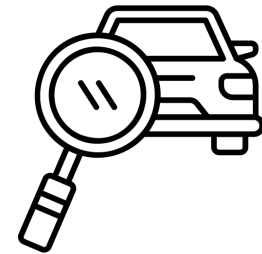


인도 중고차 시장 분석

중고차 가격 예측을 위한 통계 모델 비교

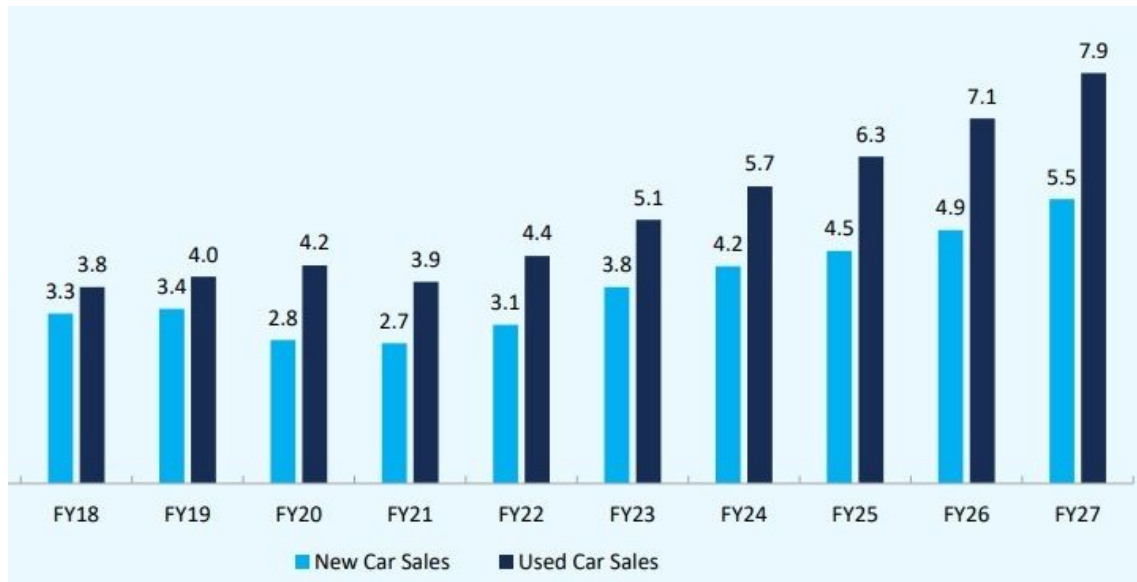


C3조

박경남 김영국 이충현
강민주 이정희 이찬영

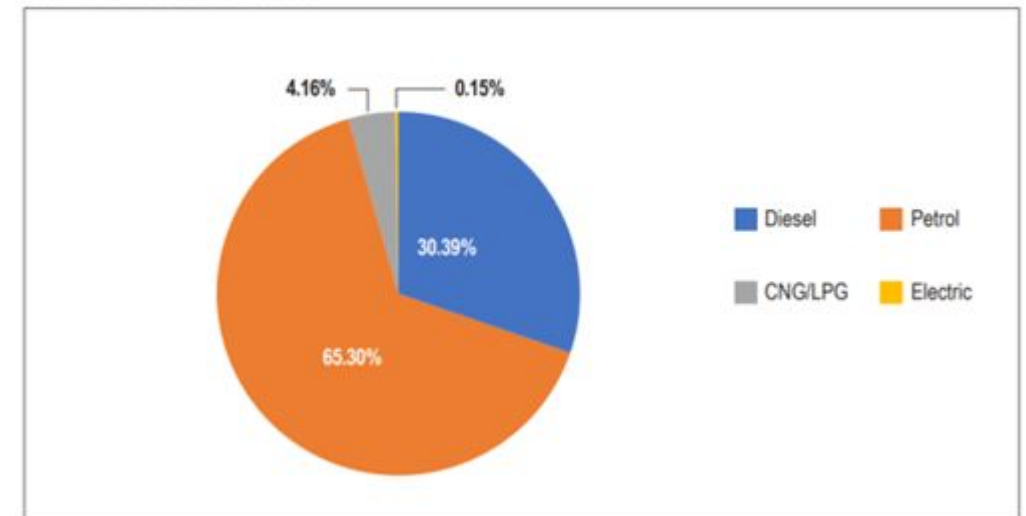
추진 배경

최근 인도에서는 중고차 시장이 신차 시장보다 지속적으로 성장하고 있다. 고객들은 차량의 세부 특성들이 좋으면서 가격이 저렴한 차량을 원하고 있다. 이에따라 POS_CAR 업체는 인도 중고차 시장에 진입을 하기위해 중고차 가격의 핵심영향인자를 도출하여 중고차 가격을 예측하는 모델을 완성하고자 한다.



<https://www.kita.net/board/overseasMarket/overseasMarketDetail.do?no=9863&boardType=1>

Fuel Wise - Used Car Sales



Source: car&bike Analysis

<https://www.kita.net/board/overseasMarket/overseasMarketDetail.do?no=9863&boardType=1>

변수 이름	목표 / 설명변수	데이터 설명
Price	목표변수	중고차 가격
Location	설명변수	자동차를 팔거나 구매할 수 있는 위치
Name	설명변수	자동차의 브랜드와 모델 이름
Year	설명변수	모델의 년도 혹은 버전
Kilometers_Driven	설명변수	이전 소유주의 차량 주행거리(KM)
Fuel_Type	설명변수	자동차의 사용연료의 종류
Transmission	설명변수	자동차의 사용 변속기의 종류
Owner_Type	설명변수	소유권이 직접 소유인지, 중고 소유인지
Mileage	설명변수	자동차 회사가 제공하는 표준 주행거리(KM/L)
Engine	설명변수	엔진의 배기량(cc)
Power	설명변수	엔진의 최대 출력
Seats	설명변수	차의 좌석 수
New_Price	설명변수	뉴모델의 가격

변수	파생 변수	비고
Year	Vehicle_Age	현재 연도 - Year
Location	major_cities	-
	small_cities	-
Name	Brand	Premium
		Regular
Transmission	Manual	-
	Automatic	-
Fuel_Type	CNG	-
	Petrol	-
	Diesel	-
	LPG	샘플 수가 적어 통계적 의미가 부족

변수	파생	비고
OwnerType	fisrt_type	-
	second	-
	third	-
	Fourth & Above	ANOVA 분석 기준 P-value가 높아 유의성이 낮아 제거함

변수	파생	비고
Engine	경 차	≤ 1000 cc
	소형	≤ 1600 cc
	중형	≤ 2500 cc
	대형	> 2500cc

01. 데이터 탐색

- 데이터 전처리
 - 이상치 및 결측치
 - 파생변수 생성
- EDA
 - 히스토그램
 - 상관계수

02. 모델 구축

- 다중선형 회귀분석
- 의사결정나무
- 랜덤 포레스트
- 그래디언트 부스팅

03. 적용 및 개선

- 알고리즘 모델 평가
- 개선안 도출

목적	분석 계획	
	분석 방법	분석 내용
탐색적 분석	Scatter Plot	두 변수 간의 관계를 시각적으로 표현
	Box Plot	데이터의 분포, 중앙값, 사분위수 (IQR), 이상치 등을 시각적으로 요약하는 그래프
	Histogram	연속형 변수의 분포를 막대 형태로 시각화한 그래프
	Heatmap	변수 간 상관관계를 색상으로 표현하는 그래프
예측 모델	다중선형 회귀분석	종속변수와 여러 독립변수 간의 선형적인 관계를 기반으로 예측
	Decision Tree	데이터를 조건에 따라 분할해 예측값을 생성하는 비선형 모델
	Random Forest	여러 개의 결정나무를 앙상블하여 예측 정확도를 향상시키는 모델
	Gradient Boosting	이전 모델의 오류를 보완해가며 순차적으로 모델을 학습시키는 앙상블 기법

