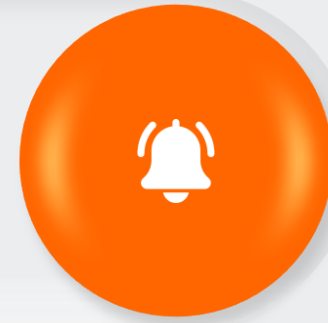


중고차 가격 예측



18101968 김범준
19102028 최우진



목차

주제 선정 배경 ▼

주제 선정 배경과 분석 목적

데이터 수집 ▼

데이터 소개

데이터 수집 방법

데이터 분석 ▼

데이터 처리 및 해석

회귀 분석

Decision Tree

결론



주제 선정 배경



다시 활기 찾은 중고차 시장..매물 늘며 가격 빠르게 하락

권진욱 기자 | news726@seoulfn.com | 승인 2023.04.20 09:30 | 댓글 0

지난해 비정상적 거래 시장 위축 현상 사라져
대형차·SUV 중심으로 중고차 거래 회복



서울 동대문구 장한평 중고차 시장. (사진=연합뉴스)

침체되었던 중고차 시장이 다시 활기를 띄기 시작

물가 상승으로 신차보다 중고차 구매 현상 증가

대학생, 사회초년생들은 첫 차로 신차보다 중고차에 더 관심

중고차 구매 시 도움이 될 수 있는 기준 마련

데이터 수집







데이터 소개

보배드림 (중고차 거래, 커뮤니티 사이트)

일반등록 1,990 대

최근 등록순 | 20개씩 |

사진+동영상	차량정보	연식	연료	주행	가격	지역 / 판매자
	세화 카니발 하이리무진 디바인 C 전동발판/확장형 루프/현금,할부,리스,렌트 가능 자동 · 6기통 · 294마력 · 36.2kgm · FF · 보험이력 · 등록증	23/05	가솔린	7km	6,500 만원	이인희 (반복) 경기 안양시 등록 05/28 조회 557
	제네시스 G90 1세대 3.3 터보 AWD 프리미엄 럭셔리 자동 · 5인승 · 6기통 · 370마력 · 52kgm · AWD · 보험이력	21/09	가솔린	7만km	150 만원 136만원 / 40개월 운용리스	양명기 (반복) 부산 강서구 등록 05/28 조회 1,277
	제네시스 GV80 3.0 디젤 5인승 자동 · 5인승 · 6기통 · 278마력 · 60kgm · FR · 보험이력	21/05	디젤	4만km	7 만원 99만원 / 36개월 운용리스	양명기 (반복) 부산 강서구 등록 05/28 조회 140
	제네시스 더 올 뉴 G80 2.5 터보 AWD 자동 · 5인승 · 4기통 · 304마력 · 42.9kgm · AWD · 보험이력	22/09	가솔린	5천km	1,950 만원 74만원 / 40개월 렌트	양명기 (반복) 부산 강서구 등록 05/28 조회 629
	현대 올 뉴 아반떼 1.6 모던 자동 · 5인승 · 4기통 · 123마력 · 15.7kgm · FF · 보험이력	21/04	가솔린	1만km	200 만원 29만원 / 36개월 운용리스	양명기 (반복) 부산 강서구 등록 05/28 조회 589
	르노코리아 뉴 QM6 2.0 LPe 2WD RE 자동 · 5인승 · 4기통 · 140마력 · 19.7kgm · FF · 보험이력	20/11	LPG	4만km	150 만원 50만원 / 30개월 렌트	양명기 (반복) 부산 강서구 등록 05/28 조회 1,095

보배드림

로그인 | 회원가입

인기

사이버매장 국산차 수입차 튜닝카 중고차시세 커뮤니티 내차팔기 오토바이 중고장터 업체검색

국산차검색 튜닝카검색 리스렌트차량 사이버매장





데이터 소개

현대 올 뉴 아반떼 1.6 모던

신차수준
↓ 최저가
리스가능

21년 04월식 | 13,590 km | 가솔린

인수비용
200만원

월리스료
29만원

잔여개월
36 / 60개월

보험료조회

보험이력 0건

운용

리스종류

자동계산기

양명기 프리랜서

판매중 12 판매완료 31

기본정보

연식	2021.04	배기량	1,598 cc (123마력)
주행거리	13,590 km	색상	흰색
변속기	자동	보증정보	60개월 / 100,000km
연료	가솔린	확인사항	

리스정보

리스종류	운용리스	보증금	0만원
인수비용	200만원	잔존가치	896만원
월리스료	29만원	미회수원금	1,044만원
리스기간	21/05 ~ 26/05 (잔여 36개월 / 총 60개월)		

차량제원

엔진 형식
1.6 I4 멀티 포인트
인젝션(MPI)

연비
15.4km/ℓ

최고출력
123마력

최대토크
15.7kg.m

차량중량
1,205kg

구동방식
전륜 FF

타이어
225/45/17

더보기

보험처리이력

<p>보험처리 0회</p>	<p>차량번호/소유자변경 0회 / 0회</p> <p>자동차보험 특수사고 전손: 0 / 침수전손: 0 / 침수분손: 0 / 도난: 0</p>	<p>보험사고(내차피해) 0회 (0원)</p> <p>보험사고(타차가해) 0회 (0원)</p>
----------------	---	---

상세보기 +


사용 변수



모델명, 연식, 주행거리, 연료, 배기량, 변속기, 연비, 최고출력,
최대 토크, 차량 중량, 구동방식, 보험처리, 중고차 가격



국산 차량 (현대, 기아) 데이터만 사용
1103개 의 차량 데이터 수집



모델명, 연식, 주행거리, 연료, 배기량, 변속기, 연비, 최고출력,
최대 토크, 차량 중량, 구동방식, 보험처리, 중고차 가격

```
Created on Wed May 13 17:05:23 2023

@author: PC
"""

from time import sleep
import requests
from bs4 import BeautifulSoup
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

totalurl=[("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=5&page={}&order=S11&view_size=20",7),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=17&page={}&order=S11&view_size=20",7),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=42&page={}&order=S11&view_size=20",6),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=21&page={}&order=S11&view_size=20",4),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=22&page={}&order=S11&view_size=20",4),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=30&page={}&order=S11&view_size=20",3),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=49&group_no=20&page={}&order=S11&view_size=20",3),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=76&page={}&order=S11&view_size=20",10),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=59&page={}&order=S11&view_size=20",5),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=45&page={}&order=S11&view_size=20",4),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=50&page={}&order=S11&view_size=20",3),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=55&page={}&order=S11&view_size=20",3),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=69&page={}&order=S11&view_size=20",3),
("https://www.bobaedream.co.kr/mycar/mycar_list.php?gubun=K&maker_no=38&group_no=44&page={}&order=S11&view_size=20",2)]

df_cars=[]

urls=[]
for i in totalurl:
    pagenum=i[1]
    for j in range(pagenum):
        url=i[0].format(str(j))
        urls.append(url)

info=["모엿열","연식","주행거리","연료","배기량","변속기","연비","최고속력","최대토크","차량종류","구동방식","보험처리","종교가격","신차"]

for url in urls:
    res=requests.get(url,verify = False,timeout=50000)
    res.raise_for_status()
    requests.adapters.DEFAULT_RETRIES = 10000
    soup=BeautifulSoup(res.text,"lxml")

    cars=soup.find_all("Li",attrs={"class":"product-item"})
    links=[]

    for car in cars:
```

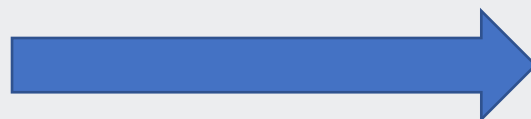

데이터 분석



데이터 처리 및 해석

```
df_cars_all.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1103 entries, 0 to 1102  
Data columns (total 14 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   Unnamed: 0    1103 non-null   int64  
1   모델명        1103 non-null   object  
2   연식          1103 non-null   object  
3   주행거리      1103 non-null   object  
4   연료          1103 non-null   object  
5   배기량        1103 non-null   object  
6   변속기        1103 non-null   object  
7   연비          815 non-null    object  
8   최고출력      907 non-null    object  
9   최대토크      864 non-null    object  
10  차량중량      843 non-null    object  
11  구동방식      896 non-null    object  
12  보험처리      1103 non-null   int64  
13  중고가격      1103 non-null   object  
dtypes: int64(2), object(12)  
memory usage: 120.8+ KB
```



