인도 중고차 가격 핵심영향인자 도출 및 가격예측모델 개발

Big Data Analysis Report

POS_Cars

C4

과제 정의

1) 인도 중고차 시장 현황

인도 중고차 시장 지속적 성장 예상

- 32조 5970억 원 (2023년 기준, $27.47 billion)
- 73조 8950억 원 (2028년 예상, $55.49 billion)
- 연평균 성장률:15.10%

2) 인도 내 차량 종류 인식 현황

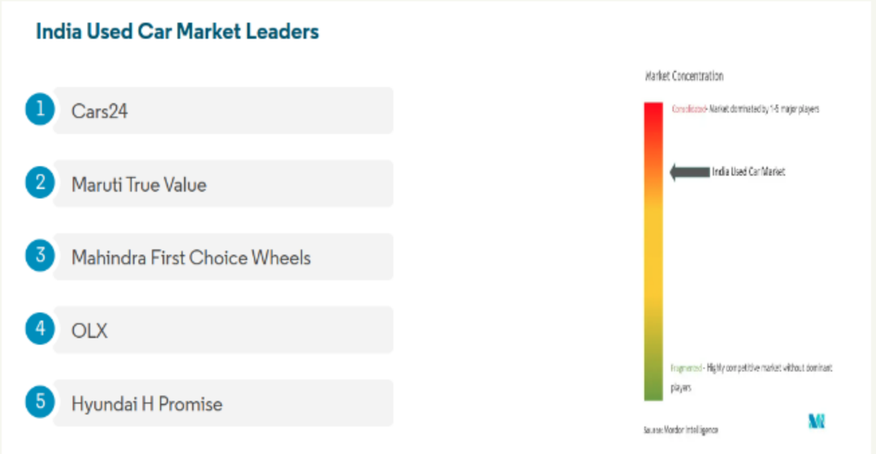
TABLE 5 - TYPE OF INDIAN BRAND CARS OWNED

Type of Indian brand car	Respondents	Percent
Mini car	32	21.3
Macro car	35	23.3
Compact car	34	22.7
Hatchback car	12	8.0
Sports utility car	10	6.7
Luxury car	27	18.0
Total	**150**	**100.0**

Source: Primary data

Source: http://ijrar.com/upload_issue/ijrar_issue_20542847.pdf

3) 인도 중고차 시장 포화 상태



▶ 데이터 분석을 통해 포화 상태인 인도 중고차 시장 내 차별화된 전략 및 서비스 제공 방안 도출

분석 계획

데이터 수집

중고차 데이터
csv파일 업로드

데이터 전처리

데이터 현황 파악 및 정제

데이터 분석

통계 분석, 머신 러닝을
사용한 패턴 분석

개선 방안 도출

분석 결과 해석 및
개선 방안 제안

데이터 현황

결측치

isnull() 함수를 이용한 파일 내 결측치 확인

Price: 1053

Mileage: 2

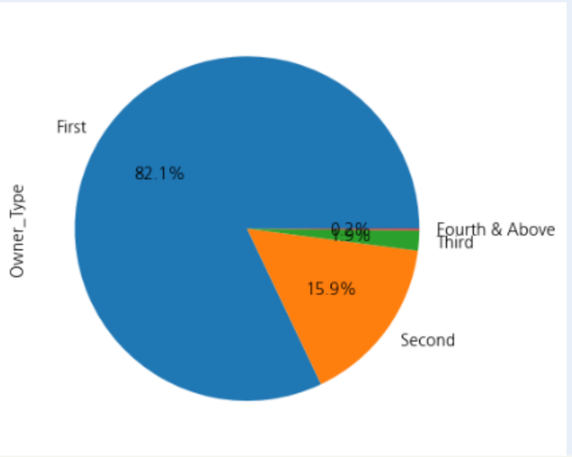
Engine: 43

Power: 43

Seats: 53

New_Price: 6247

```
Name                   0
Location               0
Price               1053
Year                   0
Kilometers_Driven      0
Fuel_Type              0
Transmission           0
Owner_Type             0
Mileage                2
Engine                46
Power                 46
Seats                 53
New_Price           6247
dtype: int64
```



