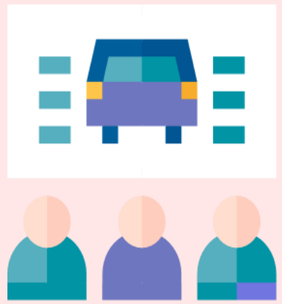
중고차 가격 예측 모델

이루지명
백지명, 강진영, 고예성, 박현식, 안영준, 조세연

목차

01 데이터셋



01

DataSet

kaggle

	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price	Brand
0	Mumbai	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	1.75	Maruti
1	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	12.50	Hyundai
2	Chennai	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	4.50	Honda
3	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	6.00	Maruti
4	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	17.74	Audi
...	...	...	...	...	...	...	...	...	...	...	...	...
6014	Delhi	2014	27365	Diesel	Manual	First	28.40	1248.0	74.00	5.0	4.75	Maruti
6015	Jaipur	2015	100000	Diesel	Manual	First	24.40	1120.0	71.00	5.0	4.00	Hyundai
6016	Jaipur	2012	55000	Diesel	Manual	Second	14.00	2498.0	112.00	8.0	2.90	Mahindra
6017	Kolkata	2013	46000	Petrol	Manual	First	18.90	998.0	67.10	5.0	2.65	Maruti
6018	Hyderabad	2011	47000	Diesel	Manual	First	25.44	936.0	57.60	5.0	2.50	Chevrolet

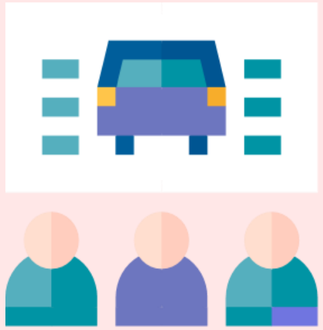
5872 rows × 12 columns

01

DataSet

Column.	Meaning.
Brand	차량 브랜드
Location	위치
Year	연도
Kilometers_Driven	주행거리
Fuel_Type	연료 종류
Transmission	변속기 종류
Owner_Type	차량 소유자 유형
Mileage	연비
Engine	엔진 배기량
Power	최대 출력
Seats	탑승 인원 수
Price	판매 가격

02 모델 설명



릿지, 라쏘

릿지(Ridge), 라쏘(Lasso) = 회귀분석에서 사용되는 통계적인 기법

과적합을 줄이고 모델의 일반화 성능을 향상시키기 위해 사용되는 방법

L1 규제: 라쏘

- L1패널티를 사용
- 가중치의 절댓값의 합을 최소화하는 방향으로 모델을 조정
- 변수 선택 기능을 갖추고 있어서 일부 변수의 가중치를 정확히 0으로 만들 수 있다
- L1 패널티를 사용해 변수 선택과 가중치 축소를 동시에 수행할 수 있음

L2규제: 릿지

- L2패널티를 사용
- 가중치의 제곱합을 최소화 하는 방향으로 모델 조정
- 모든 변수를 유지하면서 가중치를 축소시킨다
- L2패널티를 사용해 모든 변수의 가중치를 작게 만듦

엘라스틱넷

요약

릿지의 장점 + 라쏘의 장점

사용방법

규제항을 단순히 더해 사용한다. 두 규제항의 혼합 정도를 혼합비율 r을 사용하여 조절한다.
r=0 : 릿지 회귀와 같음 r=1: 라쏘 회귀와 같음

장단점

- 장점: 변수의 수가 훈련 샘플의 수보다 극단적으로 많거나 변수 몇개가 강하게 연관되어 있을 경우 사용하면 효과적인 결과를 얻을 수 있다 ! => 다중공선성이 있는 데이터셋에서 효과적
- 단점: 실행시 시간이 위의 두 규제보다 시간이 오래 걸린다

스태킹, 블랜딩

스태킹(Stacking)과 블랜딩(Blending)은 앙상블 학습에서 사용되는 방법으로
다양한 모델의 예측력을 결합해 더 강력하고 안정적인 예측 모델을 구축하는데 사용

스태킹(Stacking)