결측데이터 처리

1103개 -> 798개

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 798 entries, 0 to 1102  
Data columns (total 14 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   Unnamed: 0    798 non-null   int64  
1   모델명        798 non-null   object  
2   연식          798 non-null   object  
3   주행거리      798 non-null   object  
4   연료          798 non-null   object  
5   배기량        798 non-null   object  
6   변속기        798 non-null   object  
7   연비          798 non-null   object  
8   최고출력      798 non-null   object  
9   최대토크      798 non-null   object  
10  차량중량      798 non-null   object  
11  구동방식      798 non-null   object  
12  보험처리      798 non-null   int64  
13  중고가격      798 non-null   object  
dtypes: int64(2), object(12)  
memory usage: 93.5+ KB
```



데이터 처리 및 해석

Unnamed: 0		모델명	연식	주행거리	연료	배기량	변속기	연비	최고출력	최대토크	차량중량	구동방식	보험처리	중고가격
0	0	현대 더 뉴 그랜저 하이브리드 2.4 HEV 르블랑	2021.11 (22년형)	37,983 km	가솔린 하이브리드	2,359 cc (200마력)	자동	16.2km/ℓ	200마력	25.1kg.m	1,675kg	전륜 FF	4	3,700만원
1	1	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	2017.09 (18년형)	60,000 km	가솔린 하이브리드	2,359 cc (199마력)	자동	16.2km/ℓ	199마력	26.7kg.m	1,675kg	전륜 FF	23	2,050만원
2	2	현대 더 뉴 그랜저 하이브리드 2.4 HEV 익스클루시브	2020.03	43,818 km	가솔린 하이브리드	2,359 cc (159마력)	자동	16.2km/ℓ	159마력	21kg.m	1,675kg	전륜 FF	2	3,469만원
3	3	현대 그랜저IG 2.4 프리미엄	2018.01	61,406 km	가솔린	2,359 cc (190마력)	자동	11.2km/ℓ	190마력	24.6kg.m	1,550kg	전륜 FF	0	1,940만원
4	4	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	2018.10 (19년형)	56,098 km	가솔린 하이브리드	2,359 cc (199마력)	자동	16.2km/ℓ	199마력	26.7kg.m	1,675kg	전륜 FF	0	2,800만원

Unnamed : 열 제거

is_genesis : 새로운 열을 추가하여 모델명에 '제네시스'가 포함되면 1, 그렇지 않으면 0 을 출력

연식: 2023년 5월 기준 소요 경과 햇수 계산 ex) 202109 -> 2 (2023 - 2021)

주행거리: 숫자로만 표현 12,345km -> 12345

배기량 : cc와 심표 제거



데이터 처리 및 해석

Unnamed: 0		모델명	연식	주행거리	연료	배기량	변속기	연비	최고출력	최대토크	차량중량	구동방식	보험처리	중고가격
0	0	현대 더 뉴 그랜저 하이브리드 2.4 HEV 르블랑	2021.11 (22년형)	37,983 km	가솔린 하이브리드	2,359 cc (200마력)	자동	16.2km/ℓ	200마력	25.1kg.m	1,675kg	전륜 FF	4	3,700만원
1	1	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	2017.09 (18년형)	60,000 km	가솔린 하이브리드	2,359 cc (199마력)	자동	16.2km/ℓ	199마력	26.7kg.m	1,675kg	전륜 FF	23	2,050만원
2	2	현대 더 뉴 그랜저 하이브리드 2.4 HEV 익스클루시브	2020.03	43,818 km	가솔린 하이브리드	2,359 cc (159마력)	자동	16.2km/ℓ	159마력	21kg.m	1,675kg	전륜 FF	2	3,469만원
3	3	현대 그랜저IG 2.4 프리미엄	2018.01	61,406 km	가솔린	2,359 cc (190마력)	자동	11.2km/ℓ	190마력	24.6kg.m	1,550kg	전륜 FF	0	1,940만원
4	4	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	2018.10 (19년형)	56,098 km	가솔린 하이브리드	2,359 cc (199마력)	자동	16.2km/ℓ	199마력	26.7kg.m	1,675kg	전륜 FF	0	2,800만원

범주형 변수 (연료, 변속기, 구동방식 종류)

연료 : 'LPG' = 0, '가솔린' = 1, '디젤' = 2. '전기' = 3, 나머지는 4

변속기 : 자동 = 1 , 수동 = 0

구동방식: ' 4WD ' or ' AWD' (사륜구동) = 1 , 전륜이나 후륜 (이륜구동) = 0



데이터 처리 및 해석

Unnamed: 0		모델명	연식	주행거리	연료	배기량	변속기	연비	최고출력	최대토크	차량중량	구동방식	보험처리	중고가격
0	0	현대 더 뉴 그랜저 하이브리드 2.4 HEV 르블랑	2021.11 (22년형)	37,983 km	가솔린 하이브리드	2,359 cc (200마력)	자동	16.2km/ℓ	200마력	25.1kg.m	1,675kg	전륜 FF	4	3,700만원
1	1	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	2017.09 (18년형)	60,000 km	가솔린 하이브리드	2,359 cc (199마력)	자동	16.2km/ℓ	199마력	26.7kg.m	1,675kg	전륜 FF	23	2,050만원
2	2	현대 더 뉴 그랜저 하이브리드 2.4 HEV 익스클루시브	2020.03	43,818 km	가솔린 하이브리드	2,359 cc (159마력)	자동	16.2km/ℓ	159마력	21kg.m	1,675kg	전륜 FF	2	3,469만원
3	3	현대 그랜저IG 2.4 프리미엄	2018.01	61,406 km	가솔린	2,359 cc (190마력)	자동	11.2km/ℓ	190마력	24.6kg.m	1,550kg	전륜 FF	0	1,940만원
4	4	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	2018.10 (19년형)	56,098 km	가솔린 하이브리드	2,359 cc (199마력)	자동	16.2km/ℓ	199마력	26.7kg.m	1,675kg	전륜 FF	0	2,800만원

중고가격 : '0원', '10500000만원', '가격상담', '계약' 행 제거

연비 : 단위 km를 제거하고 소수점 아래 한자리 까지

최고 출력 : 단위제거 후 숫자만 표시. 200마력 -> 200

최대 토크 : 단위 제거

차량 중량 : 십표, 단위 제거



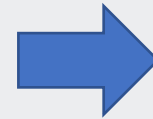
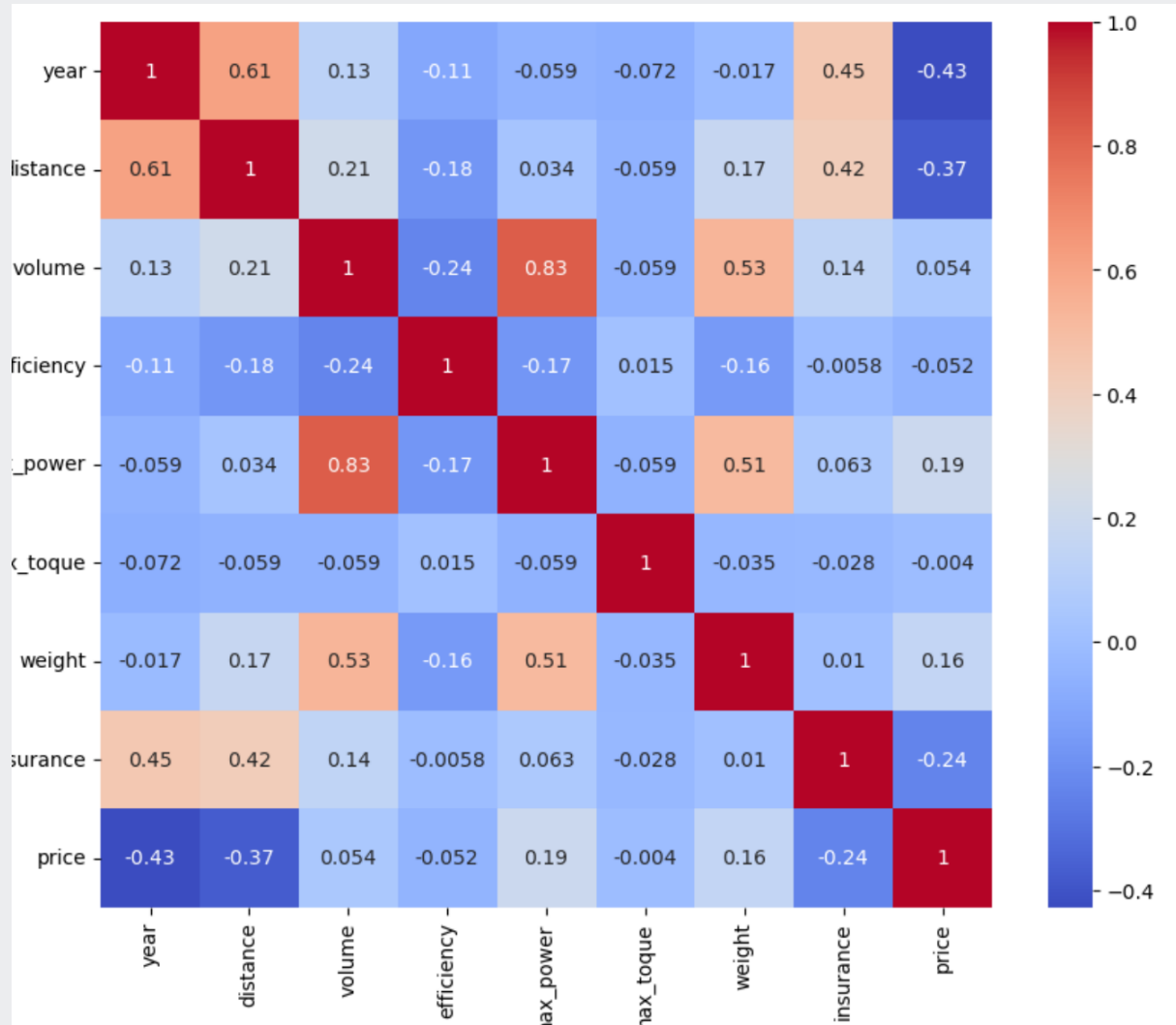
< 처리 후 데이터 >

	model	year	distance	fuel	volume	auto	efficiency	max_power	max_toque	weight	method	insurance	is_genesis	price
0	현대 더 뉴 그랜저 하이브리드 2.4 HEV 르블랑	1.0	37983.0	1.0	2359.0	1.0	16.2	200.0	25.1	1675.0	0.0	4.0	0.0	3700.0
1	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	5.0	60000.0	1.0	2359.0	1.0	16.2	199.0	26.7	1675.0	0.0	23.0	0.0	2050.0
2	현대 더 뉴 그랜저 하이브리드 2.4 HEV 익스클루시브	3.0	43818.0	1.0	2359.0	1.0	16.2	159.0	21.0	1675.0	0.0	2.0	0.0	3469.0
3	현대 그랜저IG 2.4 프리미엄	5.0	61406.0	1.0	2359.0	1.0	11.2	190.0	24.6	1550.0	0.0	0.0	0.0	1940.0
4	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	4.0	56098.0	1.0	2359.0	1.0	16.2	199.0	26.7	1675.0	0.0	0.0	0.0	2800.0
...
1096	기아 K5 2.4 프레스티지	12.0	143667.0	1.0	2359.0	1.0	13.0	201.0	25.5	1470.0	0.0	4.0	0.0	630.0
1097	기아 K5 하이브리드 2세대 2.0 HEV 노블레스 스페셜	4.0	49500.0	1.0	1999.0	1.0	17.0	191.0	27.0	1600.0	0.0	3.0	0.0	2250.0
1098	기아 K5 2.0 프레스티지	12.0	161755.0	1.0	1998.0	1.0	13.0	165.0	20.2	1415.0	0.0	4.0	0.0	460.0
1099	기아 K5 하이브리드 2세대 2.0 HEV 프레스티지	6.0	83723.0	1.0	1999.0	1.0	17.5	191.0	27.0	1580.0	0.0	2.0	0.0	1690.0
1102	기아 K5 2세대 2.0 MX 럭셔리	7.0	75235.0	1.0	1999.0	1.0	12.6	168.0	20.5	1460.0	0.0	7.0	0.0	799.0

774 rows × 14 columns



< 연속형 변수들 간 상관관계 >



중고차 가격과 연식, 주행거리 간 상관관계 0

배기량 - 최고출력간 높은 상관 관계 (0.81)
연식 - 주행거리간 높은 상관 관계 (0.61)



< 다중공선성 VIF 측정 >

	VIF Factor	features
0	6.403478	year
1	5.714142	distance
2	44.628351	volume
3	8.632383	efficiency
4	41.947636	max_power
5	1.046931	max_toque
6	26.299429	weight
7	2.375323	insurance
8	2.668645	price

Volume 제거

Weight 제거



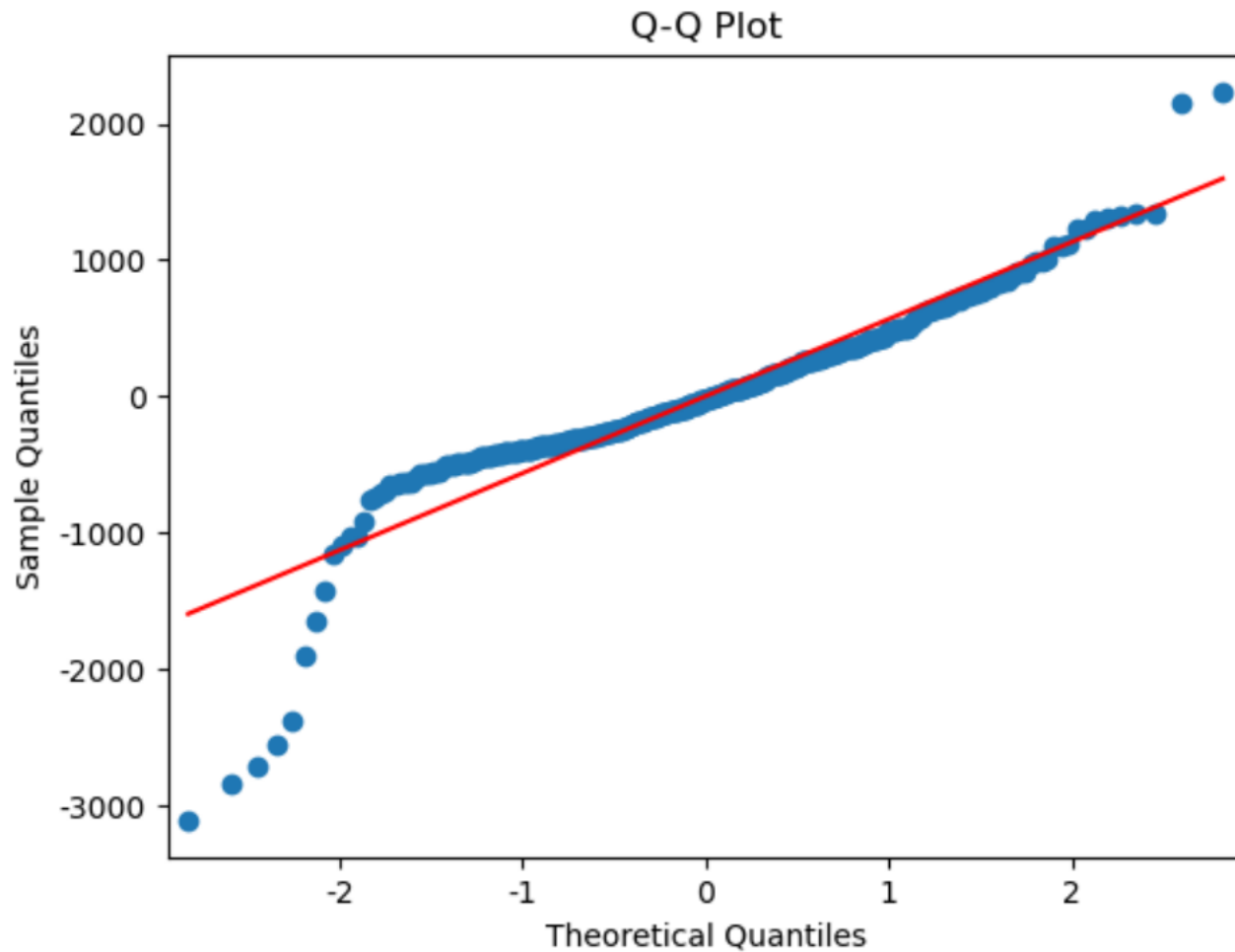
	VIF Factor	features
0	6.187143	year
1	5.142523	distance
2	6.761878	efficiency
3	8.301572	max_power
4	1.044178	max_toque
5	2.348343	insurance
6	2.498392	price

'배기량', '최고 출력', '차량중량'이 높은 VIF

'배기량'과 '차량중량' 열을 제거하여 다중공선성을 낮춤



< Q-Q plot을 이용한 정규성 검사 >



대체적으로 정규성 가정을 만족

대각선에서 크게 벗어난 이상치도 존재



데이터 처리 및 해석

< 이상치 제거 >