범주형 변수 확인

Name

Location

Owner_Type

Fuel_Type

Transmission

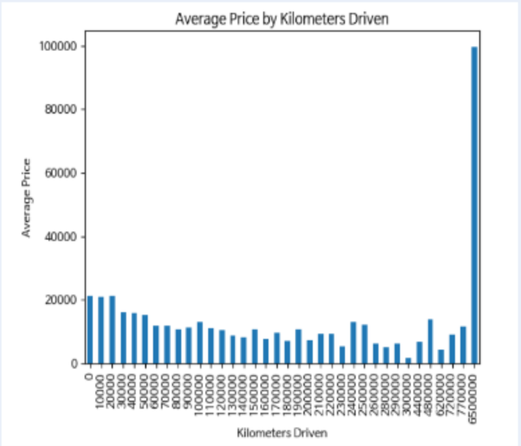
이상치

그래프를 이용한 파일 내 이상치 확인

Mileage: 29개

Kilometers_Driven: 1개

Brand: 1개



변수 현황

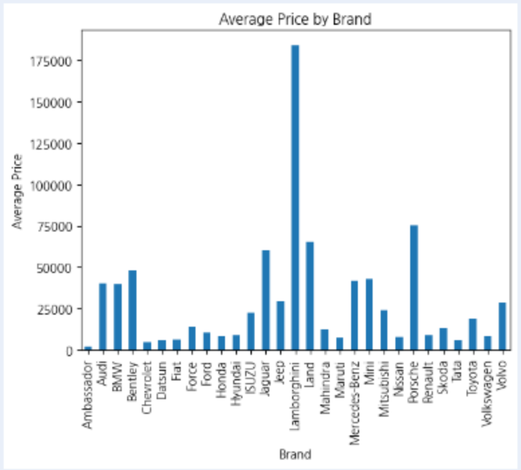
	Name	Location	Price	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price
0	Maruti Wagon R LXI CNG	Mumbai	2682.68	2010	72000	CNG	Manual	First	26.6 kmpl	998 CC	58.16 bhp	5.0	NaN
1	Hyundai Creta 1.6 CRDi SX Option	Pune	19162.00	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN
2	Honda Jazz V	Chennai	6898.32	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh
3	Maruti Ertiga VDI	Chennai	9197.76	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	27194.71	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN
5	Hyundai EON LPG Era Plus Option	Hyderabad	3602.46	2012	75000	LPG	Manual	First	21.1 kmpl	814 CC	55.2 bhp	5.0	NaN
6	Nissan Micra Diesel XV	Jaipur	5365.36	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5.0	NaN

탐색적 분석

결측치 정제

1. Price: 목표변수기 때문에 결측치를 건드리면 안된다고 판단해서 결측치 삭제
2. Mileage, Power: 차량간 편차가 심해 모든 결측치 삭제
3. Seats: 최빈도값 입력
4. New_Price: 결측치가 많아서 변수 제거

```
Name                   0
Location               0
Price               1053
Year                   0
Kilometers_Driven      0
Fuel_Type              0
Transmission           0
Owner_Type             0
Mileage                2
Engine                46
Power                 46
Seats                 53
New_Price           6247
dtype: int64
```



