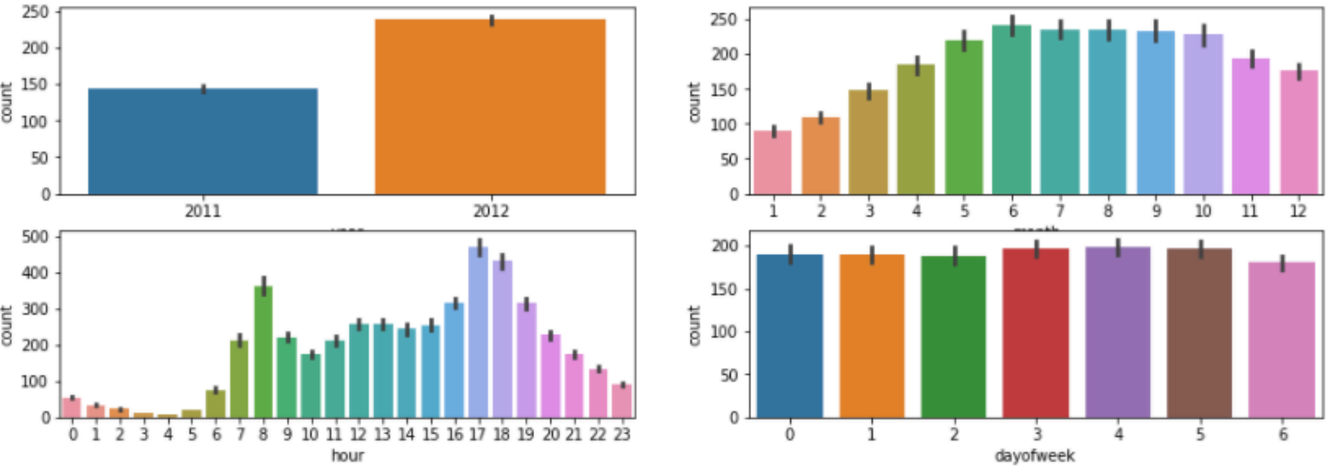
3) 평가지표
- 회귀 문제로 MAE, MSE, R2 등이 사용 가능함.
- 해당 사례에서는 R2를 기준으로 삼고 모델을 평가할 예정임.

4) 베이스라인 모델 선택
- 회귀 모델이므로 평균 대여 수를 베이스라인 모델로 선정함.

3. EDA와 데이터 전처리(1)

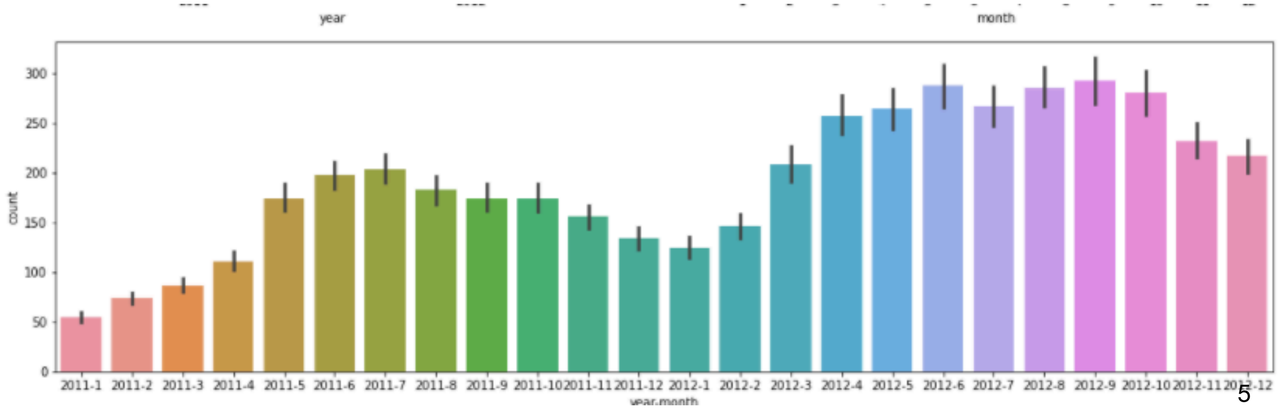
1) Datetime 분석

- 연, 월, 일, 시간, 요일



- 연도, 월, 시간에 따른 대여 수가 달라짐이 보임.
- 월 같은 경우 1월과 12월의 차이가 너무 큼. -> 계절 요인으로 추측 되었으나, 같은 겨울이라서 추가 파악이 필요함.
- 요일(0=월요일, 6=일요일)상에는 큰 차이가 없음