```

model year distance fuel auto ₩
0   현대 더 뉴 그랜저 하이브리드 2.4 HEV 르블랑 1.0 37983.0 1.0 1.0
2   현대 더 뉴 그랜저 하이브리드 2.4 HEV 익스클루시브 3.0 43818.0 1.0 1.0
3   현대 그랜저IG 2.4 프리미엄 5.0 61406.0 1.0 1.0
4   현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브 4.0 56098.0 1.0 1.0
5   현대 그랜저IG 2.4 모던 5.0 47882.0 1.0 1.0
...
1096   기아 K5 2.4 프레스티지 12.0 143667.0 1.0 1.0
1097   기아 K5 하이브리드 2세대 2.0 HEV 노블레스 스페셜 4.0 49500.0 1.0 1.0
1098   기아 K5 2.0 프레스티지 12.0 161755.0 1.0 1.0
1099   기아 K5 하이브리드 2세대 2.0 HEV 프레스티지 6.0 83723.0 1.0 1.0
1102   기아 K5 2세대 2.0 MX 럭셔리 7.0 75235.0 1.0 1.0

```

```

efficiency max_power max_toque method insurance is_genesis price
0   16.2 200.0 25.1 0.0 4.0 0.0 3700.0
2   16.2 159.0 21.0 0.0 2.0 0.0 3469.0
3   11.2 190.0 24.6 0.0 0.0 0.0 1940.0
4   16.2 199.0 26.7 0.0 0.0 0.0 2800.0
5   11.2 190.0 24.6 0.0 0.0 0.0 2099.0
...
1096 13.0 201.0 25.5 0.0 4.0 0.0 630.0
1097 17.0 191.0 27.0 0.0 3.0 0.0 2250.0
1098 13.0 165.0 20.2 0.0 4.0 0.0 460.0
1099 17.5 191.0 27.0 0.0 2.0 0.0 1690.0
1102 12.6 168.0 20.5 0.0 7.0 0.0 799.0

```

[527 rows x 12 columns]

데이터의 상위 25%, 하위 25% 제거

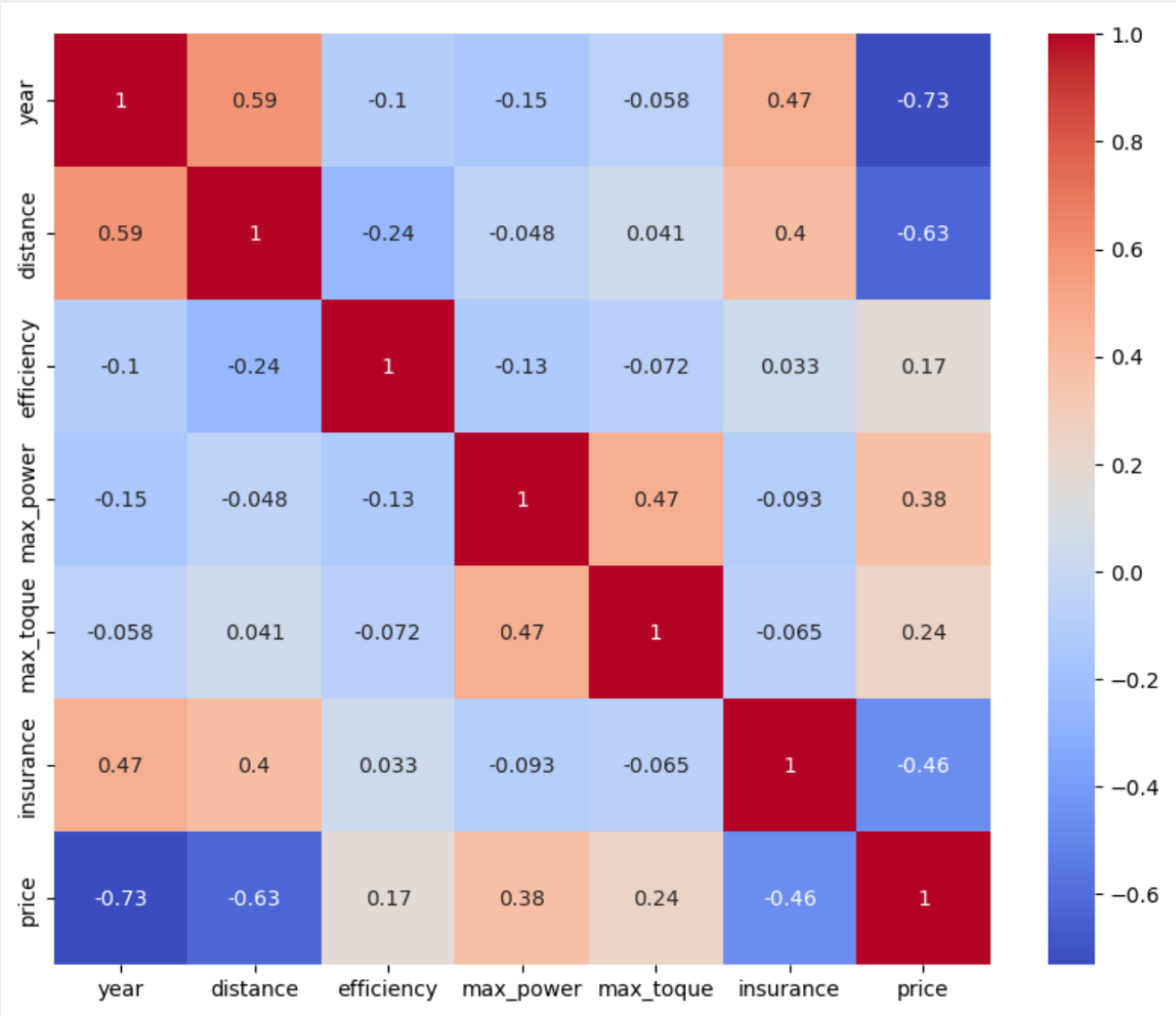
798개 -> 527개

	year	distance	fuel	auto	efficiency	max_power	max_toque	method	insurance	is_genesis	price
count	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000
mean	6.432638	93577.305503	1.277040	0.988615	12.180266	184.290323	31.214991	0.043643	2.540797	0.018975	1614.151803
std	3.871353	52937.006528	0.699745	0.106193	2.269392	29.466053	10.286093	0.204494	2.522474	0.136567	1014.137330