결측치 확인

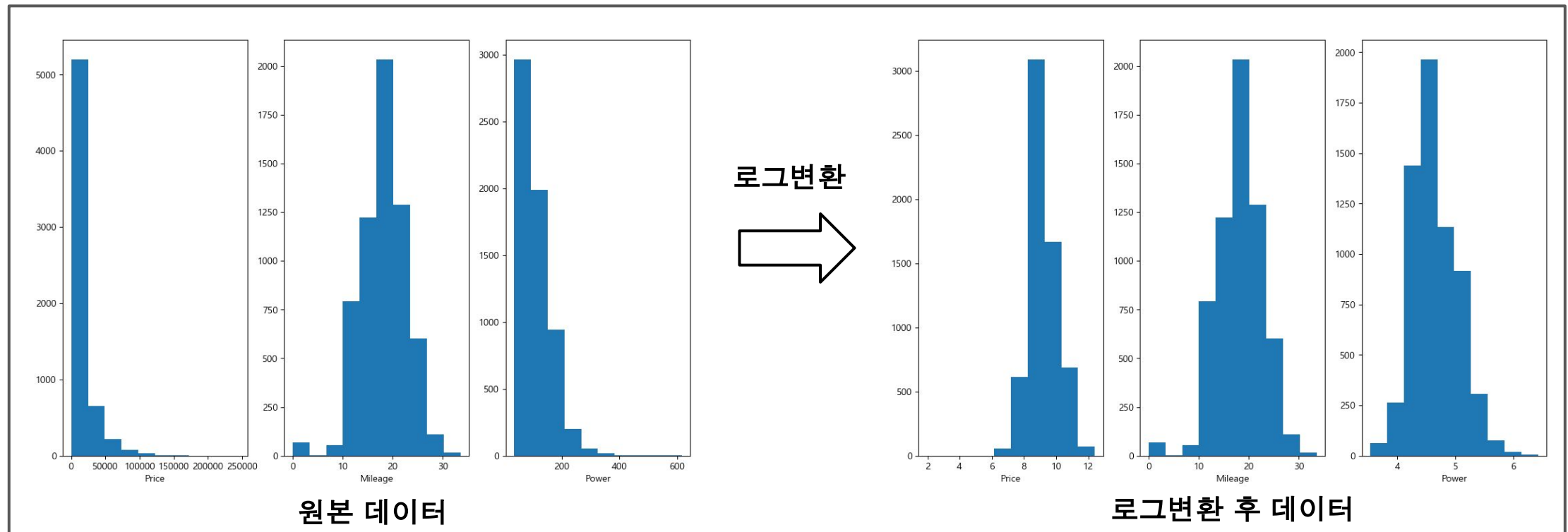
columns	결측치 개수	결측치 처리	판단 이유
Price	1053	제거	일단 제거 후 예측 모델 대체값과 비교
Mileage	2	제거	소수의 결측 영향이 적어 제거
Engine	46	평균값 대체	결측을 평균으로 처리하여 정보 손실 최소화
Power	175	평균값 대체	결측을 평균으로 처리하여 정보 손실 최소화
Seats	53	평균값 대체	결측을 평균으로 처리하여 정보 손실 최소화
New_price	6247	제거	결측 비율이 매우 높아 제거

이상치 확인

columns	이상치 개수	이상치 처리	판단 이유
Seats	1	제거	Seat 값이 0인 차량이 없어 이상치 제거

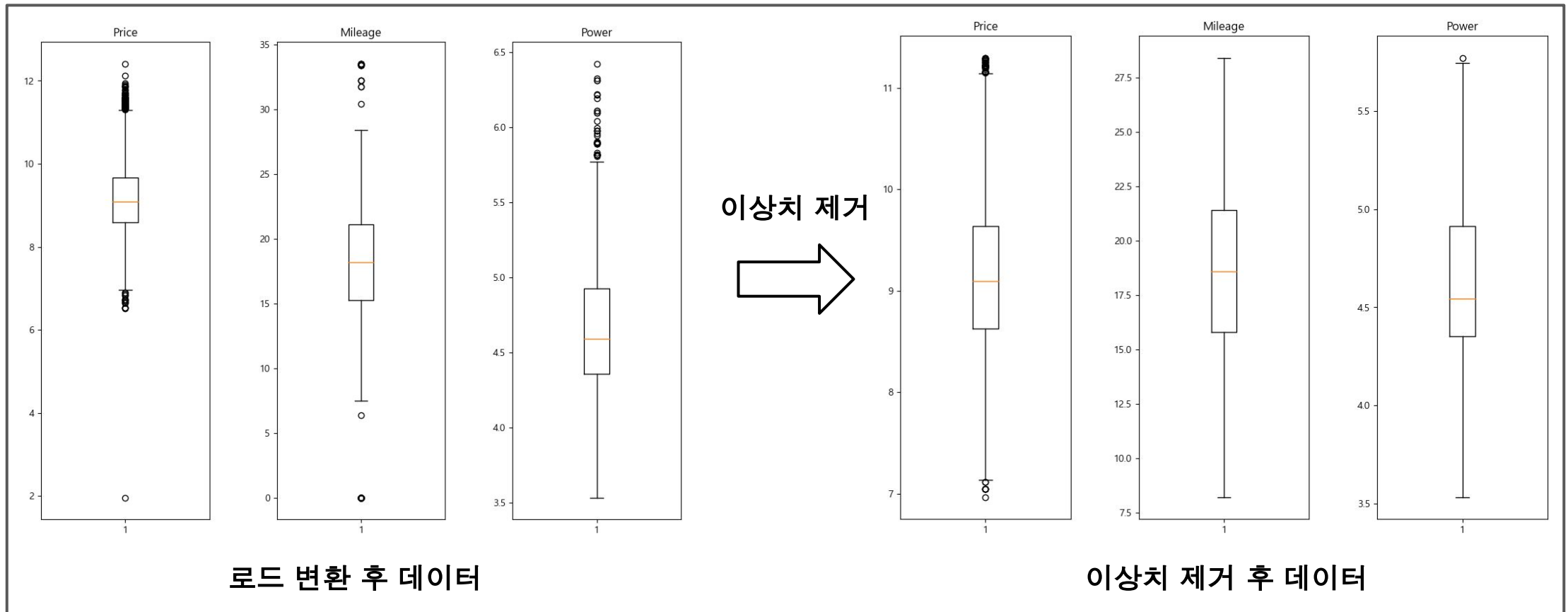
이상치 처리

Price 변수와 Power 변수가 정규분포를 띠는 것이 아닌 왼쪽꼬리분포를 띠고 있음
데이터의 치우침을 방지하기 위해 로그변환을 하여 정규분포를 보이게 함



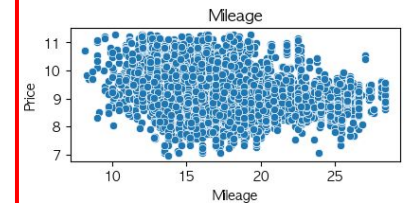
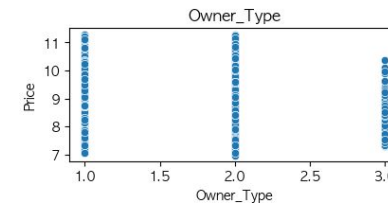
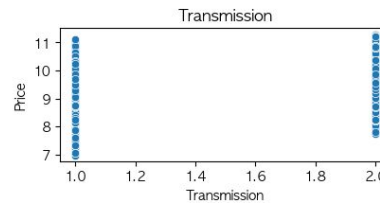
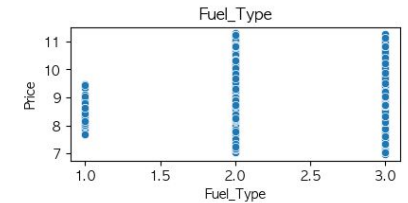
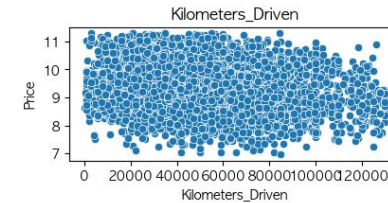
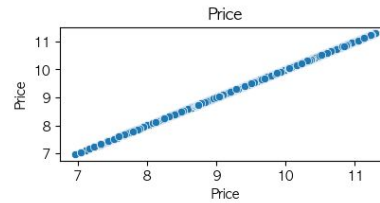
이상치 처리

로그변환 이후 해결되지 않은 극단치에 대해서 IQR을 이용하여 이상치를 제거

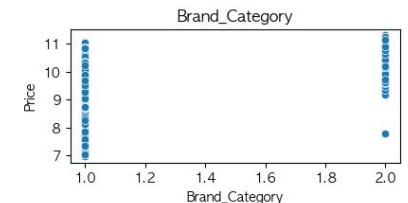
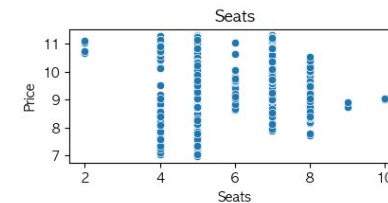
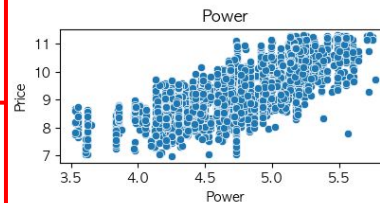


상관관계 분석

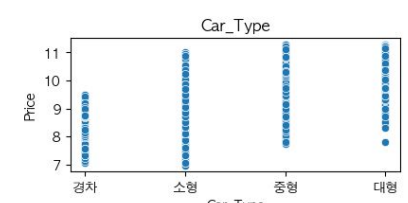
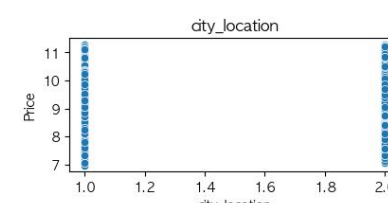
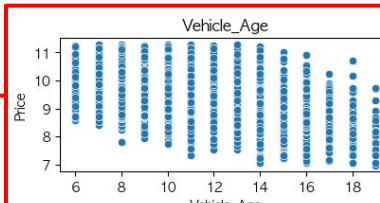
Price-Mileage: 관계성 없음