- 연도-월 비교



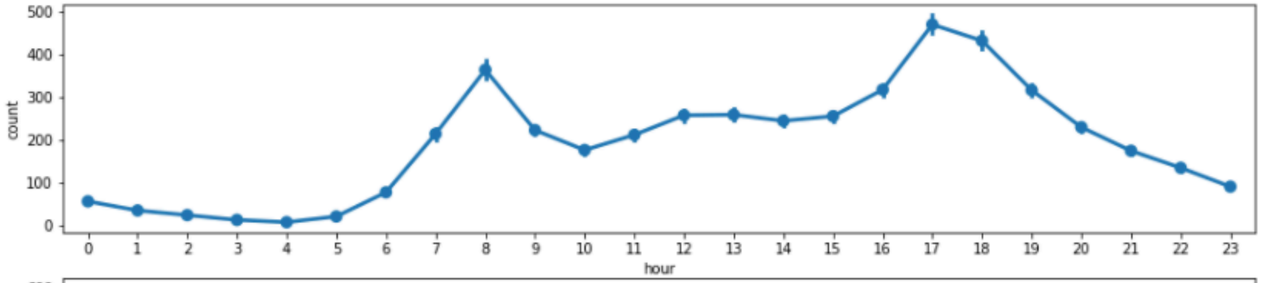
- 2011년에 비해 2012년 기업이 선정함에 따라 월별 차이가 나타남.
- 해당 변수는 과적합 위험이 있어 적절하지 않은 것으로 보임.