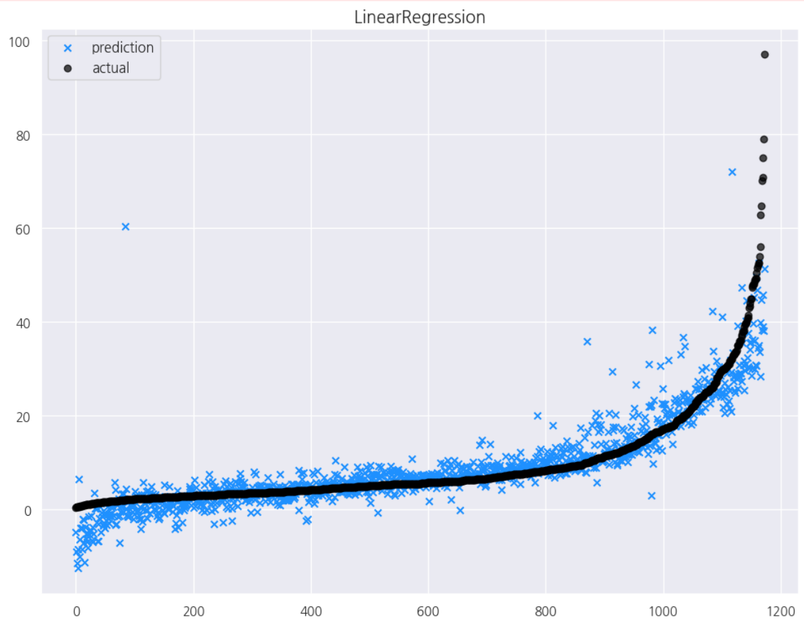
- 다양한 기본 모델의 예측결과를 활용해 최종 예측 모델을 생성하는 방법
- 기본 모델의 예측결과를 사용해 새로운 특성으로 변환한 후, 이를 다른 모델에 입력해 최종 예측을 수행한다
- cross-fold-validation 사용

블랜딩(Blending)

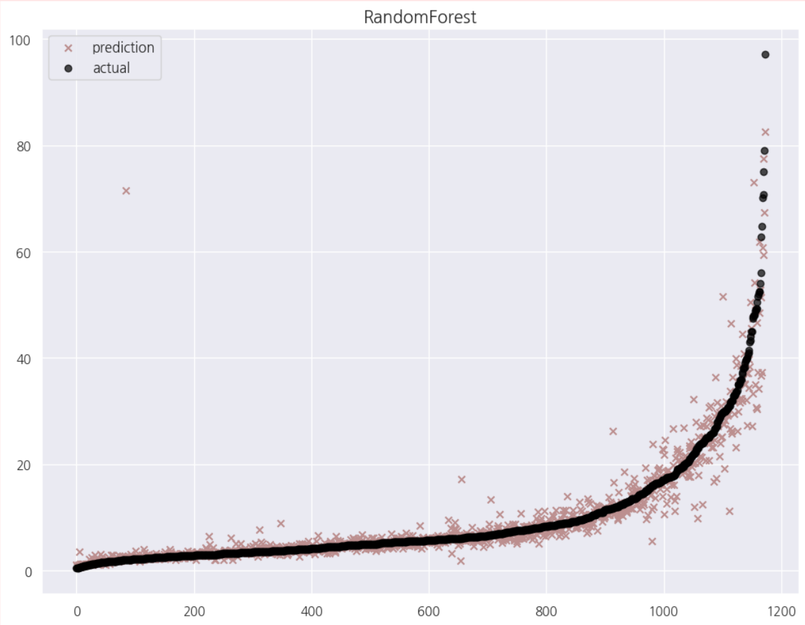
- 다양한 모델을 학습하여 가중 평균 등을 통해 최종 예측을 결합하는 방법
- 보통 학습데이터를 분할해서 일부를 사용해 모델을 학습하고, 나머지를 사용해 각 모델의 예측을 평가하여 가중치를 결정한다
- holdout validation 사용
- 모델에 대한 가중치를 조절해 최종 아웃풋을 산출하며, 가중치의 합은 1.0이 되도록 한다