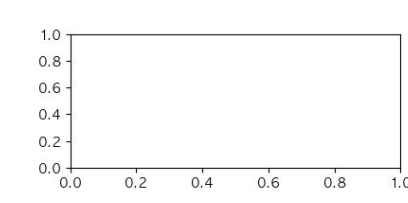
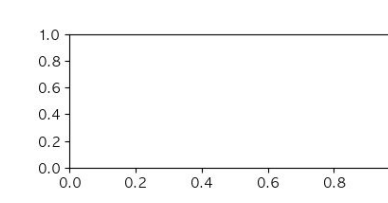
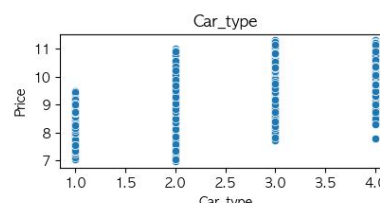
Price-Power: 양의상관관계



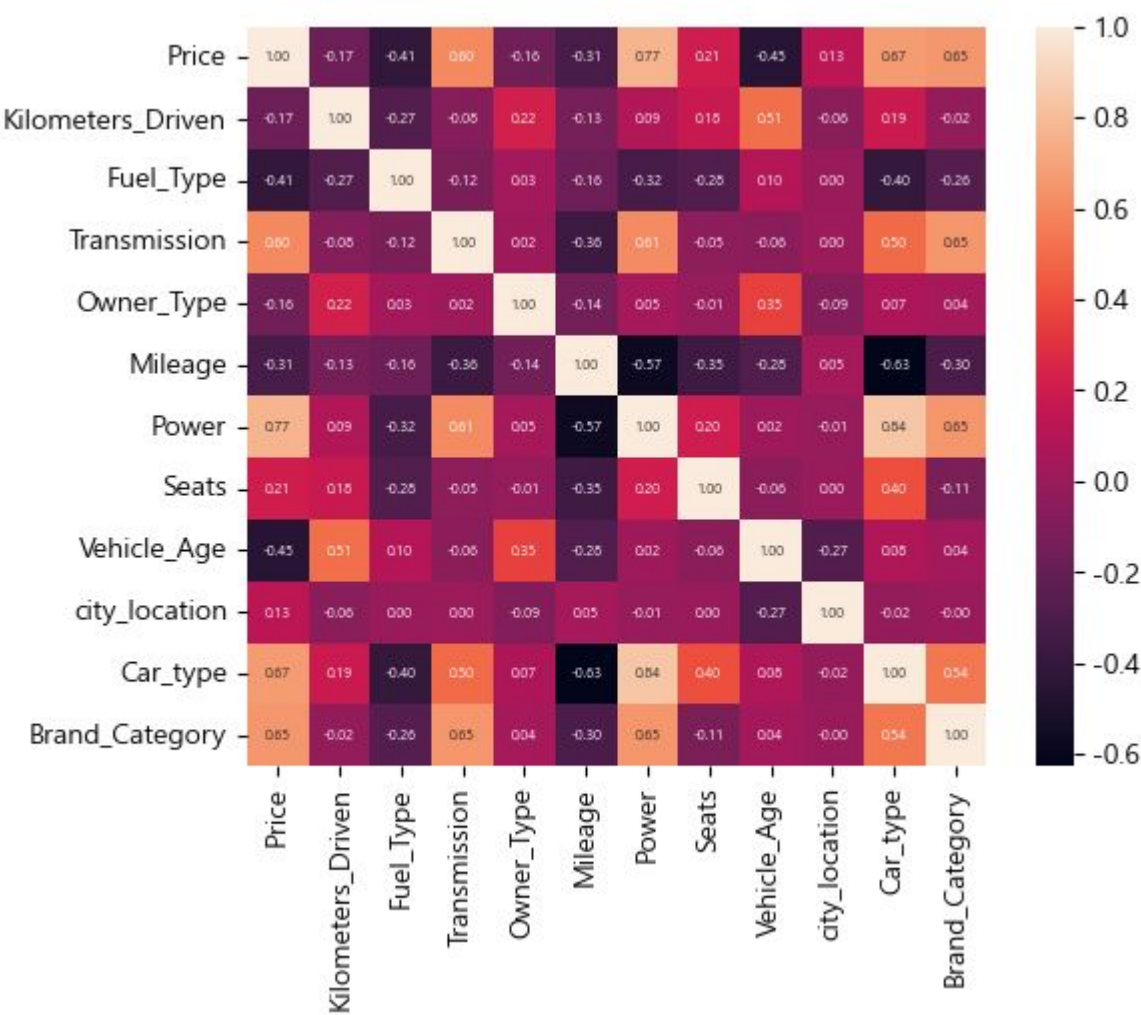
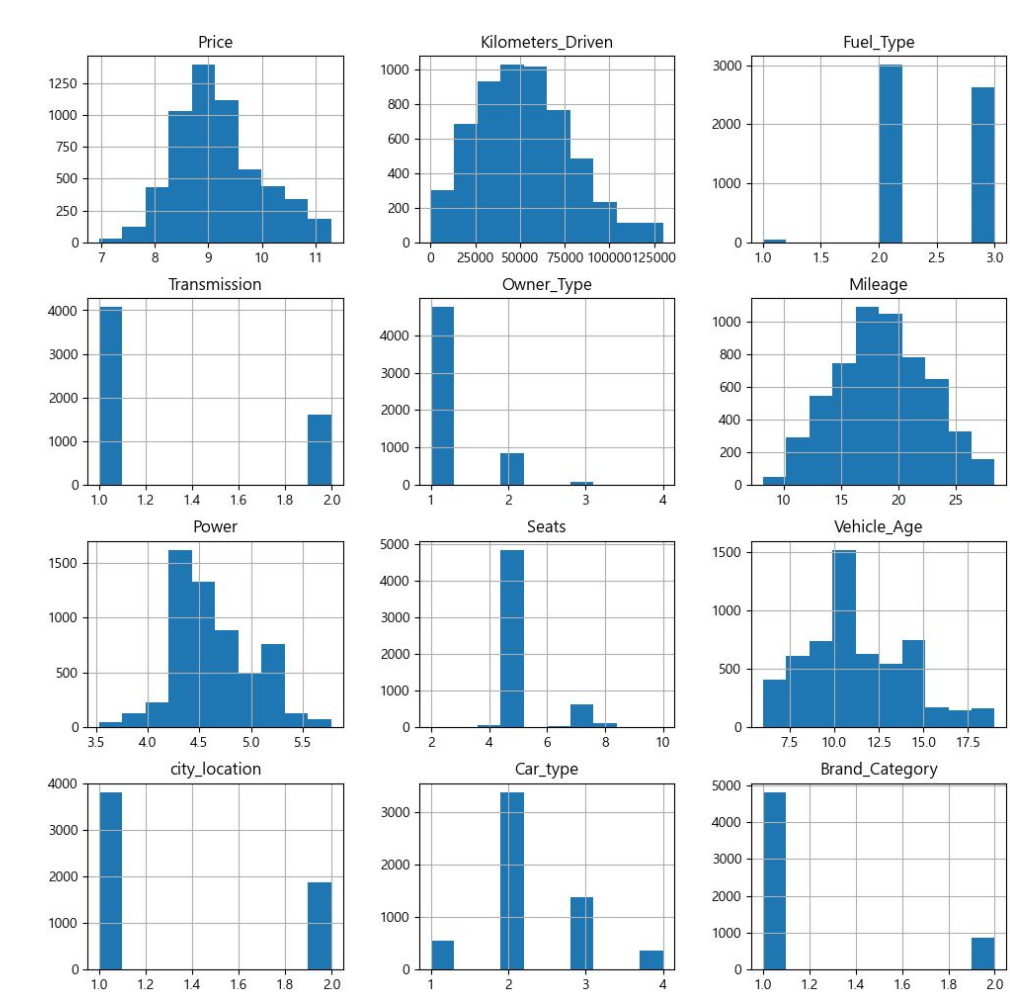
Price-Vehicle age: 음의 상관관계