min	1.000000	5.000000	0.000000	0.000000	7.300000	106.000000	14.000000	0.000000	0.000000	0.000000	40.000000
25%	3.000000	49750.000000	1.000000	1.000000	10.600000	168.000000	24.600000	0.000000	0.000000	0.000000	799.000000
50%	5.000000	85800.000000	1.000000	1.000000	11.900000	190.000000	27.000000	0.000000	2.000000	0.000000	1390.000000
75%	9.000000	131889.500000	2.000000	1.000000	13.700000	202.000000	41.000000	0.000000	4.000000	0.000000	2350.000000
max	20.000000	255056.000000	4.000000	1.000000	17.800000	262.000000	46.000000	1.000000	10.000000	1.000000	4690.000000



데이터 처리 및 해석

< 다중공선성, 이상치 처리 후 산점도 >



중고가격, 연식 -0.73 높은 상관 관계

중고가격, 주행거리 -0.63 높은 상관 관계



‘연식’ 과 ‘주행거리’는 ‘중고가격’과 높은 상관성



<여러 방법의 Regression 사용>

1. Linear Regression (선형 회귀)

- 입력 특성과 타겟 변수 간의 선형 관계를 모델링
- 주어진 입력 특성에 대해 선형 방정식을 학습하여 예측을 수행

2. Ridge Regression (릿지 회귀)

- 선형 회귀의 한 종류
- L2 규제를 사용해 모델의 복잡도를 제어. 가중치의 크기를 제한하여 과적합 방지. 일반화 성능 향상

3. Lasso Regression (라쏘 회귀)

- 선형 회귀의 한 종류
- L1 규제를 사용하여 모델의 특성 선택을 수행. 가중치의 일부를 0으로 만들어 특성 선택을 수행. 모델을 더 해석 가능하게 만듦

L1 Norm : 서로 다른 두 벡터를 나타내는 각 원소들의 차이의 절댓값의 합

L2 Norm : 서로 다른 두 벡터 사이의 유클리드 거리



<여러 방법의 Regression 사용>

4. ElasticNet Regression (엘라스틱넷 회귀)

- 릿지 회귀와 라쏘 회귀를 결합한 모델. L1 규제와 L2 규제를 동시에 적용

5. Gradient Boosting Regression (그래디언트 부스팅 회귀):

- 여러 개의 약한 학습기를 순차적으로 학습하여 예측을 수행하는 앙상블 모델
- 이전 모델의 오차를 보완하는 방향으로 새로운 모델을 학습. 예측 성능을 향상시킴

6. AdaBoost Regression (에이다부스트 회귀):

- 이전 모델이 잘못 예측한 샘플에 가중치를 부여하여 다음 모델을 학습하고, 예측 성능을 향상시킴

7. Support Vector Regression (서포트 벡터 회귀):