03 예측값과 실제값 비교

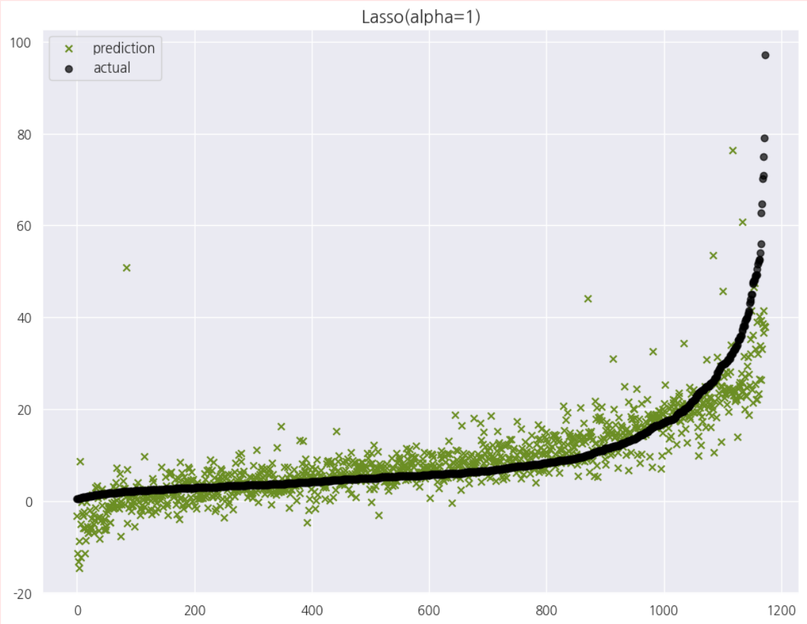
선형회귀


LinearRegression
prediction
actual
100
80
60
40
20
0
0
200
400
600
800
1000
1200

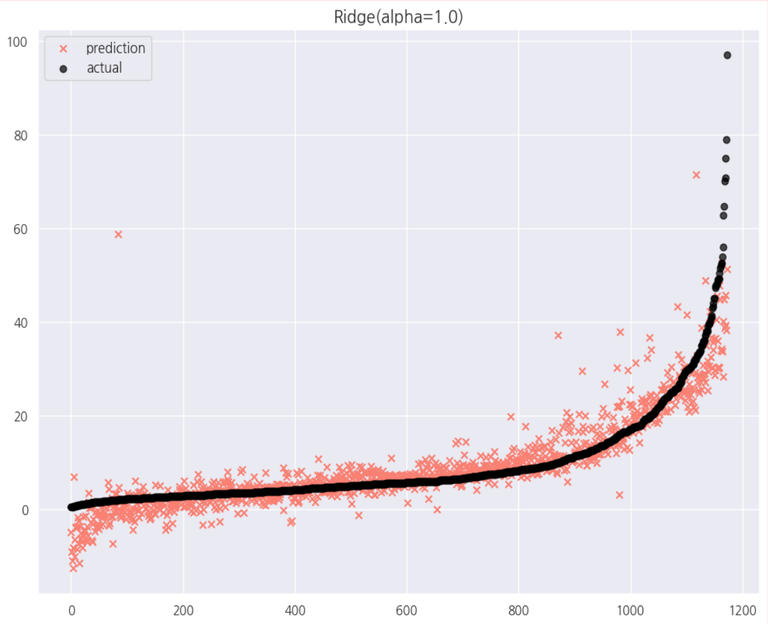
랜덤 포레스트


RandomForest
prediction
actual
100
80
60
40
20
0
0
200
400
600
800
1000
1200

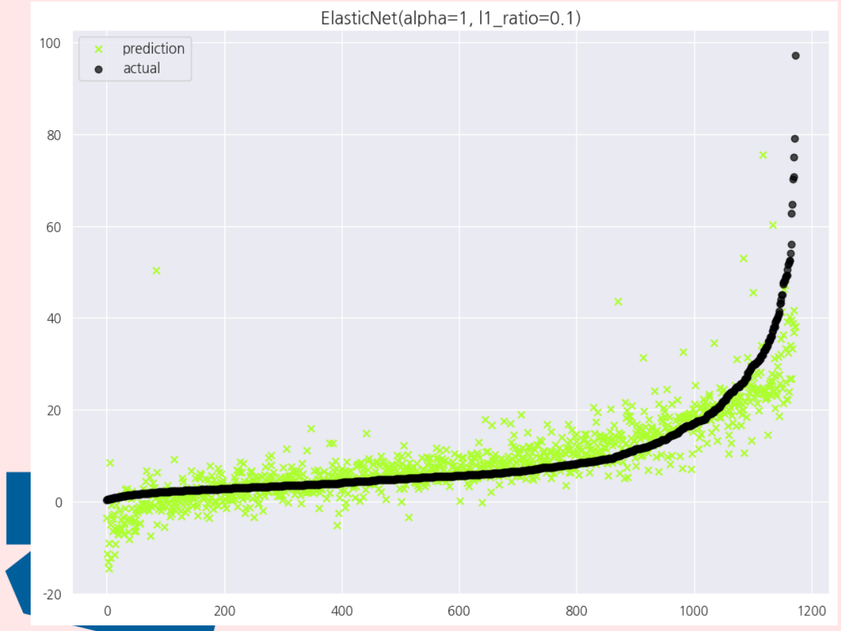
라쏘


Lasso(alpha=1)
prediction
actual
100
80
60
40
20
0
-20
0
200
400
600
800
1000
1200