대부분 범주형 변수의 산점도
그래프로 관계성 파악 불가

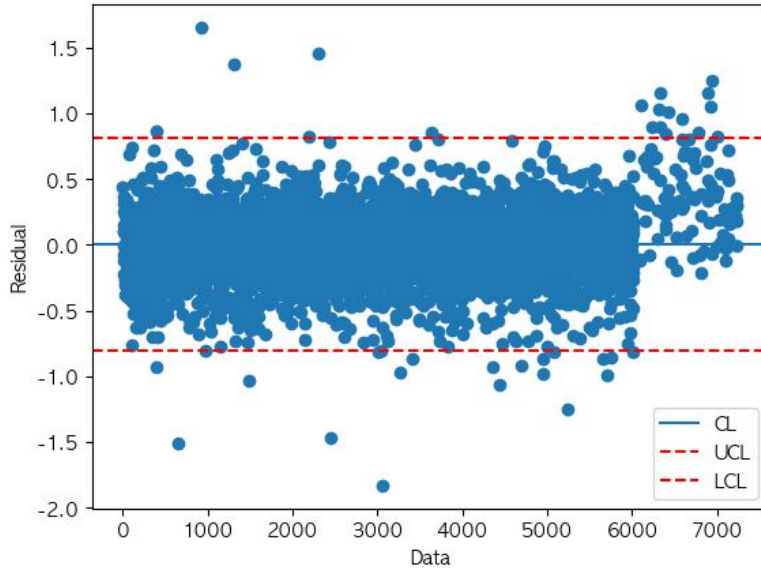


EDA

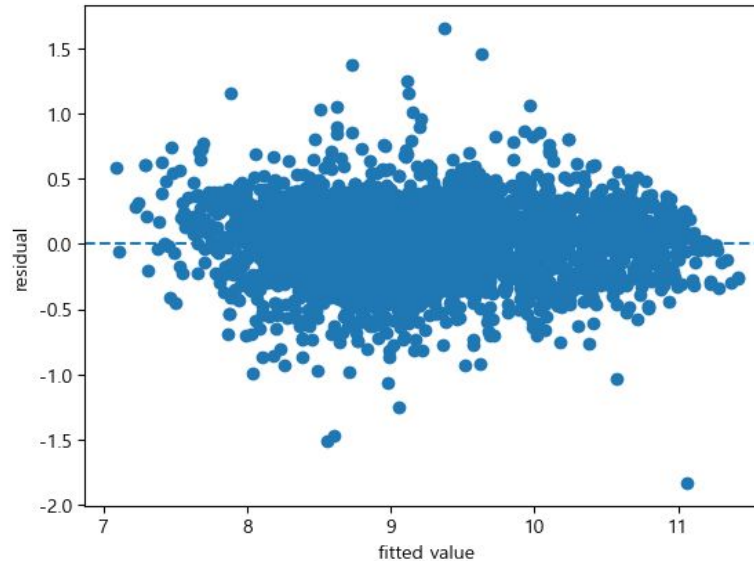


잔차에 대한 가정 검토

독립성 검토

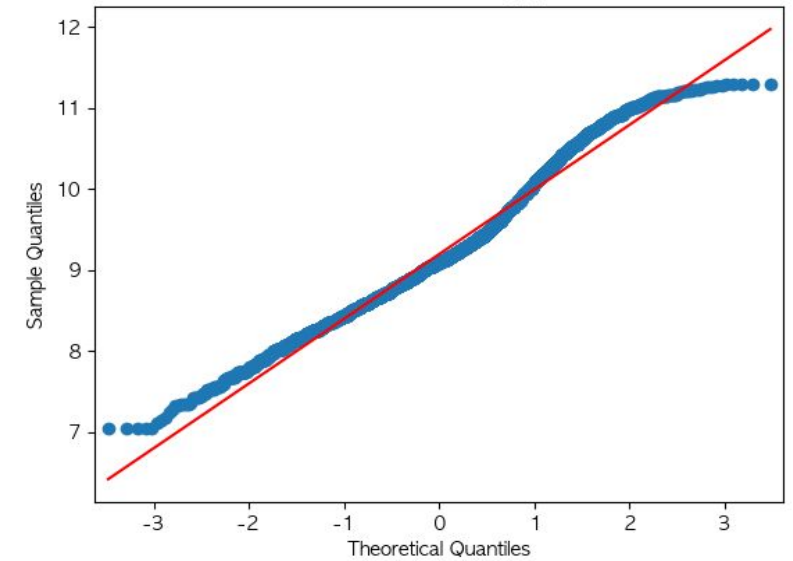


등분산성 검토



정규성 검토

Q-Q Plot - Price_log



독립성 검토 : 데이터가 특정한 패턴을 보이지 않으므로 독립성을 만족한다.

등분산성 검토 : 데이터가 특정한 패턴을 보이지 않으므로 등분산성을 만족한다.

정규성 검토 : 데이터가 직선에 적합되었기 때문에 정규성을 만족한다.

회귀모형 추정

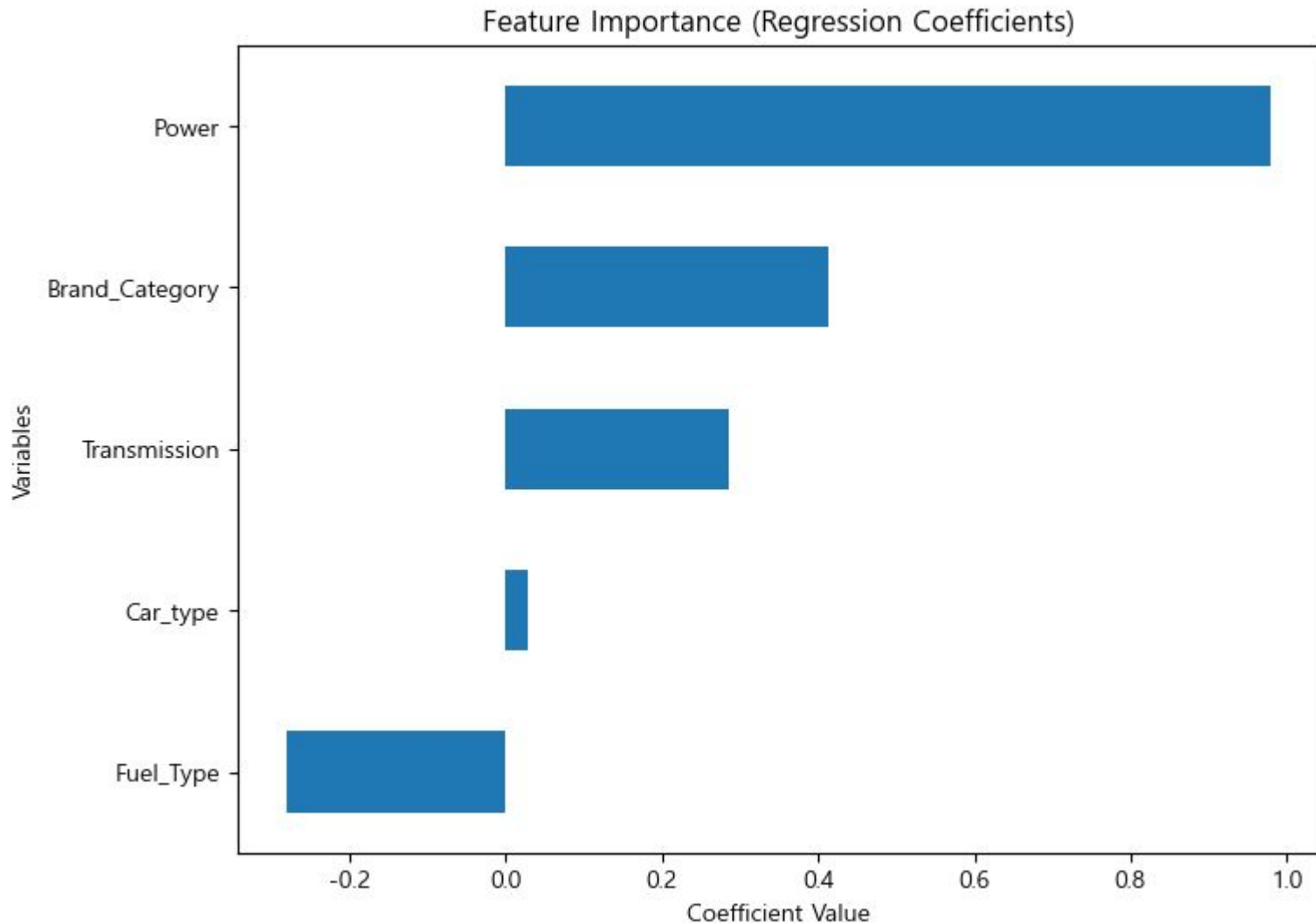
OLS Regression Results

Dep. Variable:	Price	R-squared:	0.673	→	해당 모델의 설명력은 67.3%를 가진다.
Model:	OLS	Adj. R-squared:	0.673		
Method:	Least Squares	F-statistic:	1633.	→	귀무가설 : 추정된 회귀식은 유의하지 않다
Date:	Sat, 22 Mar 2025	Prob (F-statistic):	0.00		대립가설 : 추정된 회귀식은 유의하다.
Time:	11:28:55	Log-Likelihood:	-2515.9		
No. Observations:	3969	AIC:	5044.		검정결과 : F분포의 P-Value 값이 0.05보다 작고,
Df Residuals:	3963	BIC:	5081.		F분포의 값이 높으므로 귀무가설을 기각
Df Model:	5				즉, 이 회귀식은 유의하다고 할 수 있다.
Covariance Type:	nonrobust				

	coef	std err	t	P> t	[0.025	0.975]	
Intercept	4.4466	0.142	31.220	0.000	4.167	4.726	추정 회귀식 = 4.4466 -0.2807Fuel_Type +
Fuel_Type	-0.2807	0.016	-17.923	0.000	-0.311	-0.250	0.2858Transmission + 0.9798Power +
Transmission	0.2858	0.023	12.680	0.000	0.242	0.330	0.4124Brand_Category + 0.0290Car_Trye
Power	0.9798	0.039	25.144	0.000	0.903	1.056	
Brand_Category	0.4124	0.030	13.842	0.000	0.354	0.471	
Car_type	0.0290	0.019	1.526	0.127	-0.008	0.066	Car_Type의 pvalue가 0.05를 넘기는 하지만 이를
							제외한 변수의 p-value는 유의하다고 할 수 있다.