- 서포트 벡터 회귀는 서포트 벡터 머신(SVM)을 사용하여 회귀 문제를 해결하는 모델
- 입력 특성 공간에서 서포트 벡터를 찾고, 이를 기반으로 예측을 수행



<DT , 랜덤포레스트 사용>

8. Decision Tree (의사결정나무)

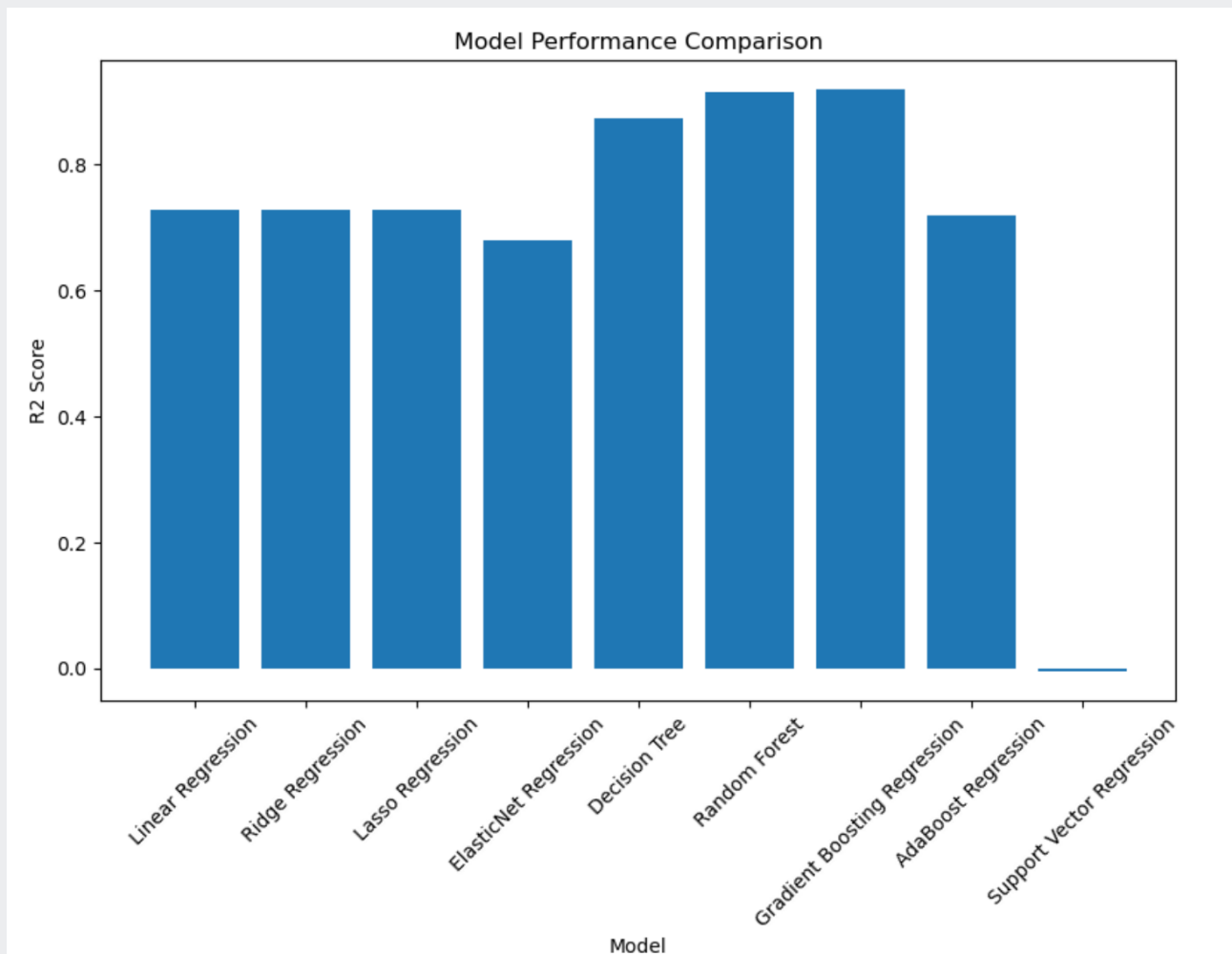
- 입력 특성을 기반으로 타깃 변수를 분류하거나 예측데이터를 분할하는 규칙을 트리 형태로 구성하여 예측을 수행

9. Random Forest (랜덤 포레스트)

- 여러 개의 의사결정나무를 앙상블한 모델로, 각 트리의 예측 결과를 평균 또는 투표하여 최종 예측을 수행
- 의사결정나무의 과적합을 줄이고 일반화 성능을 향상시킬 수 있음



각 모델을 이용한 R2 값 비교



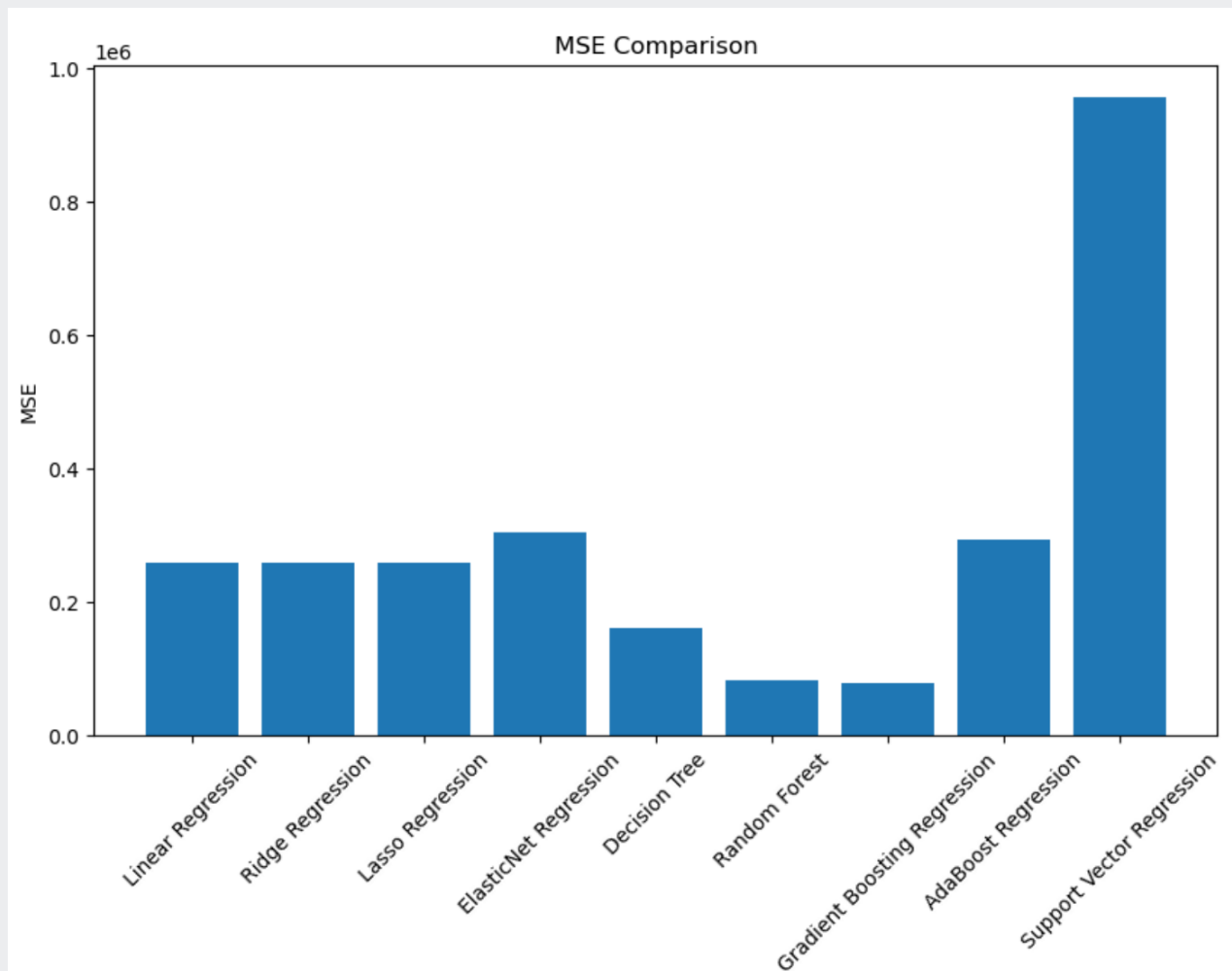
$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Linear Regression: R2 Score = 0.7286
Ridge Regression: R2 Score = 0.7283
Lasso Regression: R2 Score = 0.7276
ElasticNet Regression: R2 Score = 0.6794
Decision Tree: R2 Score = 0.8724
Random Forest: R2 Score = 0.9151
Gradient Boosting Regression: R2 Score = 0.9189
AdaBoost Regression: R2 Score = 0.7197
Support Vector Regression: R2 Score = -0.0053

'Gradient Boosting' 에서 가장 높은 R2 값



각 모델을 이용한 MSE 값 비교



$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Linear Regression: MSE = 258126.7897
Ridge Regression: MSE = 258417.5984
Lasso Regression: MSE = 259068.2246
ElasticNet Regression: MSE = 304918.3786
Decision Tree: MSE = 162157.4528
Random Forest: MSE = 82358.3096
Gradient Boosting Regression: MSE = 79075.1902
AdaBoost Regression: MSE = 294084.9063
Support Vector Regression: MSE = 956253.4568

'Gradient Boosting' 에서 가장 낮은 MSE 값



회귀 분석

	model	year	distance	fuel	auto	efficiency	max_power	max_toque	method	insurance	is_genesis	price
	현대 더 뉴 그랜저 하이브리드 2.4 HEV 르블랑	1.0	37983.0	1.0	1.0	16.2	200.0	25.1	0.0	4.0	0.0	3700.0