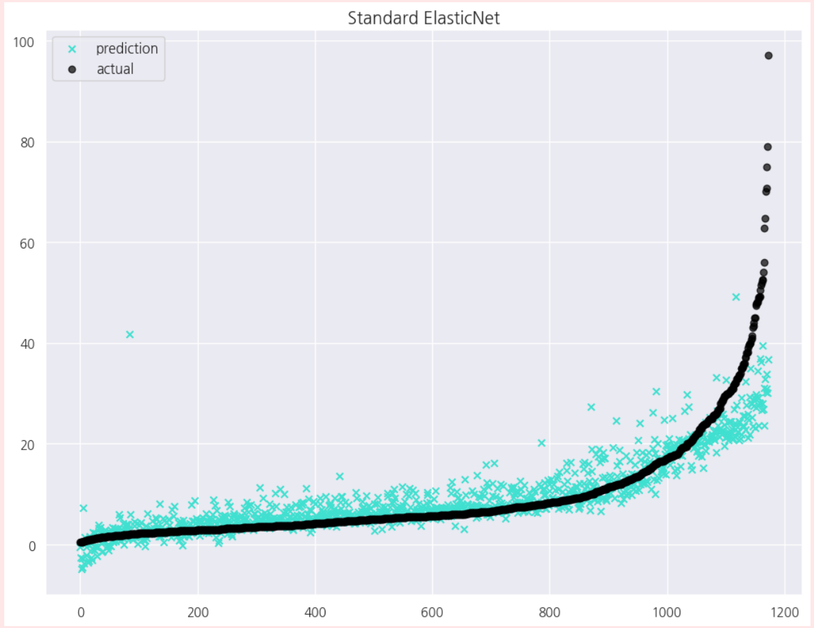
릿지


Ridge(alpha=1.0)
prediction
actual
100
80
60
40
20
0
0
200
400
600
800
1000
1200

엘라스틱넷

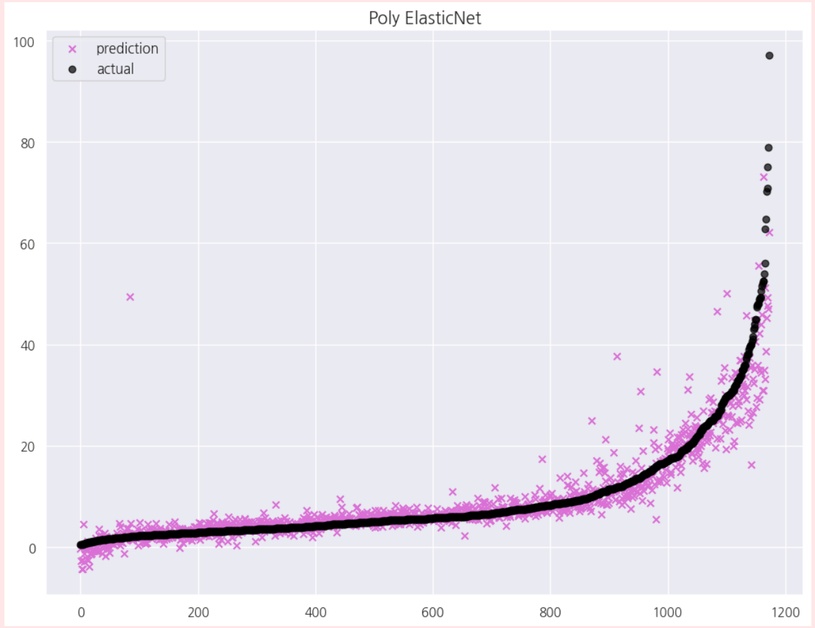