Omnibus:	171.976	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	194.668
Skew:	-0.520	Prob(JB):	5.35e-43
Kurtosis:	3.308	Cond. No.	124.

설명변수의 중요도

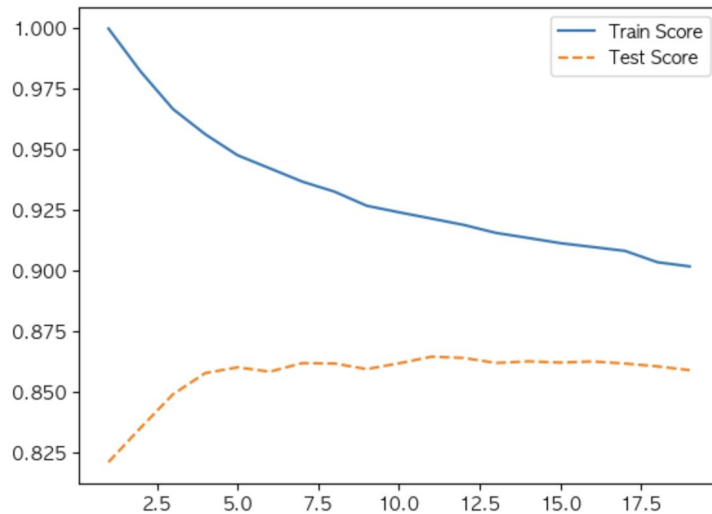


회귀분석 결과 설명변수의 중요도

- Power : 중요도가 가장 크며, 출력이 높을수록 가격이 증가한다.
- Transmission : 오토 차량이 수동보다 더 높은 가격을 형성한다.
- Brand_Category : 프리미엄 브랜드일수록 가격이 증가한다.
- Feul_Typeee : 휘발유의 차량일수록 가격이 상승한다.
- Car_Type은 p-value가 0.05를 넘기에 유의하지 않은 변수로 판단한다.

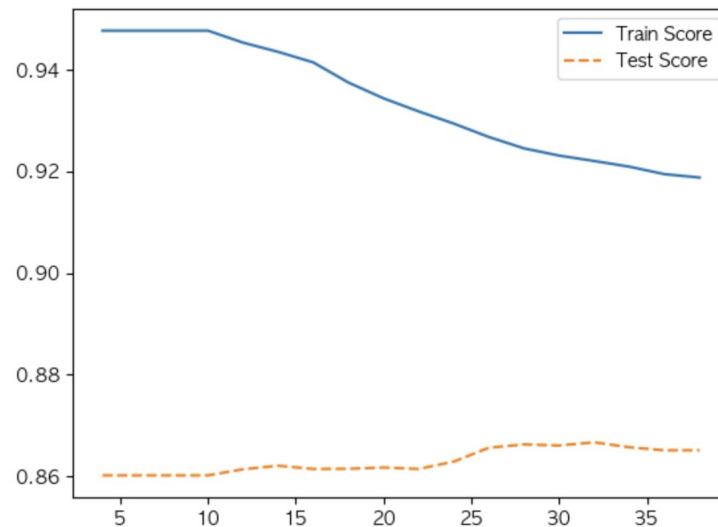
데이터분석_의사결정 나무

파라미터 값 선정기준: 1) Train score와 Test score 차이 10% 이내 2) Train score < 98%(과적합 방지)



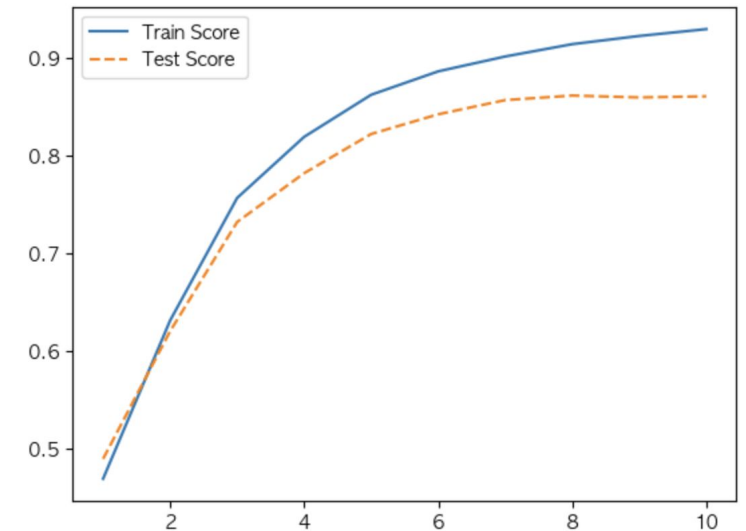
min samples leaf 설정

▶ min samples leaf =5
train score = 94.8%
test score = 86.0%



min samples split 설정

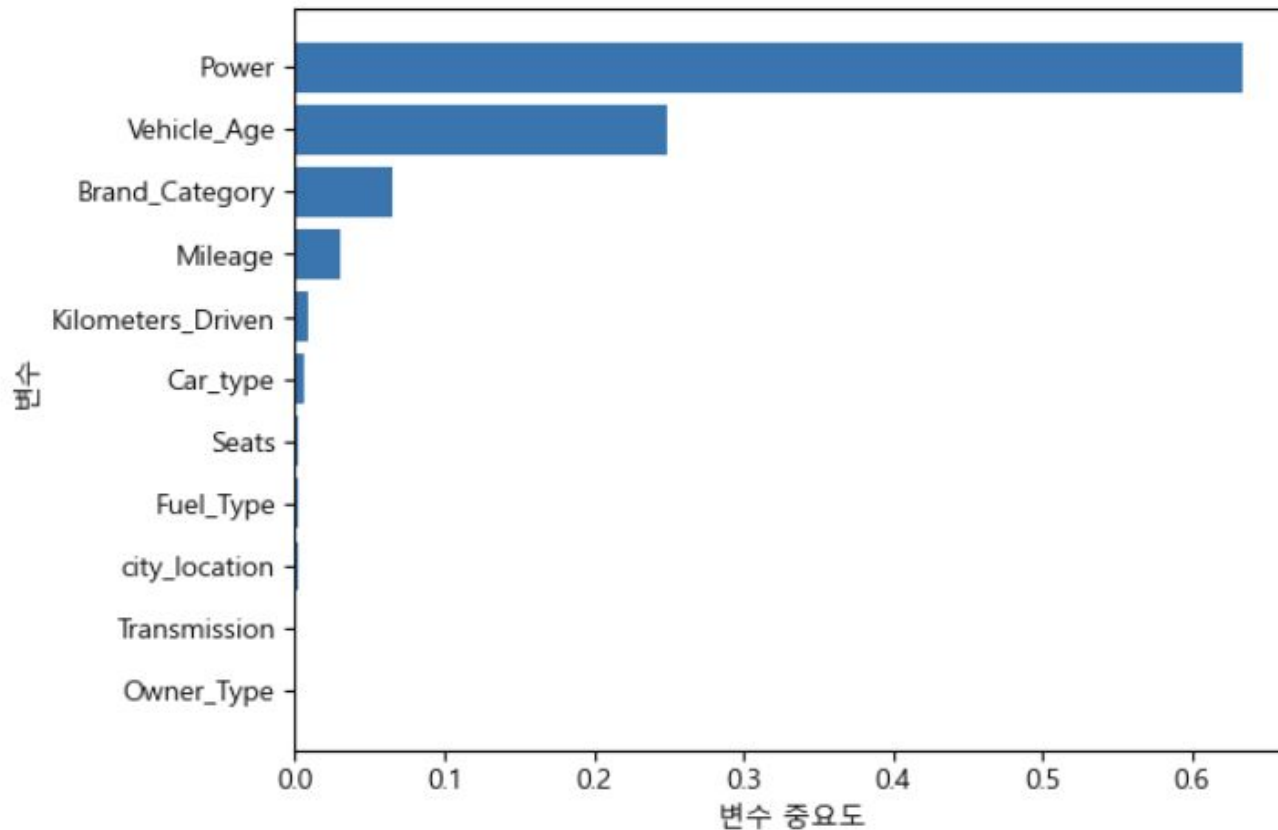
▶ min samples split =16
train score = 94.1%
test score = 86.1%