3. EDA와 데이터 전처리(2)

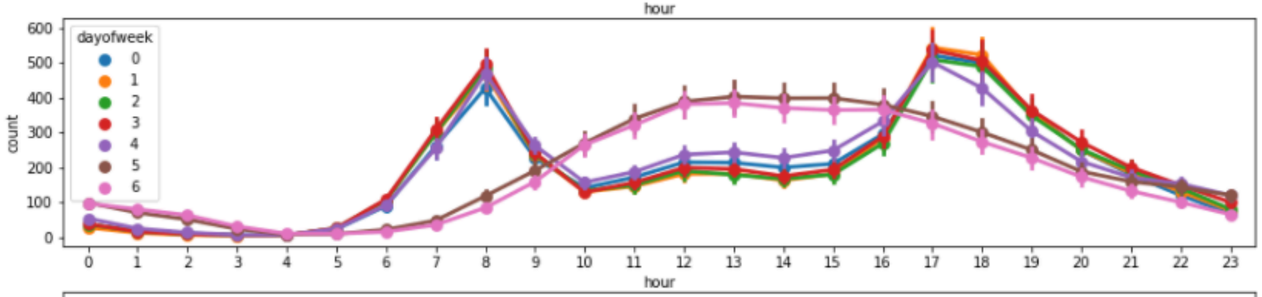
1) Datetime 분석

- 시간 분석

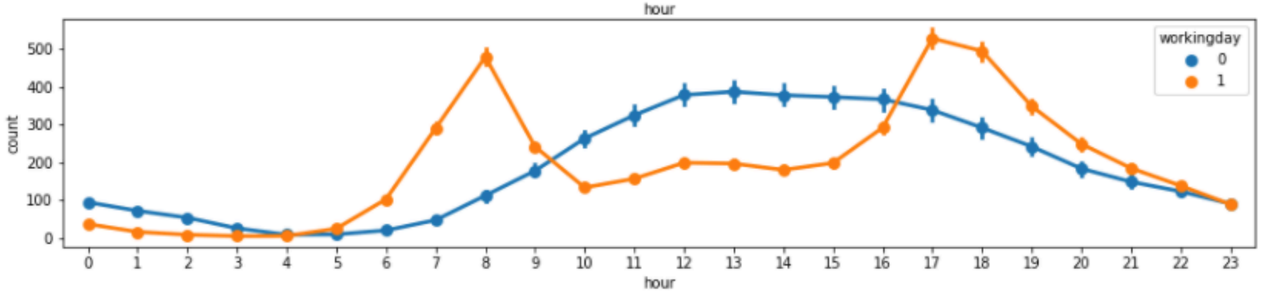
1



2



3



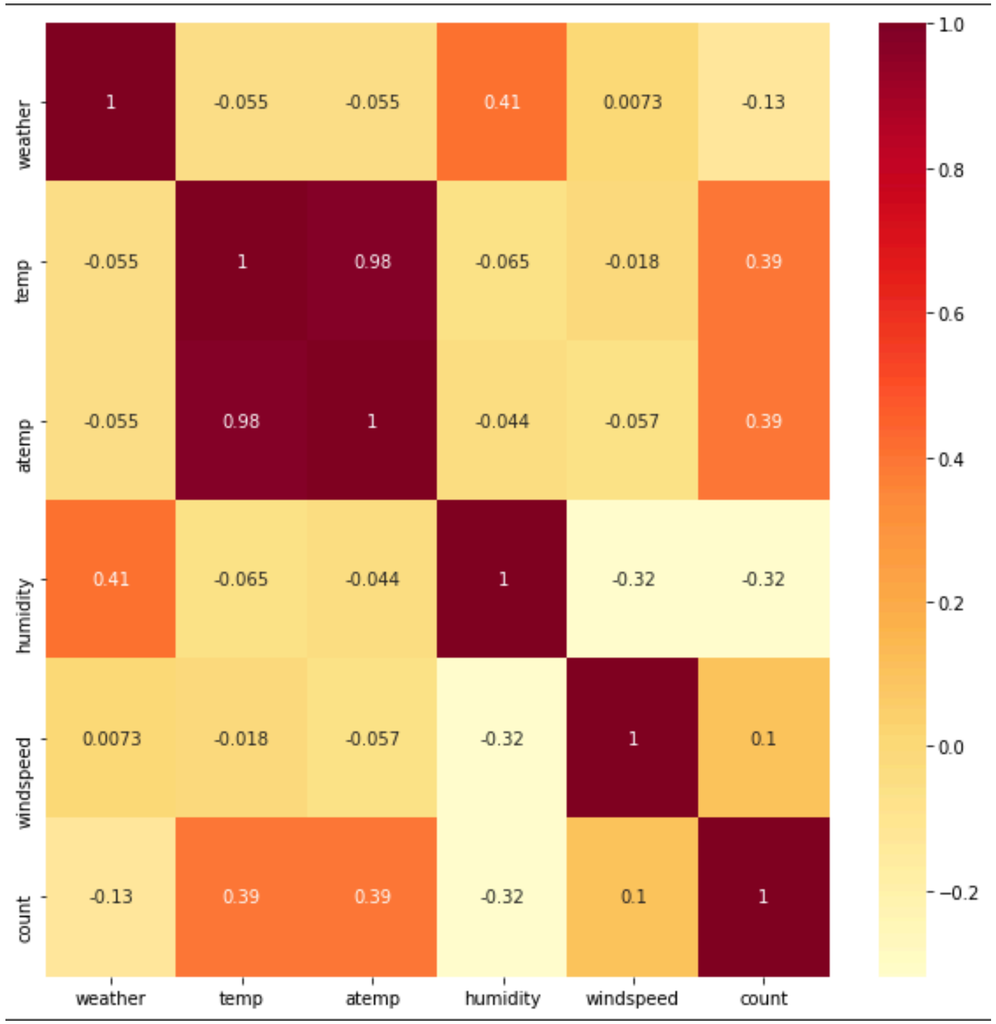
1. 시간에 따른 수요 변화
- 7~9시 수요가 크며, 17~19시 수요가 크게 나타남.

2. 요일에 따른 시간별 수요
- 평일(0~4)은 7~9시, 17~19시 수요가 크며
- 주말(5~6)은 오후대에 수요가 높음.

3. 휴일 유무에 따른 시간별 수요
- 주말, 평일과 유사한 모습을 보임.

3. EDA와 데이터 전처리(3)

2) 기후 요인 분석



1. 기후 요인과 target값인 count(대여수)와 상관관계
- season(계절), weather(맑고 흐림)과 상관관계가 작음.
- temp(기온), atemp(체감온도), humidity(습도)의 상관관계가 높음.

3. EDA와 데이터 전처리(4)

3) 데이터 전처리

- 기존 DataFrame ‘datetime’을 year, dayofweek(요일), holiday, workingday, hour 변수로 처리
- test용 dataframe이 별도로 주어졌으며 test dataframe에 없는 Casual / registered(회원 유무)는 train에 없는 특성이라 삭제함.(Leakage 방지)

- 한계점: 고객들에 대한 데이터가 매우 부족함. 경영전략을 세우기 위해선 수요 예측도 중요하나 그 수요의 주체가 되는 고객들에 대한 정보도 확보가 필요함.