스탠다드 엘라스틱넷



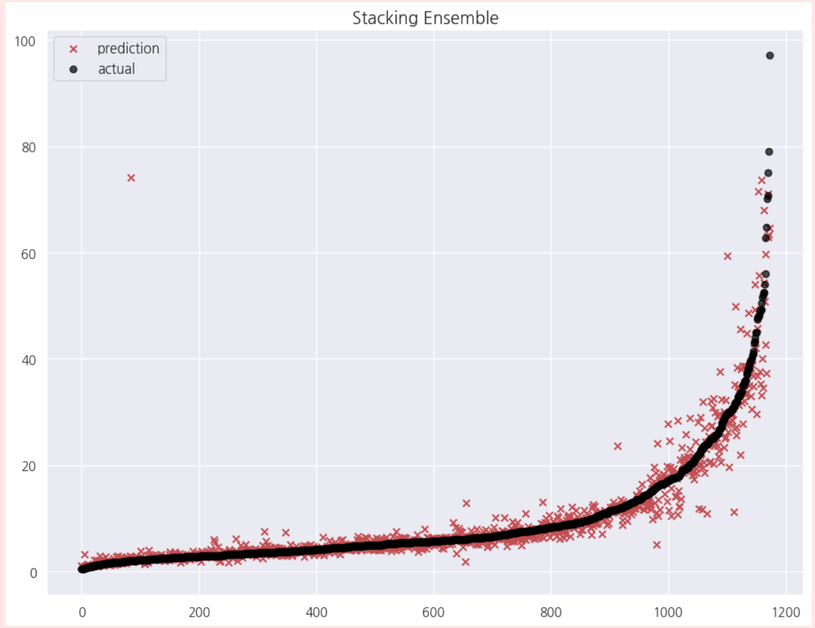
- StandardScaler로 scaling
- 엘라스틱넷을 적용

폴리 엘라스틱넷

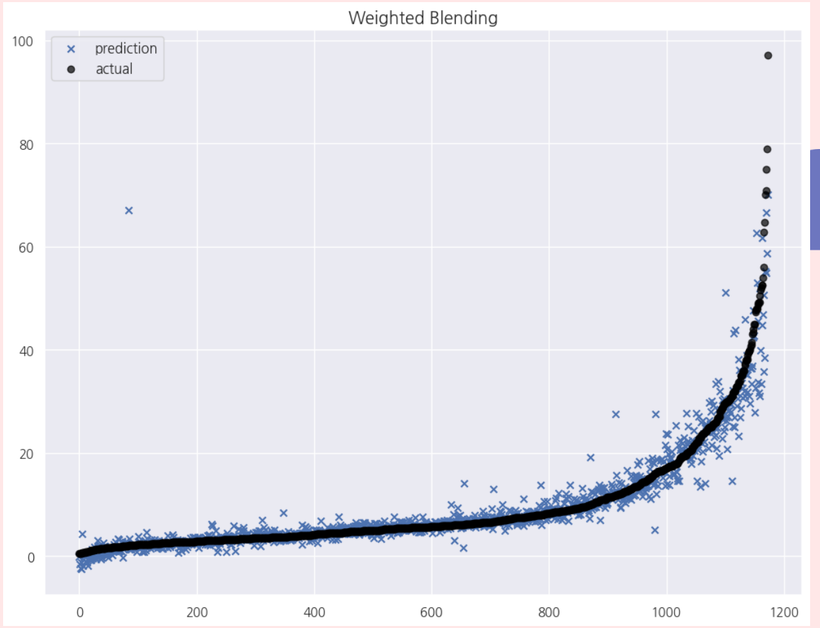


- 다항식 특성을 추가