max depth 설정

▶ max depth =10
train score = 93.0%
test score = 86.1%

데이터분석_의사결정 나무



최종 결과

Score on training set: 0.930
Score on test set: 0.861

변수 중요도

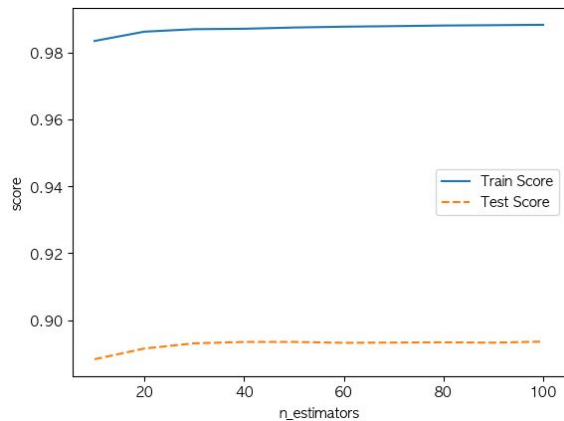
- Power의 중요도가 다른 변수들에 비해 많이 높음

결론

- Train score: 93% , Test score: 86.1%로 최적화되어 있음

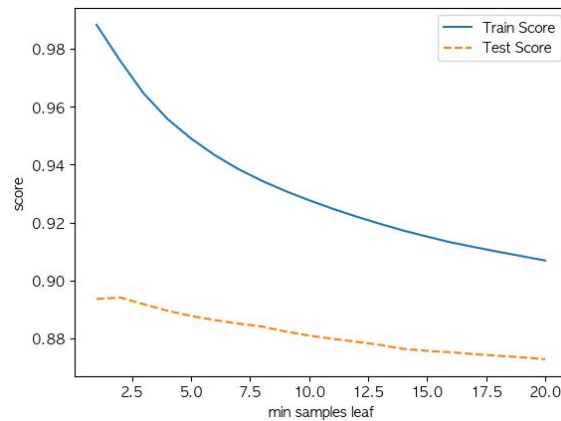
데이터분석_Random forest

파라미터 값 선정기준: 1) test score 높은 순 2) train score < 0.98 (과적합 방지)



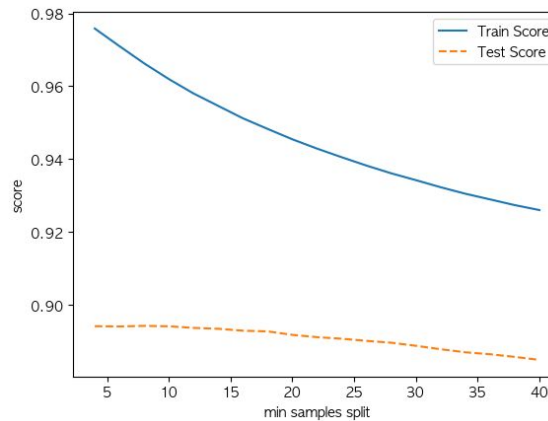
n estimators 설정

→n =100
train score = 98.8%
test score = 89.4 %



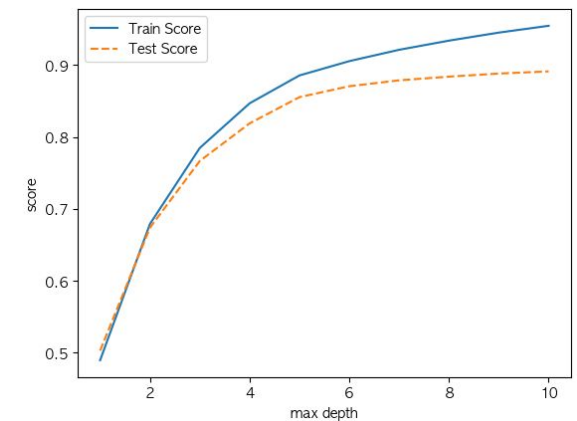
min samples leaf 설정

→min samples leaf =2
train score = 96.5%
test score = 89.2%



min samples split 설정

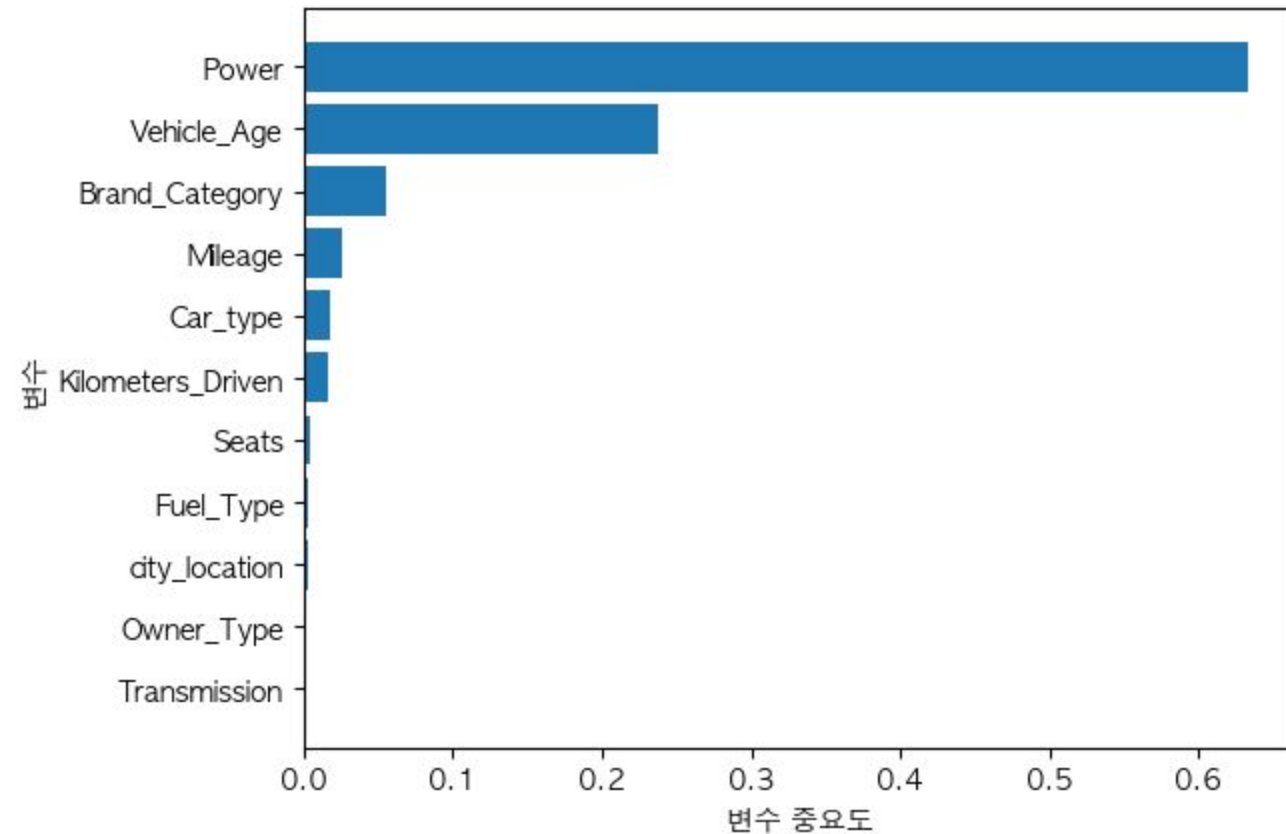
→min samples split =4
train score = 95.8%
test score =89.4%



max depth설정

→max depth =10
train score = 95.5%
test score = 89.1%

데이터분석_Random forest



최종 결과

Score on training set: 0.955
Score on test set: 0.891

변수 중요도

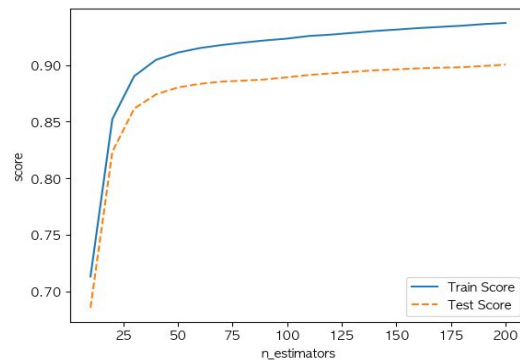
- Power의 중요도가 다른 변수들에 비해 많이 높음

결론

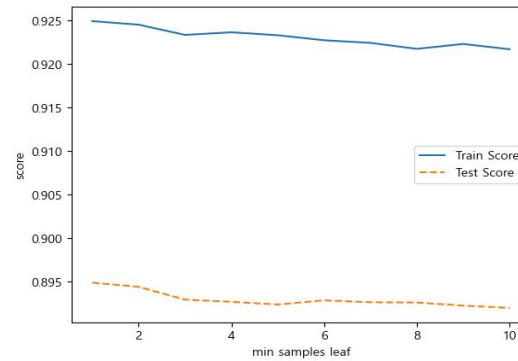
- Train score: 95% , Test score: 89.1%로 최적화되어 있음