기존 DataFrame

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	2011.1.1 0:00	1	0	0	1	9.84	14.395	81	0	3	13	16
1	2011.1.1 1:00	1	0	0	1	9.02	13.635	80	0	8	32	40
2	2011.1.1 2:00	1	0	0	1	9.02	13.635	80	0	5	27	32
3	2011.1.1 3:00	1	0	0	1	9.84	14.395	75	0	3	10	13
4	2011.1.1 4:00	1	0	0	1	9.84	14.395	75	0	0	1	1

전처리한 DataFrame

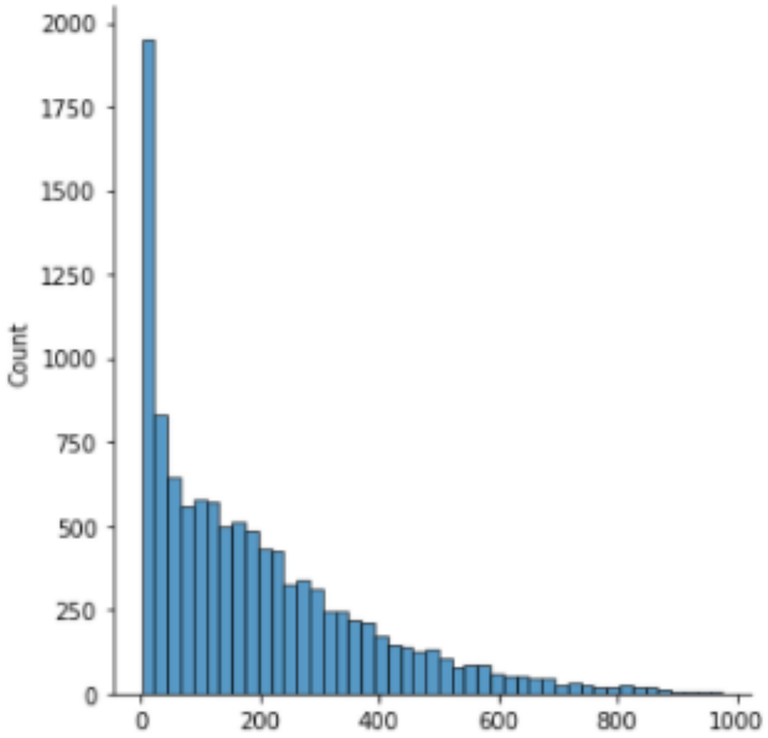
	year	dayofweek	holiday	workingday	hour	season	weather	temp	atemp	humidity	windspeed
3024	2011	3	0	1	22	3	1	26.24	31.06	57	6.0032
4605	2011	4	0	1	23	4	1	13.12	15.15	49	22.0028
8597	2012	3	0	1	14	3	1	35.26	39.395	44	6.0032
9022	2012	5	0	0	7	3	1	24.6	28.03	83	15.0013
9201	2012	3	0	1	18	3	1	29.52	34.09	70	16.9979

3. EDA와 데이터 전처리(5)

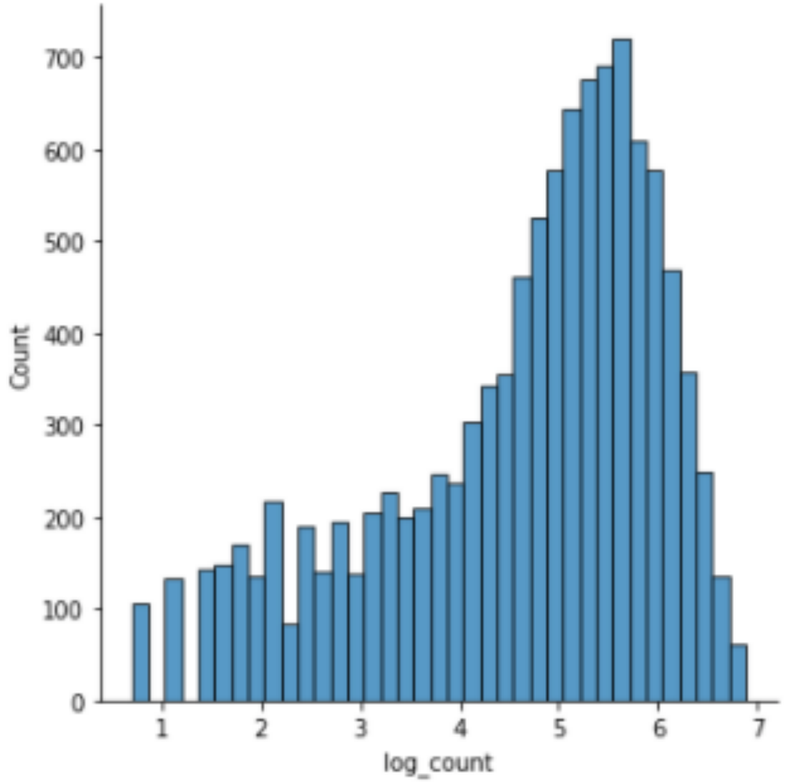
3) 데이터 전처리 (target값 로그 변환)