- StandardScaler로 scaling
- 엘라스틱넷 적용

스태킹

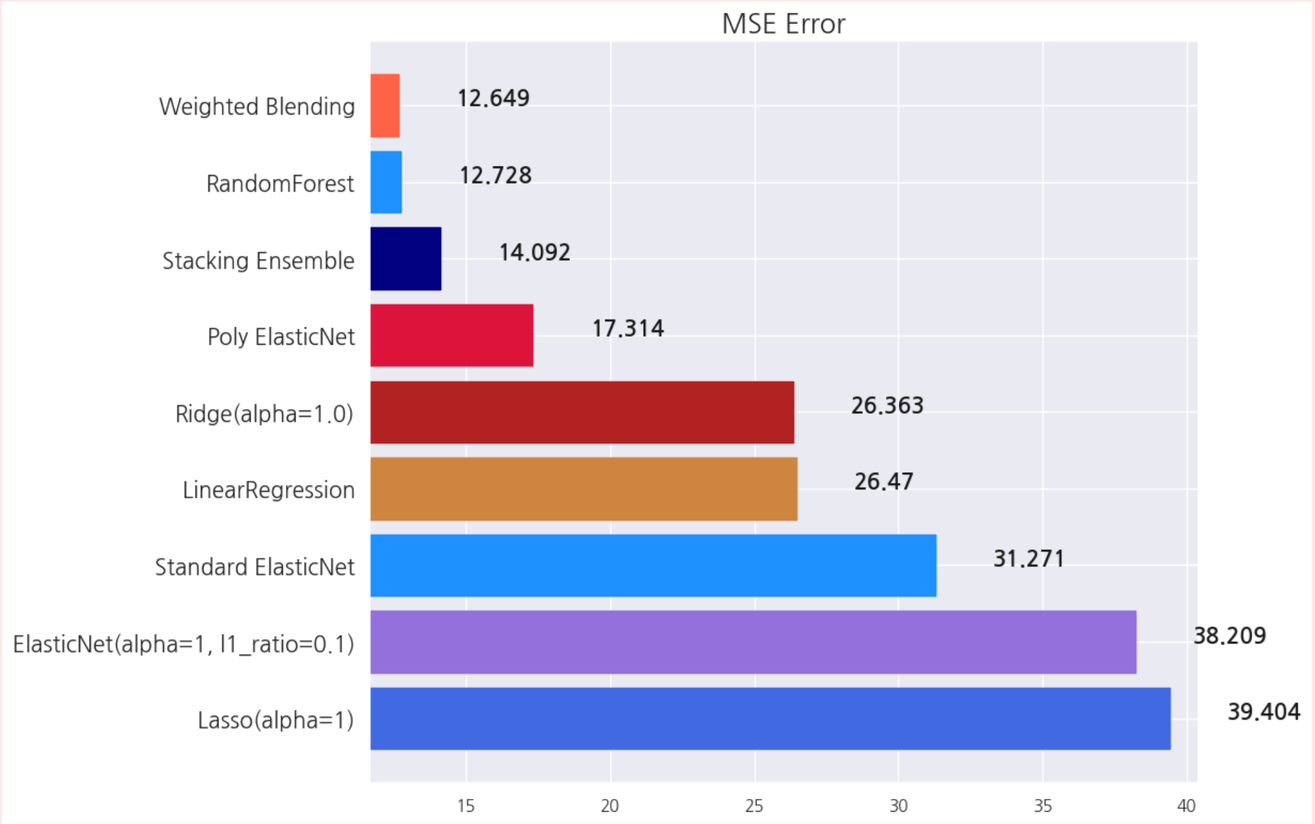

Stacking Ensemble
prediction
actual
100
80
60
40
20
0
0
200
400
600
800
1000
1200