데이터분석_Gradient Boosting

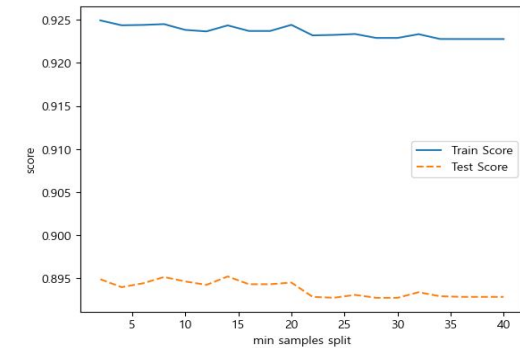
파라미터 값 선정기준: 1) Train와 Test 값의 Gap 차이를 고려 2) 모델링의 복잡성 고려



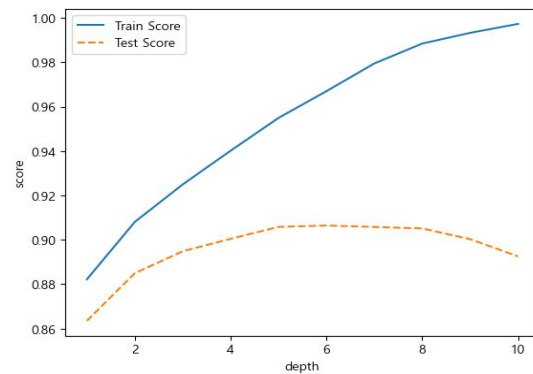
n estimators = 110



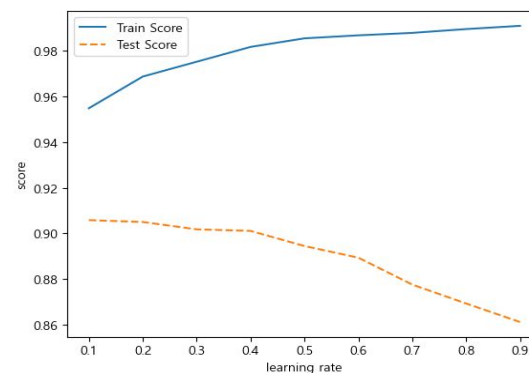
min samples leaf = 1



min samples split = 2

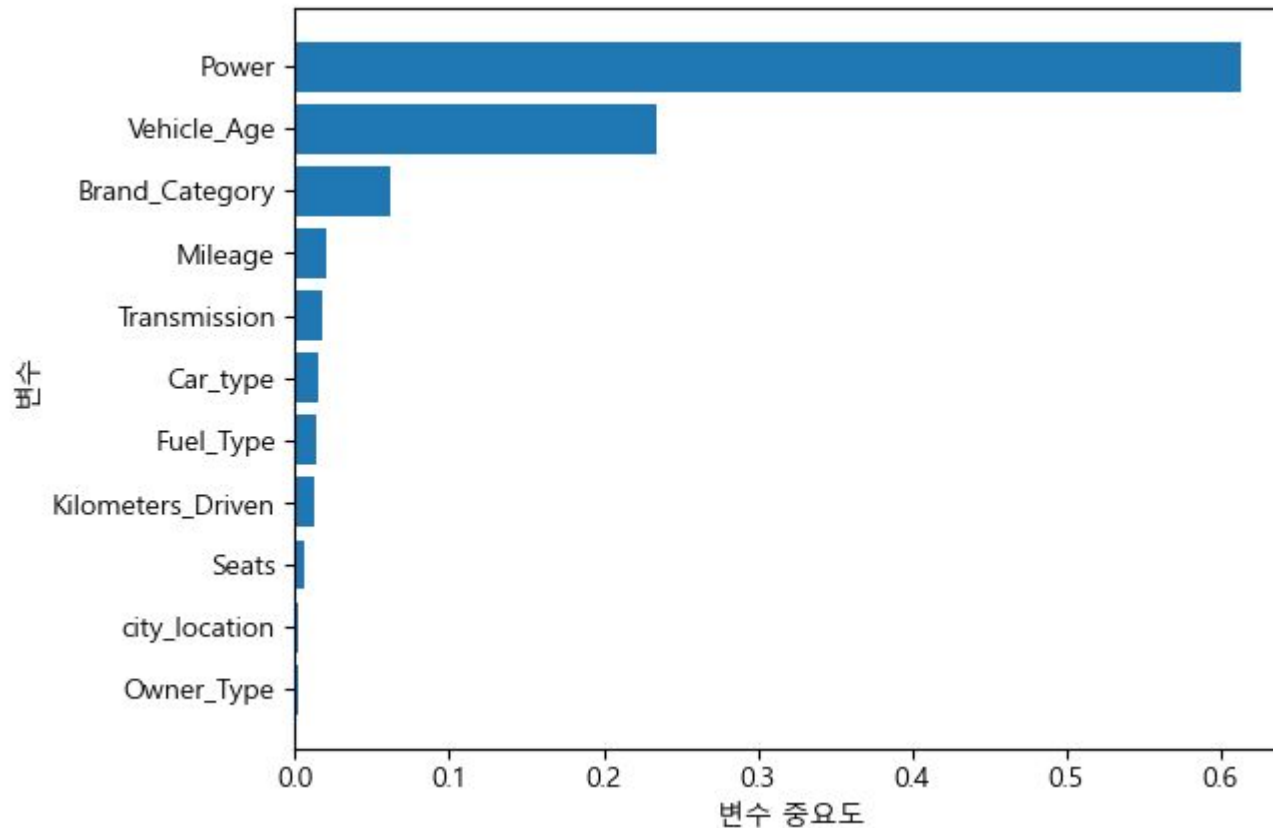


max depth = 5



learning rate = 0.1

데이터분석_Gradient Boosting



최종 결과

Score on training set: 0.953
Score on test set: 0.906

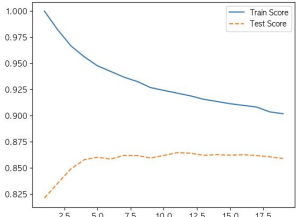
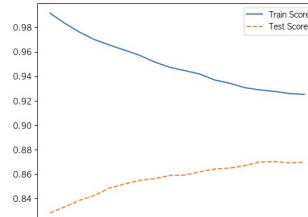
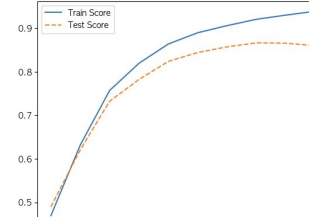
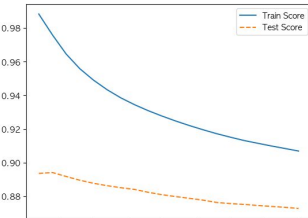
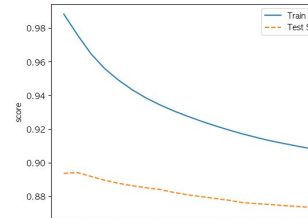
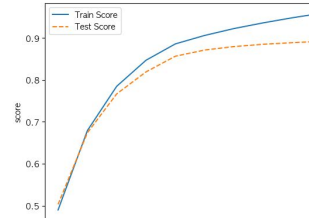
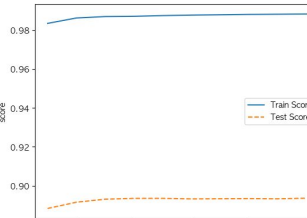
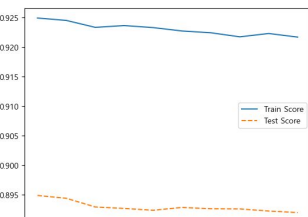
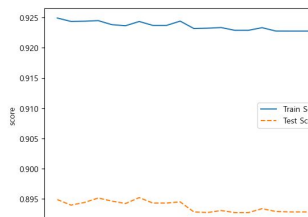
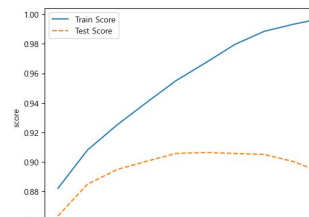
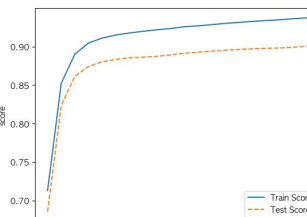
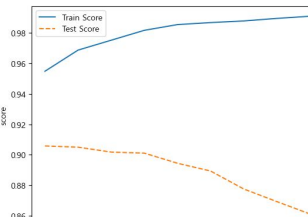
변수 중요도