	현대 더 뉴 그랜저 하이브리드 2.4 HEV 익스클루시브	3.0	43818.0	1.0	1.0	16.2	159.0	21.0	0.0	2.0	0.0	3469.0
	현대 그랜저IG 2.4 프리미엄	5.0	61406.0	1.0	1.0	11.2	190.0	24.6	0.0	0.0	0.0	1940.0
	현대 그랜저IG 하이브리드 2.4 HEV 익스클루시브	4.0	56098.0	1.0	1.0	16.2	199.0	26.7	0.0	0.0	0.0	2800.0
	현대 그랜저IG 2.4 모던	5.0	47882.0	1.0	1.0	11.2	190.0	24.6	0.0	0.0	0.0	2099.0

	기아 K5 2.4 프레스티지	12.0	143667.0	1.0	1.0	13.0	201.0	25.5	0.0	4.0	0.0	630.0
	기아 K5 하이브리드 2세대 2.0 HEV 노블레스 스페셜	4.0	49500.0	1.0	1.0	17.0	191.0	27.0	0.0	3.0	0.0	2250.0
	기아 K5 2.0 프레스티지	12.0	161755.0	1.0	1.0	13.0	165.0	20.2	0.0	4.0	0.0	460.0
	기아 K5 하이브리드 2세대 2.0 HEV 프레스티지	6.0	83723.0	1.0	1.0	17.5	191.0	27.0	0.0	2.0	0.0	1690.0
	기아 K5 2세대 2.0 MX 럭셔리	7.0	75235.0	1.0	1.0	12.6	168.0	20.5	0.0	7.0	0.0	799.0

이전 데이터에서 293개 차량 종류를
기아차(1) / 현대차(2) 로 나누고
'is_genesis' 변수를 제거하여 수행

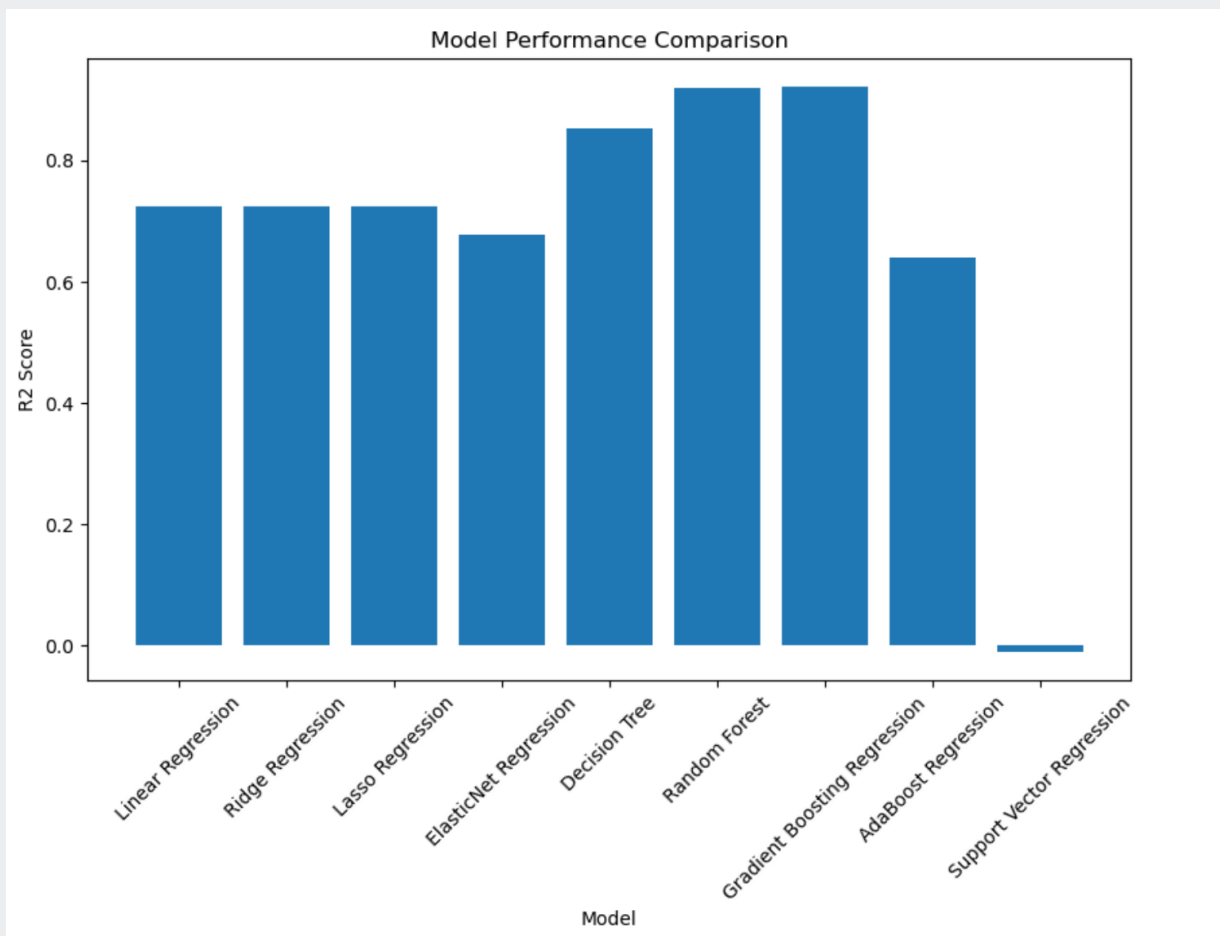
'기아' / '현대' 차 로 구분

	model	year	distance	fuel	auto	efficiency	max_power	max_toque	method	insurance	is_genesis	price
0	현대	1.0	37983.0	1.0	1.0	16.2	200.0	25.1	0.0	4.0	0.0	3700.0
1	현대	3.0	43818.0	1.0	1.0	16.2	159.0	21.0	0.0	2.0	0.0	3469.0
2	현대	5.0	61406.0	1.0	1.0	11.2	190.0	24.6	0.0	0.0	0.0	1940.0
3	현대	4.0	56098.0	1.0	1.0	16.2	199.0	26.7	0.0	0.0	0.0	2800.0
4	현대	5.0	47882.0	1.0	1.0	11.2	190.0	24.6	0.0	0.0	0.0	2099.0
...
522	기아	12.0	143667.0	1.0	1.0	13.0	201.0	25.5	0.0	4.0	0.0	630.0
523	기아	4.0	49500.0	1.0	1.0	17.0	191.0	27.0	0.0	3.0	0.0	2250.0
524	기아	12.0	161755.0	1.0	1.0	13.0	165.0	20.2	0.0	4.0	0.0	460.0
525	기아	6.0	83723.0	1.0	1.0	17.5	191.0	27.0	0.0	2.0	0.0	1690.0