블렌딩

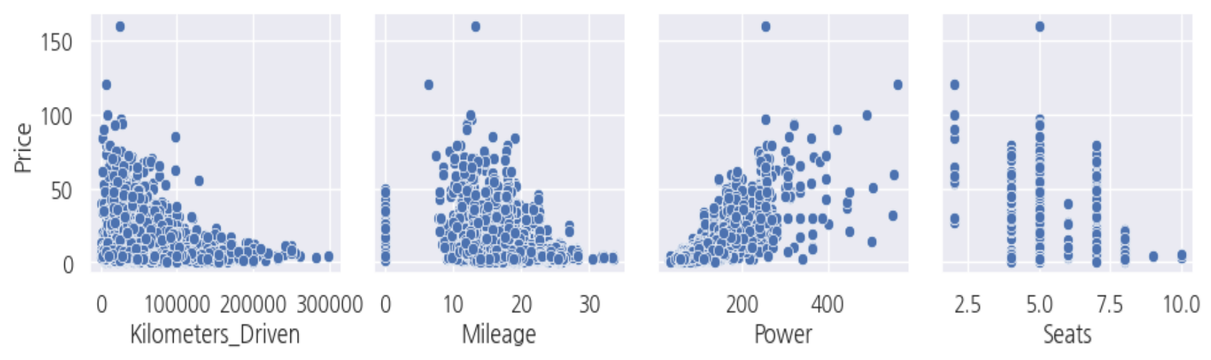

Weighted Blending
prediction
actual
100
80
60
40
20
0
0
200
400
600
800
1000
1200

04 결론

모델별 예측값 비교


MSE Error
Weighted Blending
12.649
RandomForest
12.728
Stacking Ensemble
14.092
Poly ElasticNet
17.314
Ridge(alpha=1.0)
26.363
LinearRegression
26.47
Standard ElasticNet
31.271
ElasticNet(alpha=1, l1_ratio=0.1)
38.209
Lasso(alpha=1)
39.404
15
20
25
30
35
40

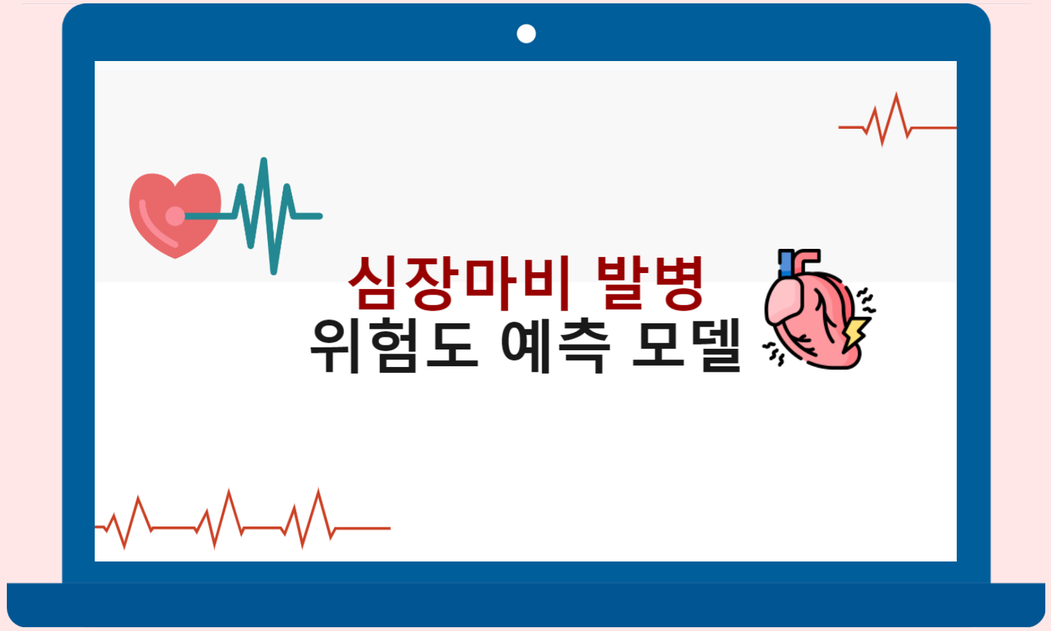
결론



- Random Forest의 MSE 값이 선형 회귀보다 더 낮은 이유는 Power를 제외한 나머지 데이터들이 선형이 아니기 때문이라 예측됨
- Weighted Blending은 가중치를 조정하여 각 모델의 중요도를 조절하여 전체 예측 성능을 향상시킴
- Weighted Blending은 다른 ensemble 모델과 달리 가중치를 조정하여 앙상블의 효과를 극대화시킴

05 트러블 슈팅

트러블슈팅



심장마비 예측모델에서
중고차가격 예측모델로 옮긴 이유

- 심장마비 데이터가 양이 적었다.
- 종속변수 이진분류여서 회귀모델을 적용하기 부적합했다.

감사합니다