- **Power** 와 **Vehicle_Age** 는 모델 예측에 가장 큰 영향을 미치는 주요 변수로, 이 변수들을 중심으로 한 예측 모델이 유효한 것으로 확인
- **Owner-Type** 변수의 중요도는 매우 낮은 것으로 확인

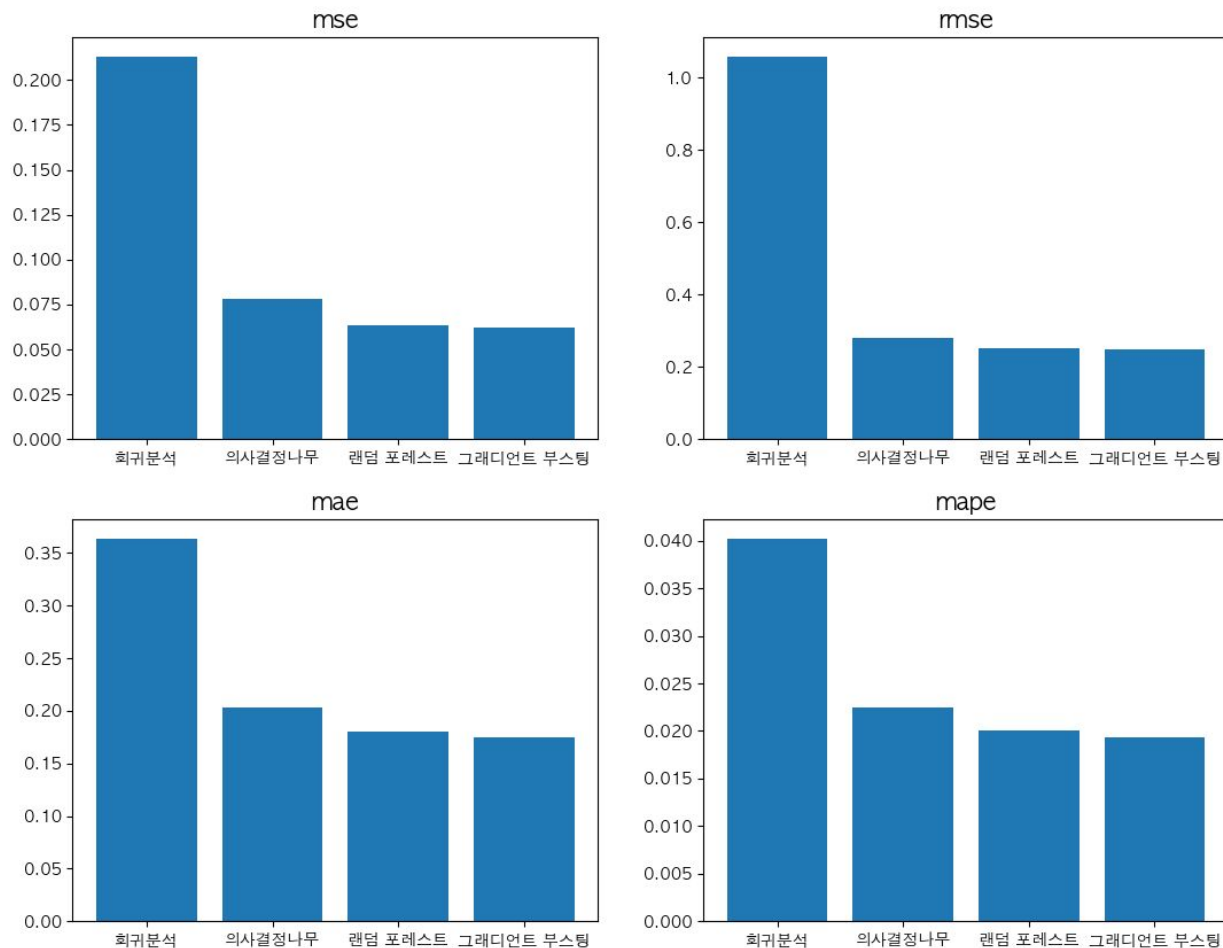
결론

- Train score: 95.3% , Test score: 90.6%로 예측 성능이 안정적인 모델을 확보

분석결과정리

예측 모델	min_samples_ leaf	min_samples_ split	max_depth	n_estimators	learning_rate	Score	
						Train	Test
의사결정 나무						93.2%	86.0%
랜덤 포레스트						95%	89.1%
그래디언 트 부스팅						95.3%	90.6%

모델평가 그래프



모델평가 분석

> 결과 요약

회귀분석은 가장 단순하지만 예측 성능은 가장 낮다.

트리 기반 모델(랜덤포레스트, 그래디언트 부스팅)이 우수한 성능을 보인다.

중고차 가격은 변수 간 복잡한 상호작용이 존재한다.

그래디언트 부스팅이 모든 오차 지표에서 가장 낮다.

> 최종 모델 선정 : 그래디언트 부스팅

- 개선안도출/선정 : 예측 모델 개선

- 개선안 구체화 : Price 결측값 회귀모델 예측

- Pilot test

: Car_type 변수의 유의성을 반영하기 위해, 모델 복잡도 증가를 수용하고 해당 변수를 포함하는 방향의 개선안 선택

=====						
Dep. Variable:	Price	R-squared:	0.673			
Model:	OLS	Adj. R-squared:	0.673			
Method:	Least Squares	F-statistic:	1633.			
Date:	Sat, 22 Mar 2025	Prob (F-statistic):	0.00			
Time:	14:14:14	Log-Likelihood:	-2515.9			
No. Observations:	3969	AIC:	5044.			
Df Residuals:	3963	BIC:	5081.			
Df Model:	5					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	4.4466	0.142	31.220	0.000	4.167	4.726
Fuel_Type	-0.2807	0.016	-17.923	0.000	-0.311	-0.250
Transmission	0.2858	0.023	12.680	0.000	0.242	0.330
Power	0.9798	0.039	25.144	0.000	0.903	1.056
Brand_Category	0.4124	0.030	13.842	0.000	0.354	0.471
Car_type	0.0290	0.019	1.526	0.127	-0.008	0.066
=====						
Omnibus:	171.976	Durbin-Watson:	1.991			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	194.668			
Skew:	-0.520	Prob(JB):	5.35e-43			
Kurtosis:	3.308	Cond. No.	124.			
...						
=====						

개선 전 모델

OLS Regression Results						
Dep. Variable:	Price	R-squared:	0.672			
Model:	OLS	Adj. R-squared:	0.671			
Method:	Least Squares	F-statistic:	2318.			
Date:	Sat, 22 Mar 2025	Prob (F-statistic):	0.00			
Time:	15:59:46	Log-Likelihood:	-3613.0			
No. Observations:	5670	AIC:	7238.			
Df Residuals:	5664	BIC:	7278.			
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	4.5169	0.119	37.897	0.000	4.283	4.751
Fuel_Type	-0.2836	0.013	-21.572	0.000	-0.309	-0.258
Transmission	0.2910	0.019	15.317	0.000	0.254	0.328
Power	0.9650	0.033	29.623	0.000	0.901	1.029
Car_type	0.0373	0.016	2.339	0.019	0.006	0.069
Brand_Category	0.3890	0.025	15.685	0.000	0.340	0.438
Omnibus:	270.864	Durbin-Watson:	1.915			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	322.705			
Skew:	-0.519	Prob(JB):	8.43e-71			
Kurtosis:	3.536	Cond. No.	123.			

개선 후 모델

인도 시장 현황

- 2018년 기준 인도의 자동차의 보급률은 0.022 수준이다. 인도의 인구대비 자동차 보유 비율은 상당히 낮으며, 점차 비율이 확대될 것으로 전망된다. 팬데믹 이후 보건의 사유로 인도내 자동차 구입 수요가 증가하고 있으며, 글로벌 반도체 수급 부족으로 중고차의 지속적으로 성장하는 매력적인 시장이다.

경쟁요인

- 인도 중고차 경쟁업체 현황 (오프라인 / 온라인)

Offline Dominant (OEM Backed)



Source: car&bike Analysis

Online Dominant (Tech Start-Ups)



Source: car&bike Analysis

시장적합성

- 분석 결과 중고차 가격에 가장 큰 요인은 마력과 자동차의 연식이 상관관계가 있음을 보였다.
- 이에따라 Pos_Car 업체에서는 중고차를 매입할 때, 마력이 높고, 자동차의 연식이 낮은 자동차를 매입을 하는 것을 우선으로 하는 전략을 설정해야 한다.