- Right skewed 되어 있음.
- log 변화 후 정규분포가 조금 가까워 짐

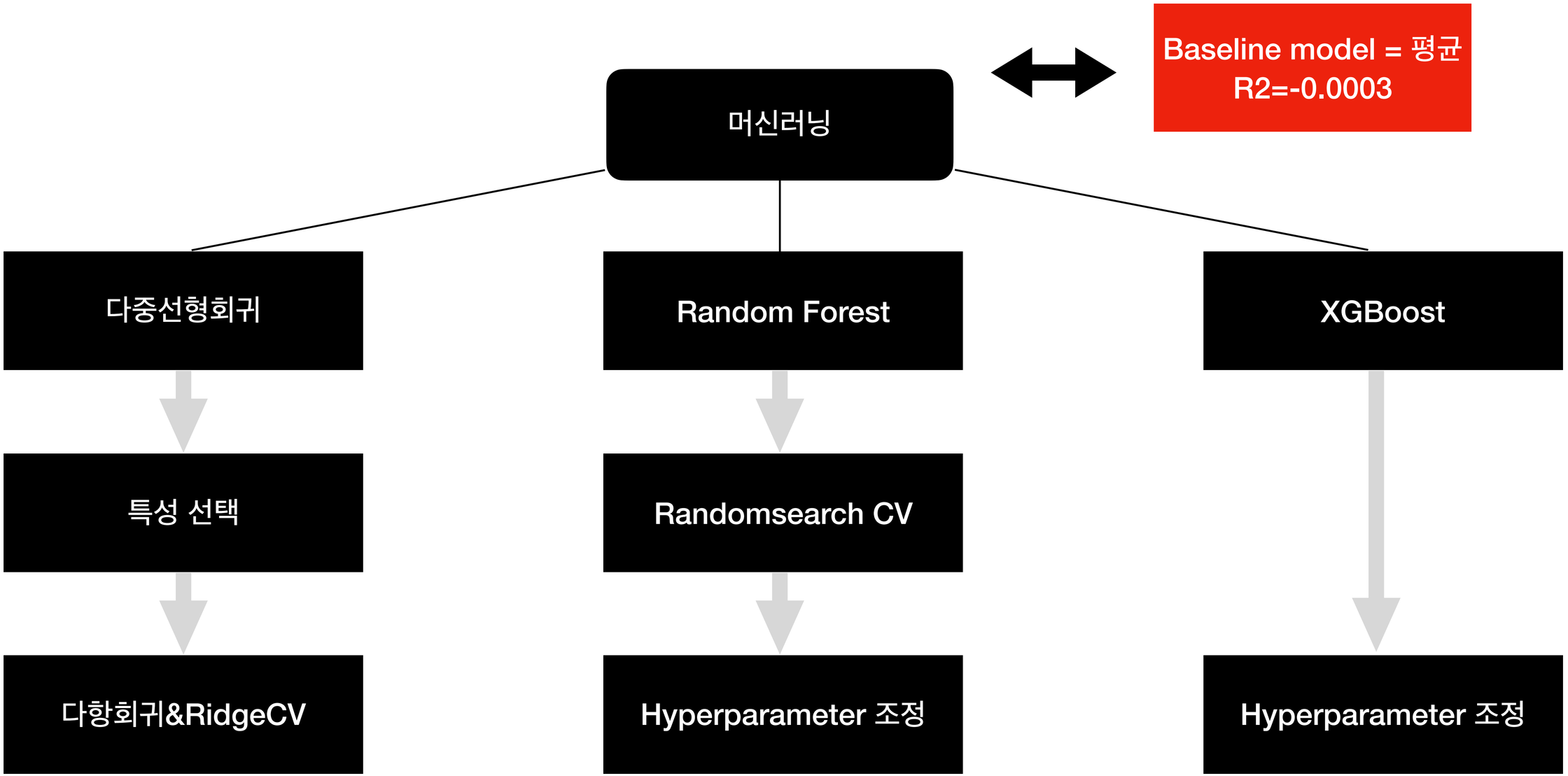
변환 전



변환 후



4. 머신러닝_진행방식


Baseline model = 평균
R2=-0.0003
머신러닝
다중선형회귀
Random Forest
XGBoost
특성 선택
Randomsearch CV
다항회귀&RidgeCV
Hyperparameter 조정
Hyperparameter 조정