526	기아	7.0	75235.0	1.0	1.0	12.6	168.0	20.5	0.0	7.0	0.0	799.0

527 rows × 12 columns



각 모델을 이용한 R2 값 비교



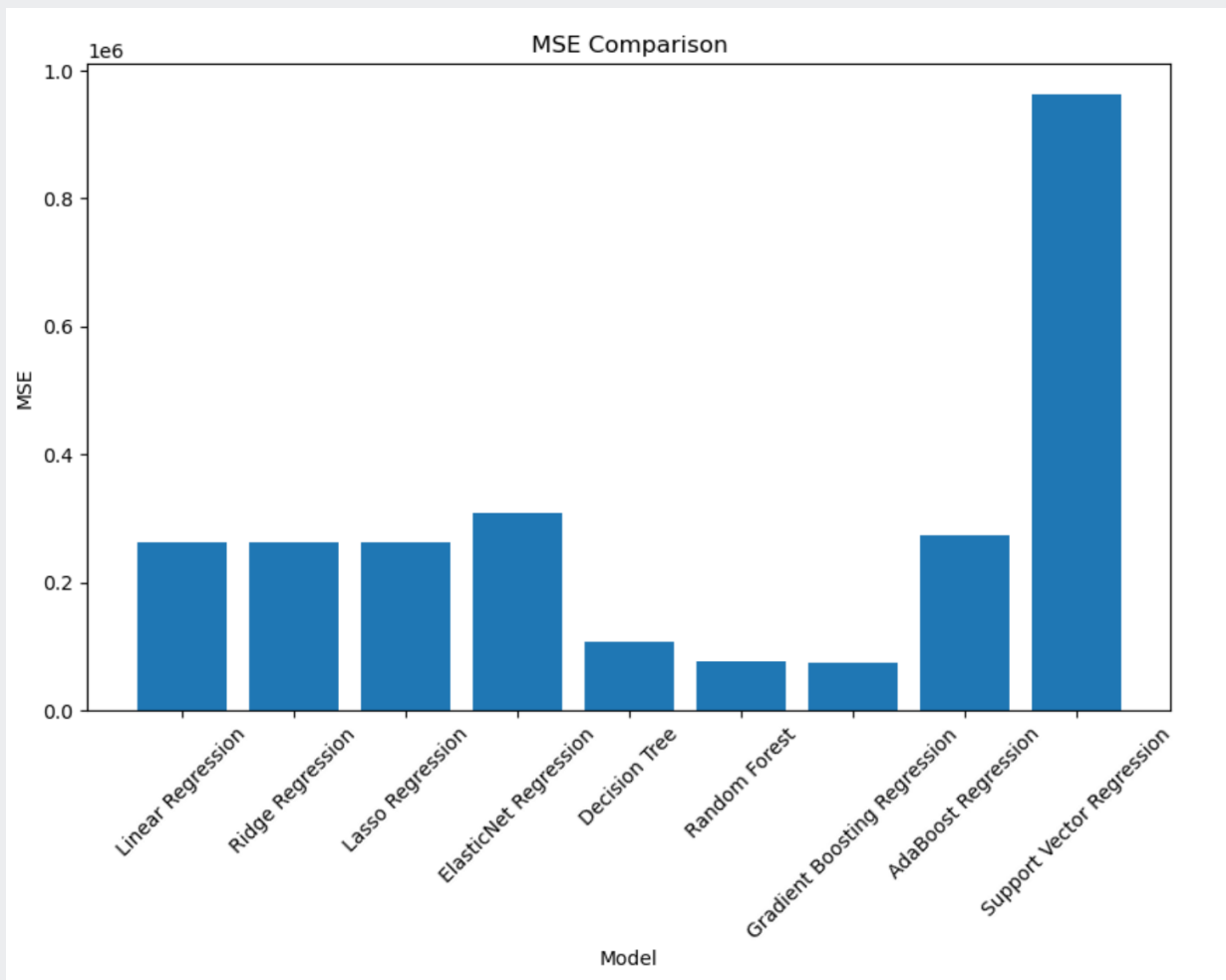
$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Linear Regression: R2 Score = 0.7240
Ridge Regression: R2 Score = 0.7236
Lasso Regression: R2 Score = 0.7232
ElasticNet Regression: R2 Score = 0.6766
Decision Tree: R2 Score = 0.8534
Random Forest: R2 Score = 0.9206
Gradient Boosting Regression: R2 Score = 0.9214
AdaBoost Regression: R2 Score = 0.6389
Support Vector Regression: R2 Score = -0.0116

'Gradient Boosting' 에서 가장 높은 R2 값



각 모델을 이용한 R2 값 비교



$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Linear Regression: MSE = 262574.2086
Ridge Regression: MSE = 262889.7169
Lasso Regression: MSE = 263328.2535
ElasticNet Regression: MSE = 307639.2386
Decision Tree: MSE = 106562.9057
Random Forest: MSE = 77720.4845
Gradient Boosting Regression: MSE = 74914.1266
AdaBoost Regression: MSE = 272916.3340
Support Vector Regression: MSE = 962212.4495

'Gradient Boosting' 에서 가장 낮은 MSE 값



Gradient Boosting 알고리즘의 높은 성능

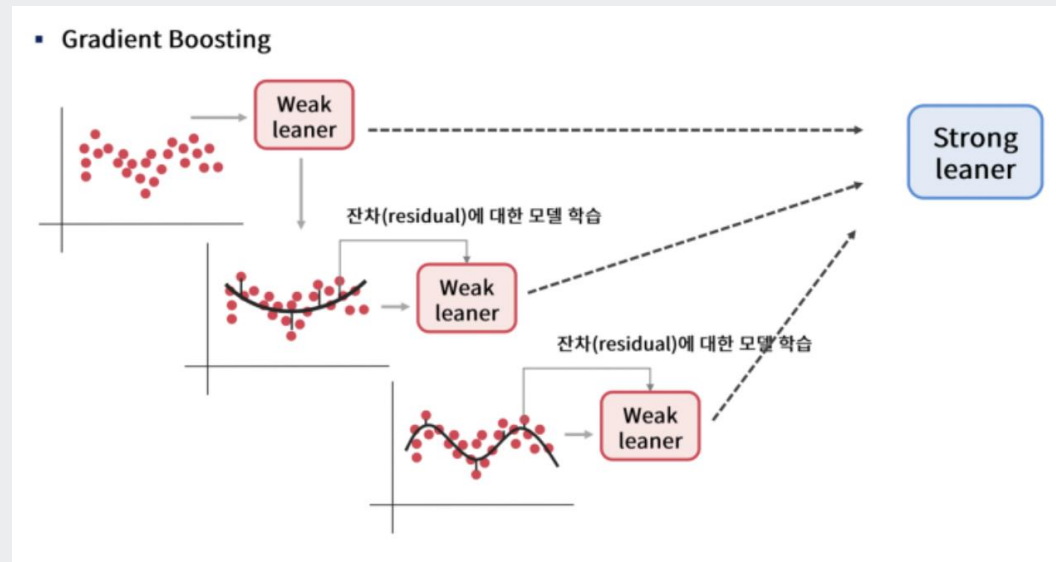
왜 Gradient Boosting 알고리즘이 성능이 가장 좋을까?

Boosting 기법

여러개의 알고리즘이 순차적으로 학습-예측하면서 이전에 학습한 알고리즘의 예측을 통해 다음 알고리즘에 반영

Gradient Boosting

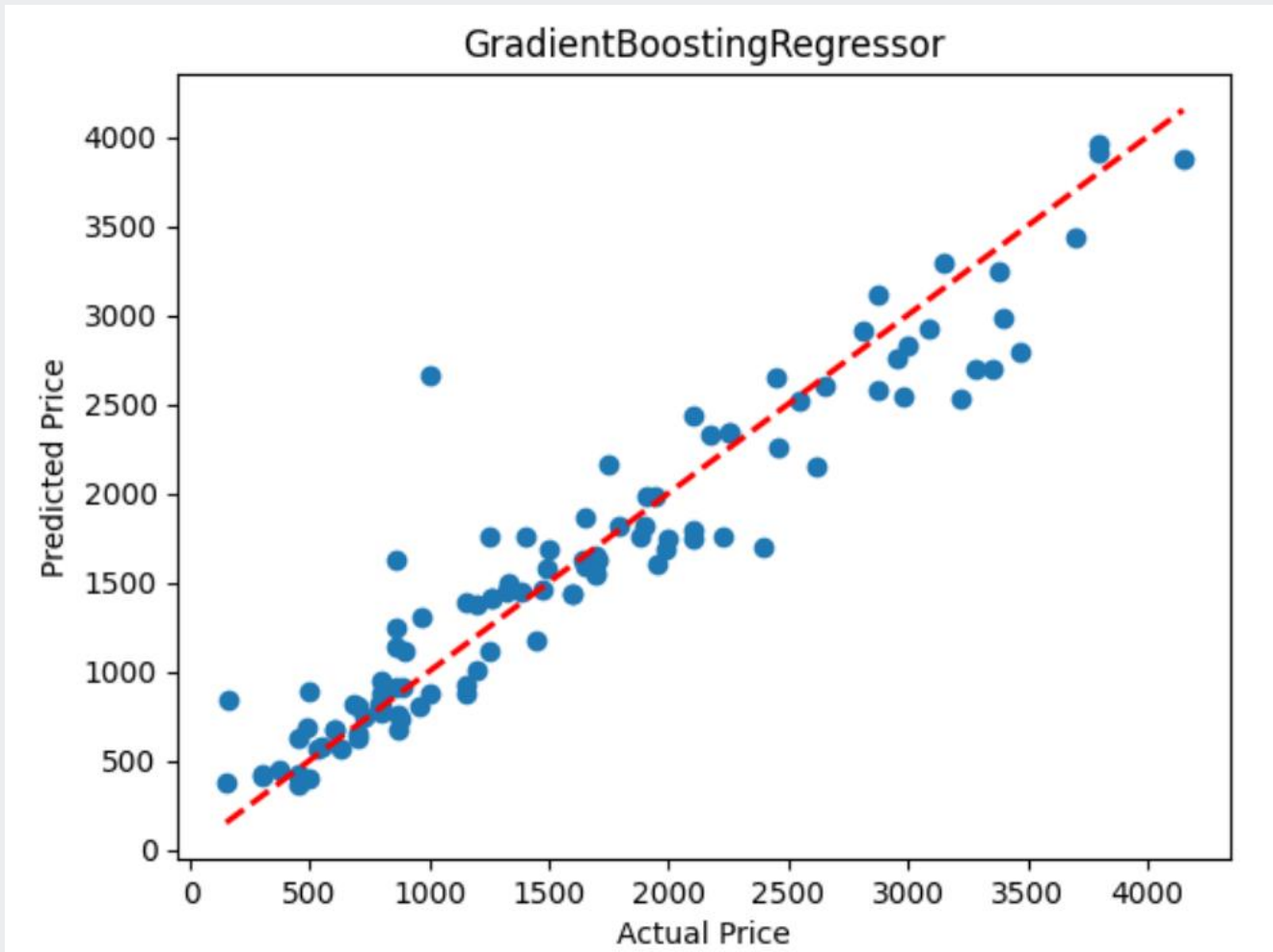
이전 예측기가 만든 잔여오차에 새로운 예측기를 학습 시킴
가중치 업데이트를 경사하강법 기법으로 최적화된 결과 얻음





Gradient Linear Regression

성능이 좋았던 Gradient Linear Regression를 이용하여
더미 변수들인 is_genesis, auto, method, fuel 변수들 제거 후 데이터 스케일링 후 그래프 생성