범주형 변수 변환

1. Name 변수에서 Brand명 추출 후 그룹화(평균값 기준)

2. Location 변수 그룹화 (인구수 기준)

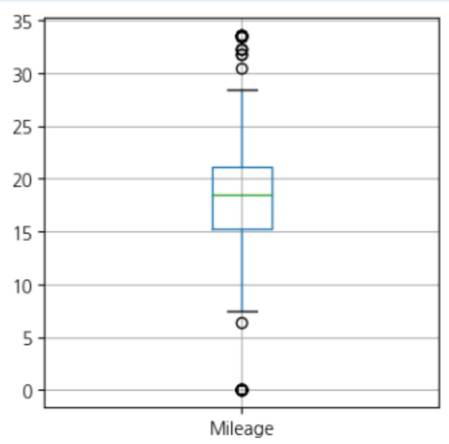
3. ①, ② 포함 모든 범주형 설명변수 더미 변환

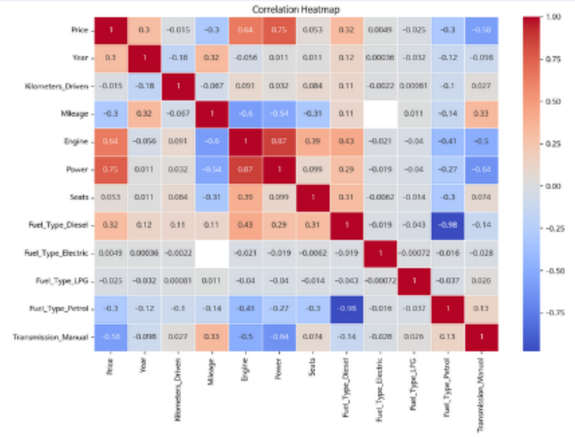
이상치 제거

1.Box Plot으로 Mileage가 0인 이상치 발견 후 제거

2. Bar Chart로 Kilometers_Driven가 6500000인 이상치 발견 후 제거

3. Bar Chart로 Brand에서 OpelCorsa의 값이 0인 것을 확인 후 제거





변수 제거

1. Engine과 Power의 상관관계가 0.75로 다중공선성 방지를 위해서 Engine 제거

2. F 분포 ANOVA 검정을 통해서 Owner_Type간 가격차이가 유의미하지 않다고 판단 후 제거 (p-value: 0.15)

모델링 & 요약

다중선형 회귀 분석

R-Squared: 0.716
Adj. R-Squared: 0.715

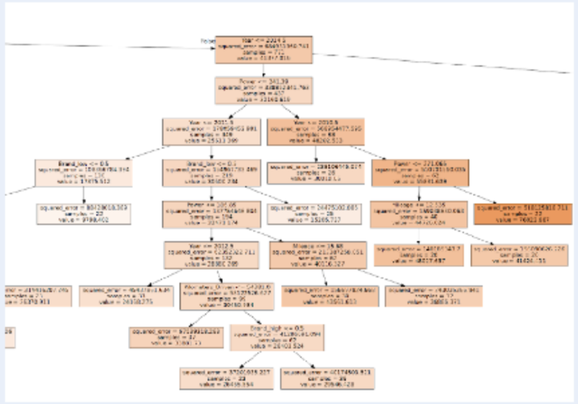
OLS Regression Results

Dep. Variable:	Price	R-squared:	0.716
Model:	OLS	Adj. R-squared:	0.715
Method:	Least Squares	F-statistic:	1375.
Date:	Sun, 05 Nov 2023	Prob (F-statistic):	0.00
Time:	01:03:33	Log-Likelihood:	-63677.
No. Observations:	6025	AIC:	1.274e+05
Df Residuals:	6013	BIC:	1.275e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.985e+06	9.55e+04	-31.248	0.000	-3.17e+06	-2.8e+06
Kilometers_Driven	-0.0269	0.004	-6.803	0.000	-0.035	-0.019
Year	1484.7018	47.509	31.251	0.000	1391.567	1577.837
Location_Urban	-885.2494	248.344	-3.567	0.000	-1371.700	-398.798
Transmission_Automatic	771.4292	381.053	2.024	0.043	24.428	1518.430
Mileage	-173.7572	42.667	-4.072	0.000	-257.400	-90.115
Power	147.7657	4.089	36.139	0.000	139.750	155.781
Fuel_Type_CNG	2424.2483	1303.081	1.860	0.063	-130.257	4978.754
Fuel_Type_Diesel	2583.3001	300.040	8.610	0.000	1995.134	3171.486
Fuel_Type_LPG	5766.6265	2989.046	1.929	0.054	-92.575	1.16e+04
Brand_high	1.071e+04	628.311	17.051	0.000	9481.814	1.19e+04
Brand_low	-5433.6040	494.555	-10.987	0.000	-6403.109	-4464.099

Omnibus:	4813.854	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	485591.895
Skew:	3.177	Prob(JB):	0.00
Kurtosis:	46.519	Cond. No.	5.35e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.35e+07. This might indicate that there are strong multicollinearity or other numerical problems.