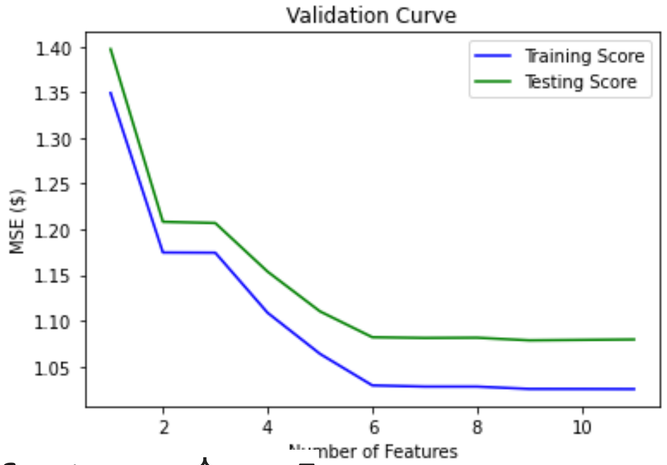
4. 머신러닝_Linear

다중선형회귀 → 특성 선택 → 다항회귀&RidgeCV

사용특성(11개)
- 'year', 'dayofweek', 'holiday', 'workingday', 'hour', 'season', ‘weather', 'temp', 'atemp', 'humidity', 'windspeed'

R2 : 0.47385

Kaggle 점수: 1.01901*
(*RMSLE* 숫자가 작을 수록 성능이 우수)




```
feature 수:  7
Selected names:[year,hour,
season,temp,atemp,humidity,
windspeed]
Test R2: 0.47304622629198123
```

사용특성 : 120
Alpha : 0.02
R2 : 0.6316

Kaggle 점수: 0.86612

4. 머신러닝_Treebased

Random Forest → Randomsearch CV → Hyperparameter 조정

사용특성(11개)
- 'year', 'dayofweek', 'holiday', 'workingday', 'hour', 'season', 'weather', 'temp', 'atemp', 'humidity', 'windspeed'

R2 : 0.95089

Kaggle 점수: 0.39894
($RMSLE$ 숫자가 작을 수록 성능이 우수)

Best Parameters:
simpleimputer__strategy': 'median',
n_estimators': 150
max_features': 8,
max_depth: 17

R2: 0.95141

Train, val로 분류한 데이터를 합쳐 원래의 train 데이터를 만듬.
-> Best Parameter를 적용하여 재학습.

Kaggle 점수: 0.39909

왜 교차검증을 했는데도 점수는 더 떨어질까?
-> 자전거 대여 수요 예측에서 시계열 정보가 가장 중요한데 교차검증을 할 경우 시계열 데이터를 교란하기 때문이지 않을까 추측함.

4. 머신러닝_Gradient boosting

XGBoost → Hyperparameter 조정

사용특성(11개)
- 'year', 'dayofweek', 'holiday', 'workingday', 'hour', 'season', ‘weather', 'temp', 'atemp', 'humidity', 'windspeed'

R2 : 0.92582

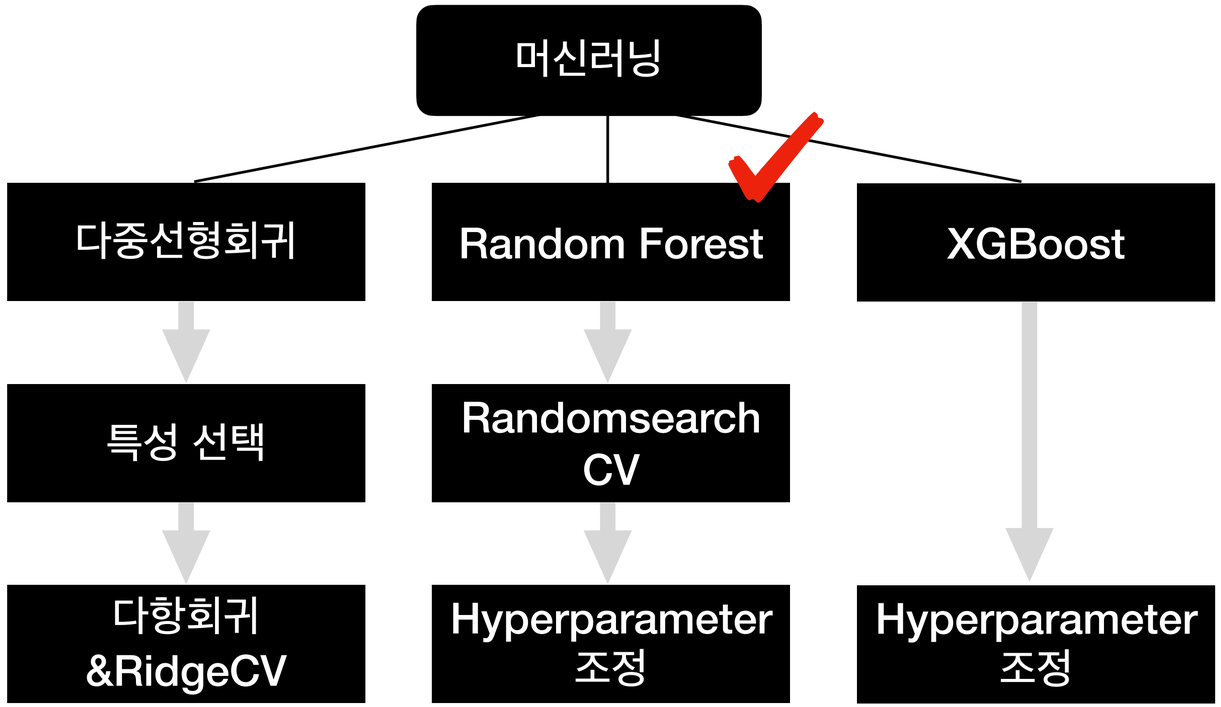
Kaggle 점수: 0.44043
($RMSLE$ 숫자가 작을 수록 성능이 우수)