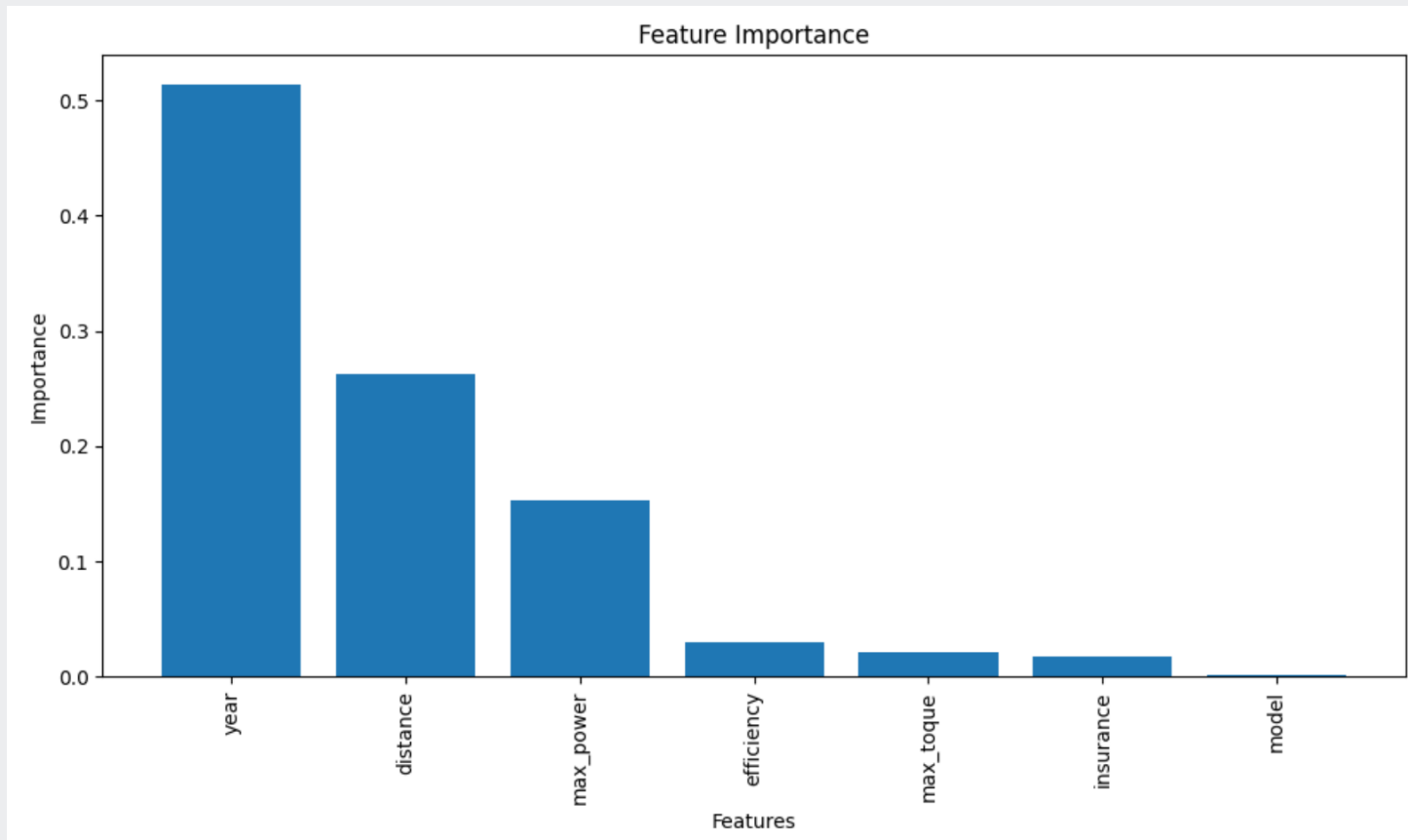
실제 가격과 예측 가격 비교

선형적인 그래프 모양



Gradient Linear Regression

변수 중요도 측정

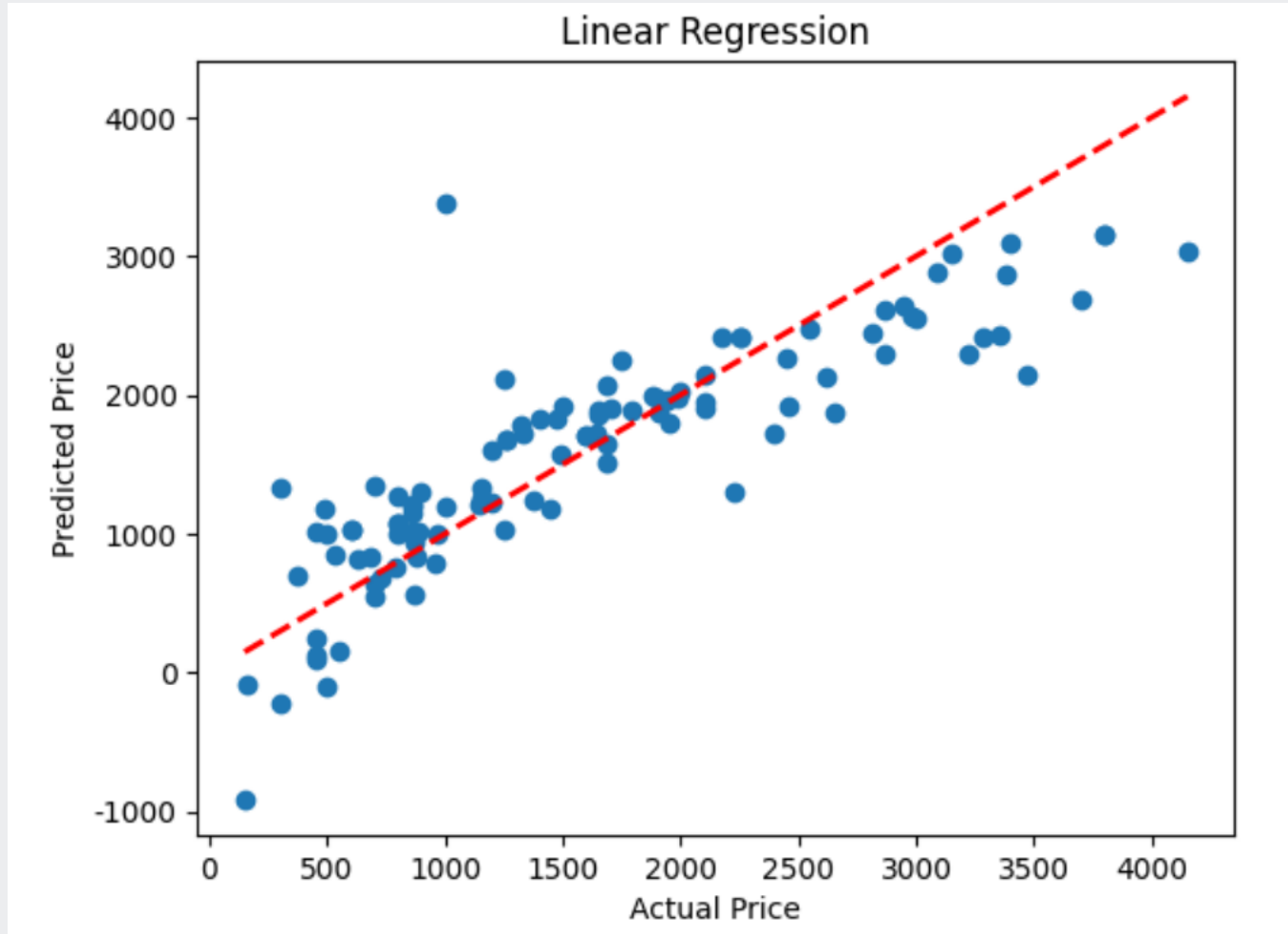


연식이 가장 중요한 변수

Model 변수는 매우 낮은 중요도



Linear Regression



실제 가격과 예측 가격 비교

선형적인 그래프 모양

Linear Regression를 이용하여

더미 변수들인 is_genesis, auto, method, fuel 변수들 제거 후 데이터 스케일링 후 그래프 생성



변수		p-value
model	0.0535	
year	0.0000	
distance		0.0000
efficiency		0.0003
max_power		0.0000
max_torque		0.0057
insurance		0.0002

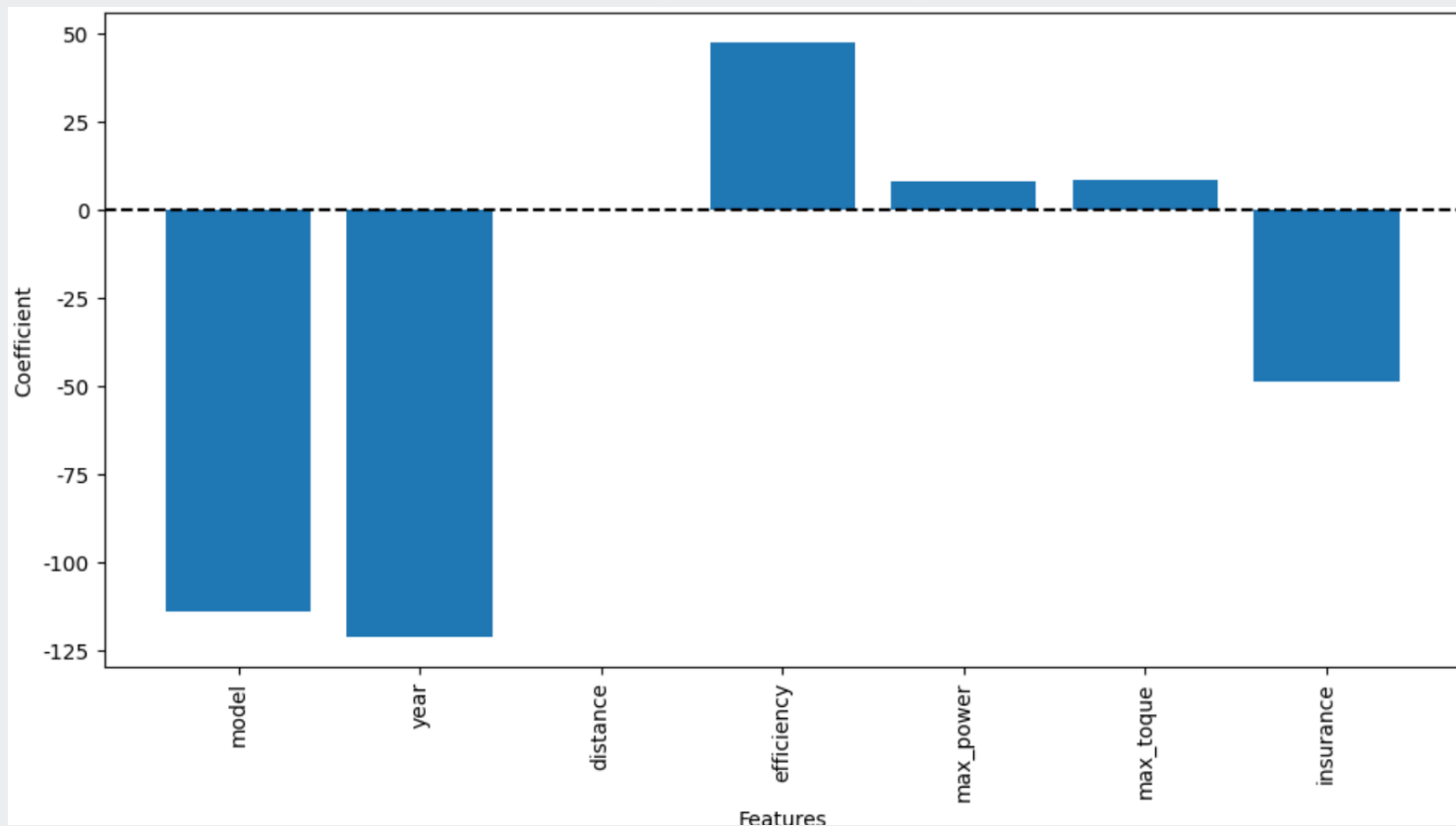
선형회귀분석에서 각각의 변수들은 p-value 값이 낮으므로 대체적으로 유의미함

특히, '연식'과 '주행거리', '배기량'이 가장 유의미함



Linear Regression

Linear Regression 계수 측정



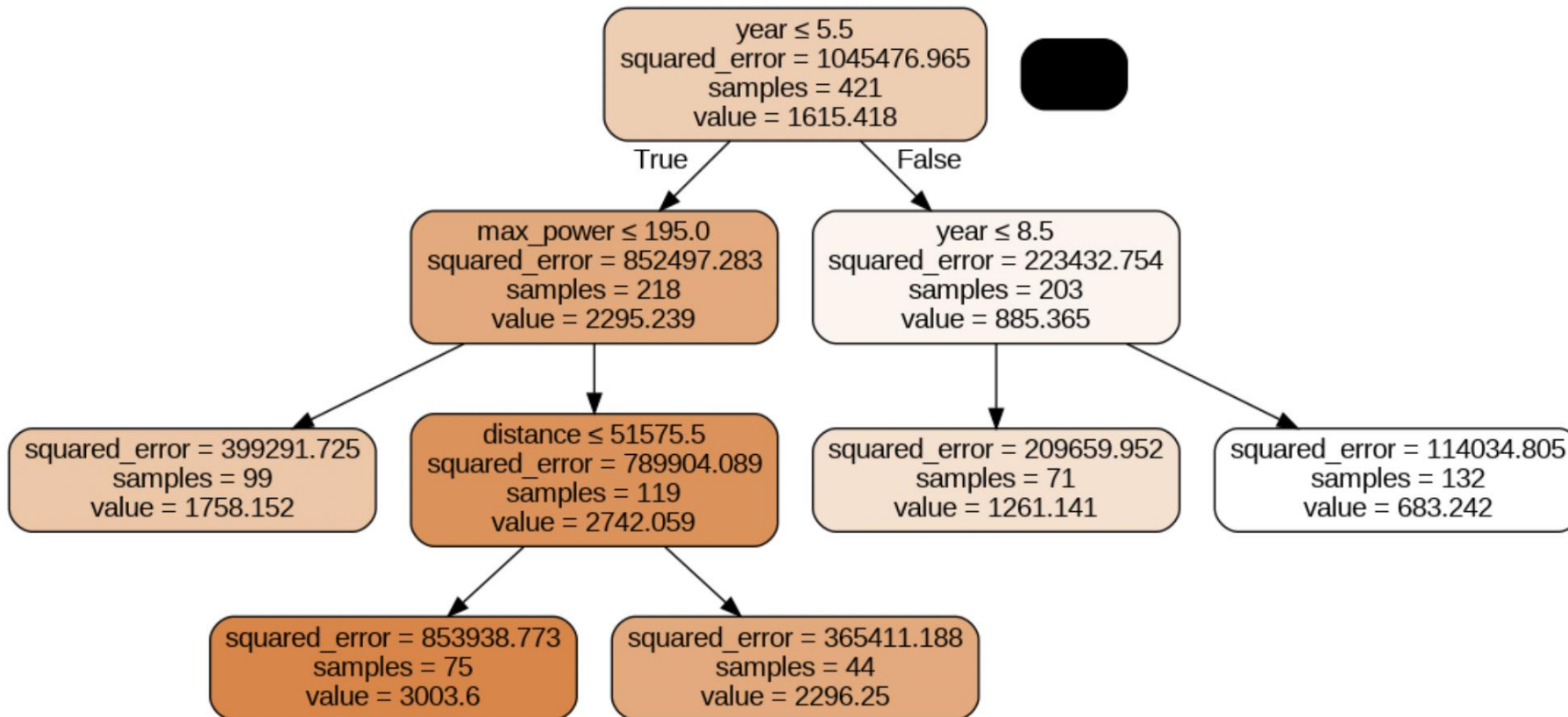
```
Intercept: 767.9977091115721  
model : -113.88633618423331  
year : -121.36737133726247  
distance : -0.004894508254892571  
efficiency : 47.50269513360232  
max_power : 8.353083058261015  
max_toque : 8.557726903531186  
insurance : -48.502028491727366
```

'모델 종류'와 '연식'이 중요한 설명 변수



Decision Tree

Decision Tree 를 이용한 해석



연식이 낮을 수록

주행거리가 짧을수록

최고출력이 높을수록



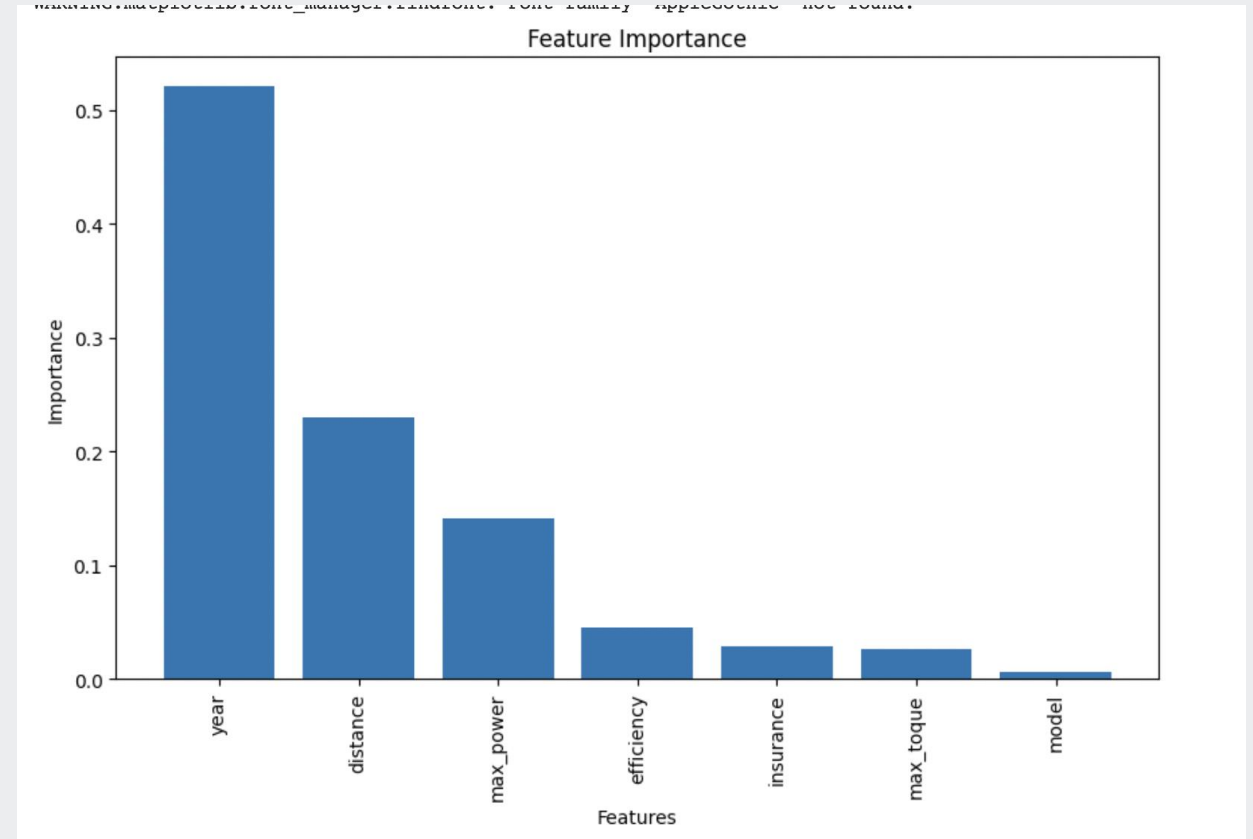
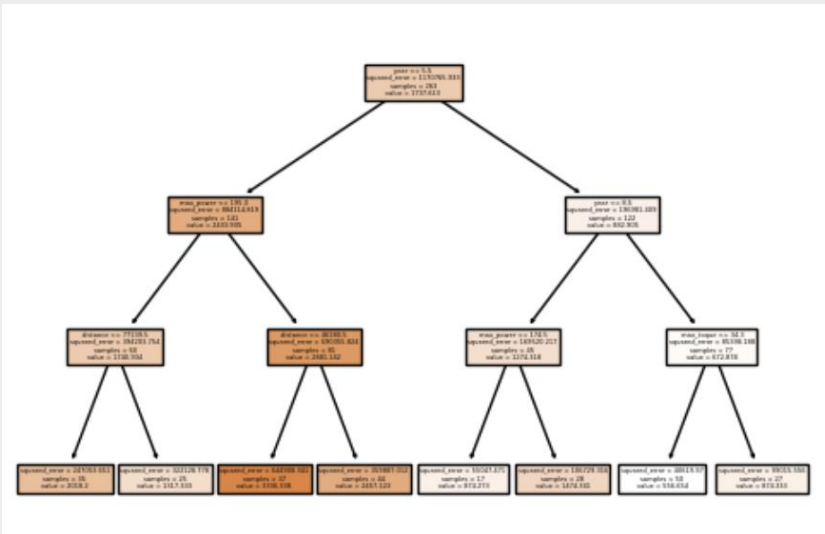
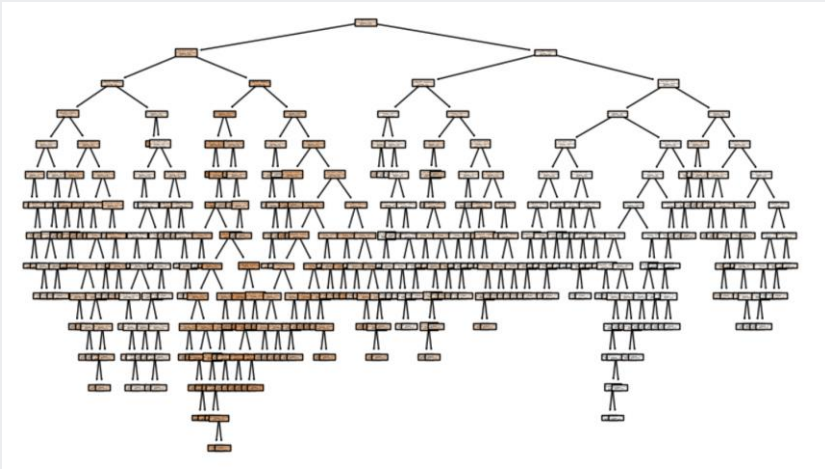
중고가격이 높음

연식이 5.5년 이하, 최고 출력이 195 이상, 주행거리가 51,575km 이하인 차량의 가격이 높음



Random Forest

Random Forest 를 이용한 해석

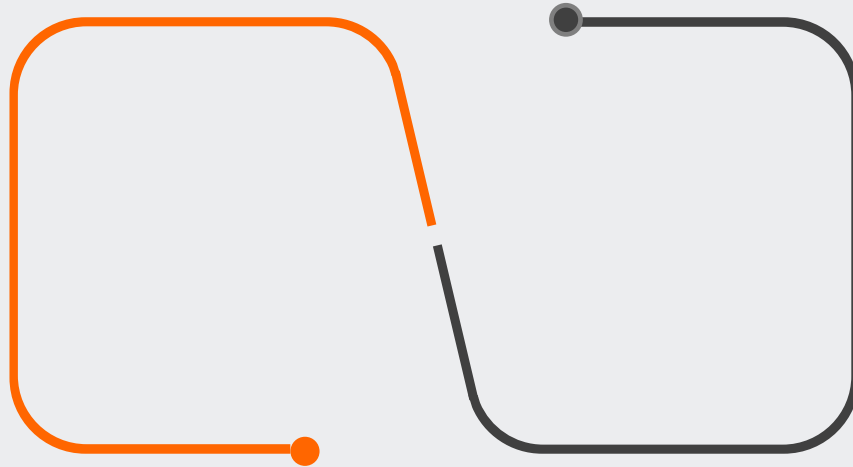


연식이 가장 중요한 변수



Gradient Boosting Model이 낮은 MSE값과 높은 R2 값으로 좋은 성능을 보였고 배기량, 연식, 주행거리 등이 중고차 가격에 영향을 미친다.

'연식' 과 '주행거리'는 중고차 가격에 가장 큰 영향을 끼침



연식이 낮은, 주행거리가 적은 차량이 가격이 높음

감사합니다