Decision Tree

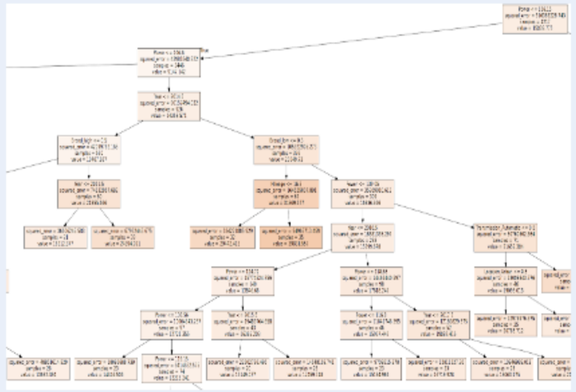
Hyper Parameter:

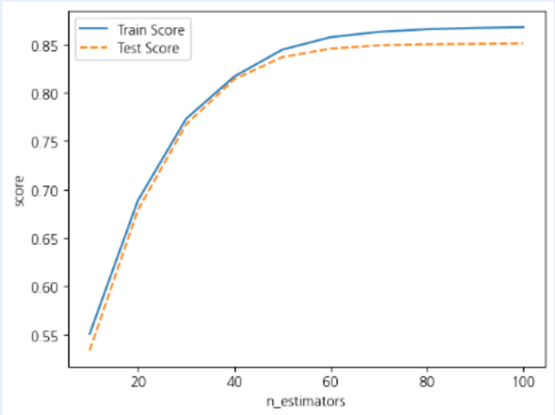
min_sample_leaf: 20
min_sample_split: 30
max_depth: 9

Random Forest

Hyper Parameter:

n_estimators:10
min_sample_leaf: 12
min_sample_split: 40
max_depth: 7





Gradient Boosting

Hyper Parameter:

n_estimators:100
min_sample_leaf: 40
min_sample_split: none
max_depth: 6
learning rate: 0.1