Parameter 설정
- n_estimators=100
- max_depth=13,
- learning_rate=0.2

R2: 0.95651

Kaggle 점수: 0.39936

4. 머신러닝_모델선택



	다중선형회귀	Random Forest	XGBoost
R2	0.6316	0.95089	0.95651
Kaggle	0.86612	0.39894	0.39936

R2는 높을 수록, Kaggle(RMSLE)는 작을 수록 성능이 좋음.

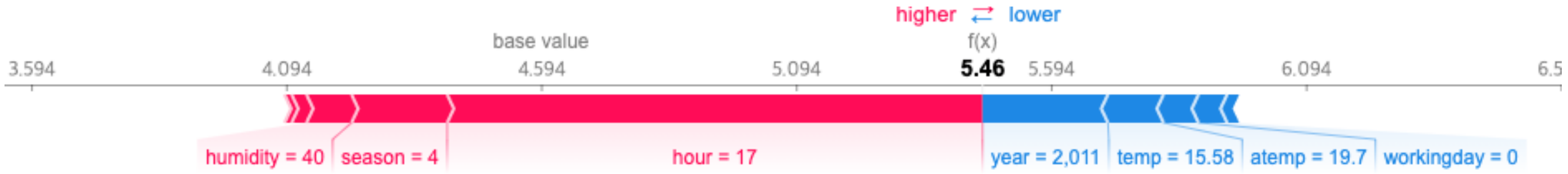
4. 머신러닝_모델해석(Shap 사용)

1) force plot 확인

-예측값(log) : 5.814131

-실제값(log): 5.4601

	year	dayofweek	holiday	workingday	hour	season	weather	temp	atemp	humidity	windspeed
4623	2011	5	0	0	17	4	1	15.58	19.695	40	11.0014