모델링 & 요약

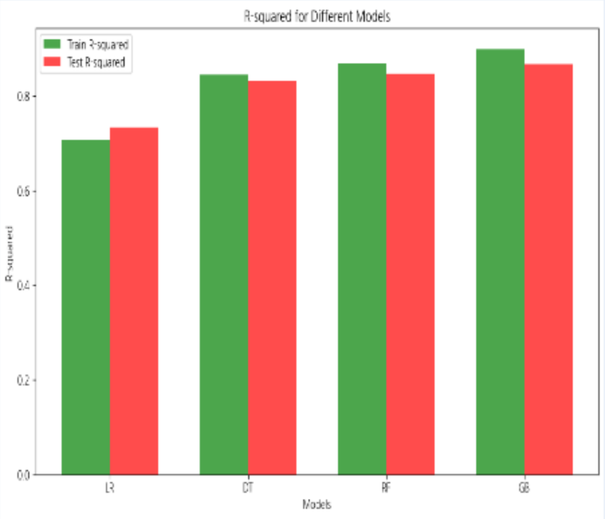
R-Squared

다중선형 회귀분석:
Train R-Squared = 0.707
Test R-Squared = 0.734

Decision Tree:
Train R-Squared = 0.845
Test R-Squared = 0.832

Random Forest:
Train R-Squared = 0.869
Test R-Squared = 0.847

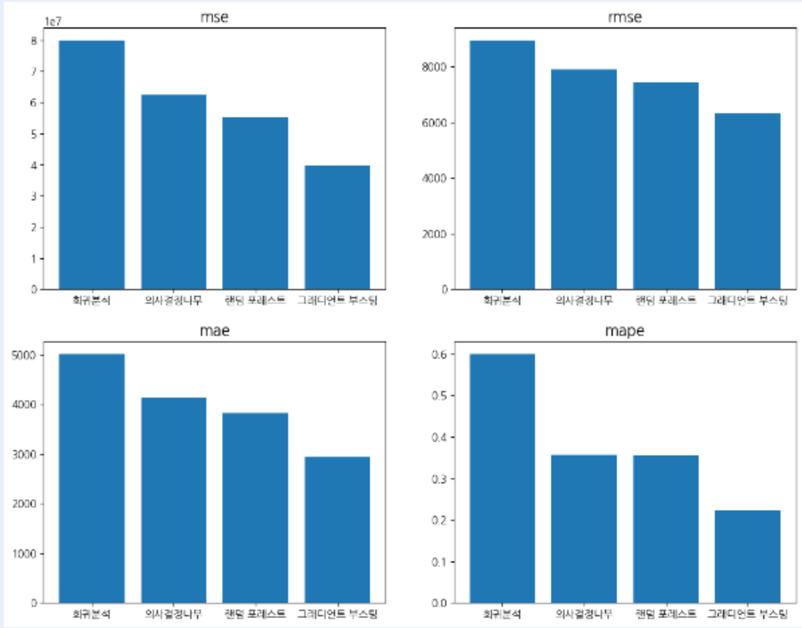
Gradient Boosting:
Train R-Squared = 0.898
Test R-Squared = 0.878



R-Squared 순위

Gradient Boosing > Random Forest > Decision Tree > 회귀분석

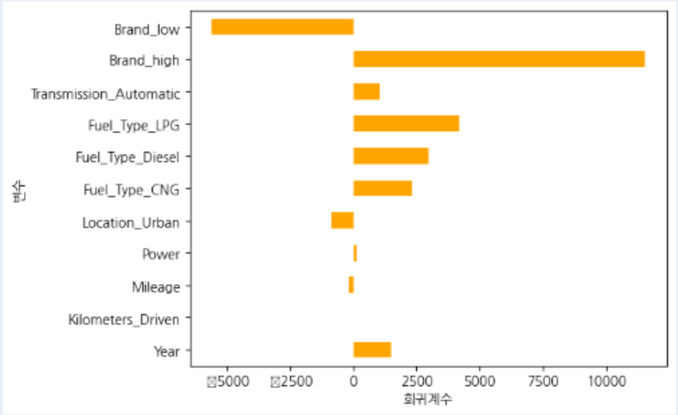
모델 평가 지표



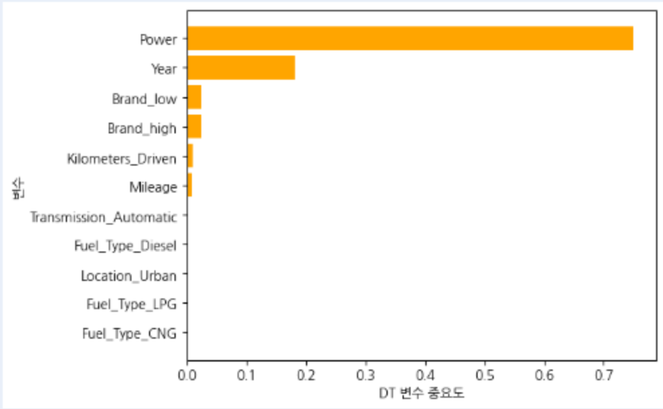
mse (Mean Squared Error): LR > DT > RF > GB
rmse (Root Mean Square Error): LR > DT > RF > GB
mae (Mean of Absolute Value of Error): LR > DT > RF > GB
mape (Mean Absolute Percentage Error): LR > DT > RF > GB