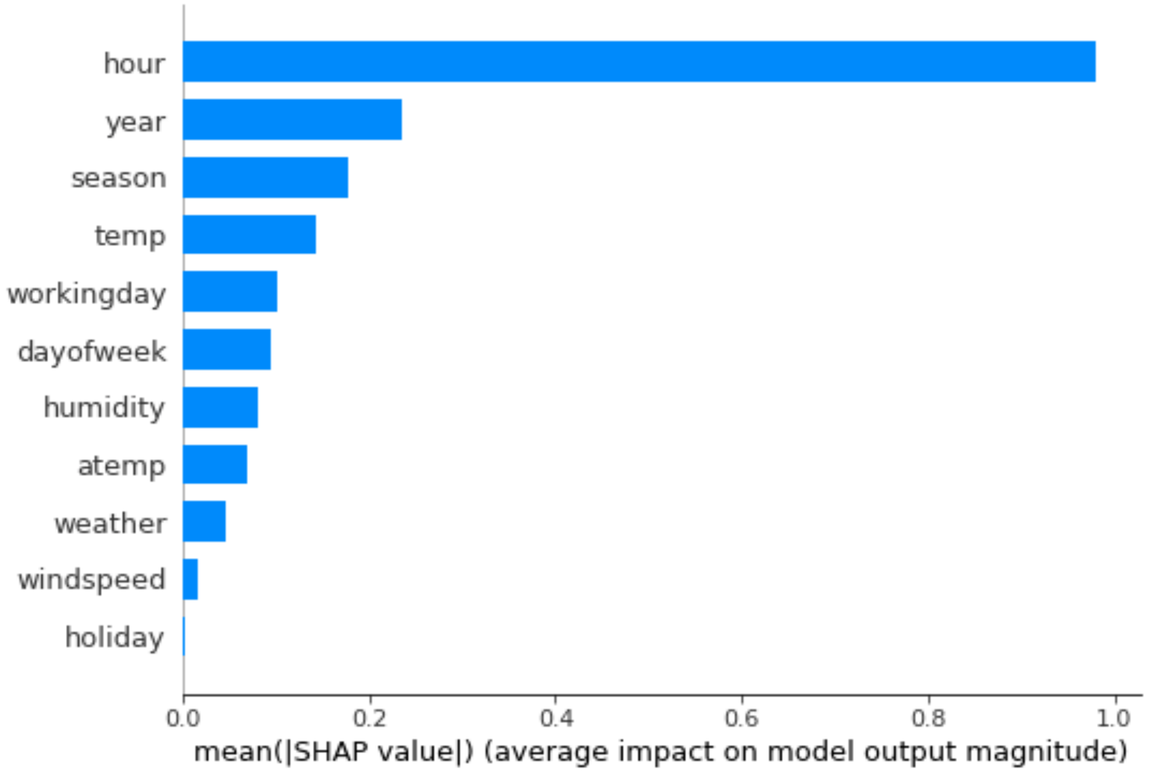


해석

- 수요 예측에 있어 hour, season, humidity가 자전거 수요를 높임
- 반면, year, temp, atemp, working day 는 값을 낮추는 효과로 나타남.

4. 머신러닝_모델해석(Shap 사용)

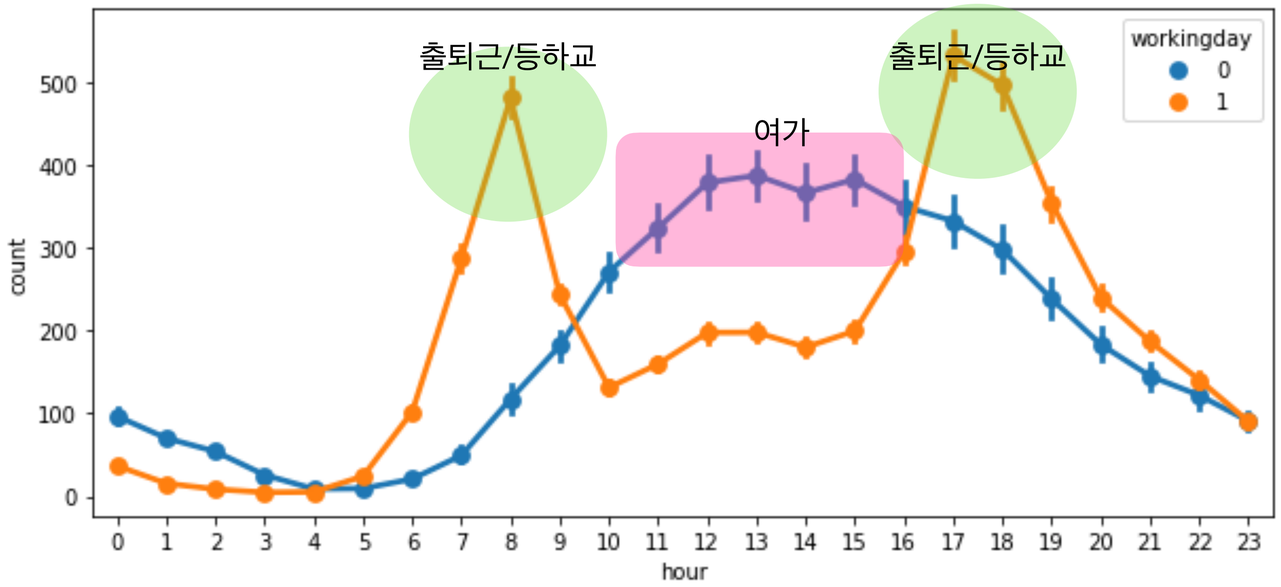
2) 각 특성이 미치는 영향력 분석



해석

- 자전거 대여 수요에 가장 큰 영향력을 미치는 요소는 hour 임.
- 이는 앞 EDA과정을 통해서 추측한 것과 동일한 결과를 보여줌
- year요소는 해당 기업이 성장하는 것을 보여줌.
- Season, temp 와 같은 기후 요소도 영향을 주지만 크지 않음.

5. 결론



○ 경영전략 측면

- 자전거 대여 수요는 시간의 영향을 가장 많이 받음.
- 이는 고객층을 출퇴근/등하교로 사용하는 고객과 여가/취미로 사용하는 고객이 다르게 형성되어 있음을 알려줌.
- 두 고객층을 동시에 공략하면 좋겠지만 자원의 한계를 고려하였을 때 출퇴근/등하교 고객층(직장인/학생)에 먼저 집중이 필요
- 마케팅 전략 방향을 직장인/학생을 중심으로 설정 필요

○ 운영측면

- 자전거 정거장을 회사나 학교 중심으로 이동 및 확장 필요
- 자전거 관리는 평일은 수요가 적은 이른 오전, 오후, 늦은 저녁 중심으로 진행 필요. 휴일은 오후 시간대를 피하는 것이 효율적임.