모델별 변수 중요도

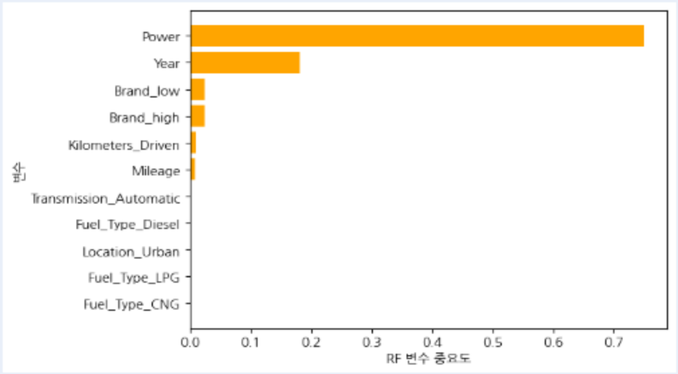
회귀분석



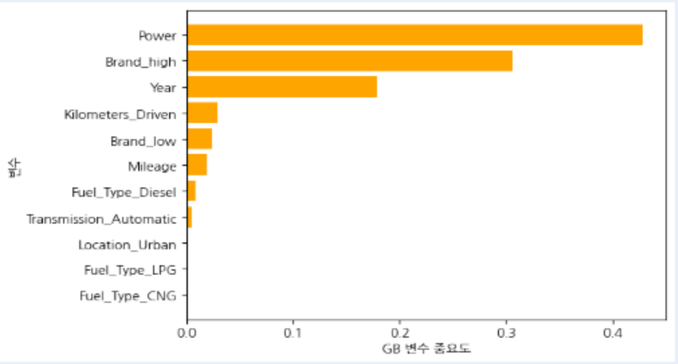
Decision Tree



Random Forest



Gradient Boosting



변수 중요도

변수	회귀분석	DT	RF	GB	총점	순위	선정
Year	6	2	2	3	3.25	3	O
Kilometers_Driven	-	5	5	4	4.67	**5**	O
Mileage	8	6	**6**	**6**	6.5	6	O
Power	9	1	1	1	3	2	O
Location_Urban	7	-	-	-	-	-	X
Fuel_Type_CNG	5	-	-	-	-	-	X
Fuel_Type_Diesel	4	-	-	7	-	-	X
Fuel_Type_LPG	3	-	-	-	-	-	X
Transmission_ Automatic	-	-	-	8	-	-	X
Brand_high	1	4	**4**	2	2.75	1	O
Brand_low	2	3	3	5	4.33	4	O

결론 및 전략 제시

결론 1

가격에 주요한 영향을
끼치는 요소는 브랜드,
파워(엔진), 연식이다

결론 2

보편적인 인식과 달리
주행 거리가 차량 가격에
큰 영향을 끼치지 않는다

결론 3

인도 내 고급 차에 대한
인식이 높아지고 있고
시장 규모 또한 성장 중이다

"The Indian luxury car market studied was valued at USD 1.06 billion in 2021. It is expected to reach a value of over USD 1.54 billion by 2027"

-2021년 기준 인도 내 고급 차량 시장 규모: $1.06 billion
-2027년 기준 인도 내 고급 차량 시장 규모 예상: $1.54 billion

Source: https://www.mordorintelligence.com/industry-reports/india-luxury-car-market

POS_Cars(주)는 고급 브랜드와 모델에 초점을 맞추며, 주행거리보다는 브랜드, 파워, 연식을 강조하는 전략으로 고객들에게 럭셔리 경험을 제공해야 합니다.

한계점 및 제언

중고차 구매 및 가격 선정에서 가장 중요한 요소인 사고 및 보험 관련 변수 없이 예측을 진행했기 때문에 모델의 신뢰성이 부족하다.

데이터셋에서 제공한 변수 외에 사고 내역, 보험 처리 등의 다양한 설명변수 데이터를 수집하여 분석을 진행하면 보다 정확한 예측을 진행할 수 있을 것이다.

럭셔리 중고차 스타트업 특성상 구매할 수 있는 고객층이 한정되어 있어 매출의 안정성이 보장되지 않는다.

고급화 전략을 시작으로 기업의 브랜드 가치와 신뢰성을 높인 후 폭 넓은 가격대의 차량을 도입하여 주 고객층을 확장시킨다.

Lesson Learned

01 데이터 분석을 통해 목표 예측값에 대한 예상치 못한 요인을 발견할 수 있기 때문에 흔히 알려진 사실과 일치하지 않을 수 있음을 확인했다.

02 모델마다 변수 중요도가 다르기 때문에 다양한 모델을 활용한 교차검증의 필요성을 확인했다.

03 프로젝트 주제에 대한 도메인 지식이 갖춰져 있어야 수치로 이루어진 데이터 분석 결과와 비즈니스 목표를 유기적으로 연결할 수 있다.