중고자동차 시세 예측

Team 1
2017036835 조희승
2016033045 박태은
2015019898 이동환

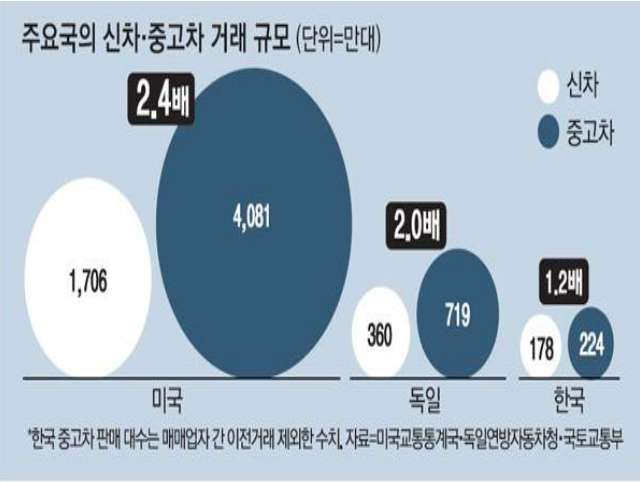
목차

- 문제 정의
- 데이터 전처리
- 모델링
- 결과 확인

문제 정의

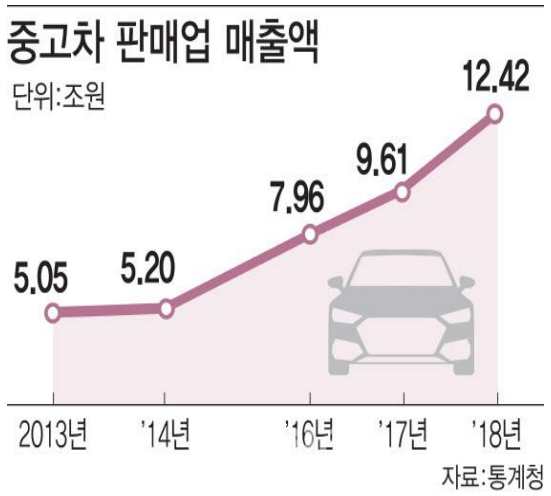
1. 배경

주요국 중고차 거래 규모

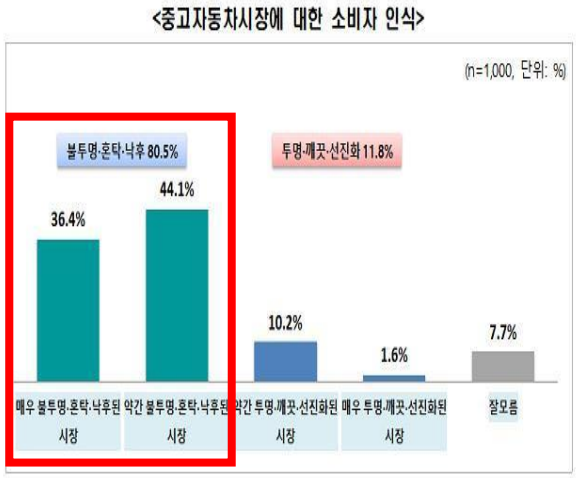


2019년

국내 중고차의 가파른 성장



불투명한 중고차 시장



2020년

2. 목적

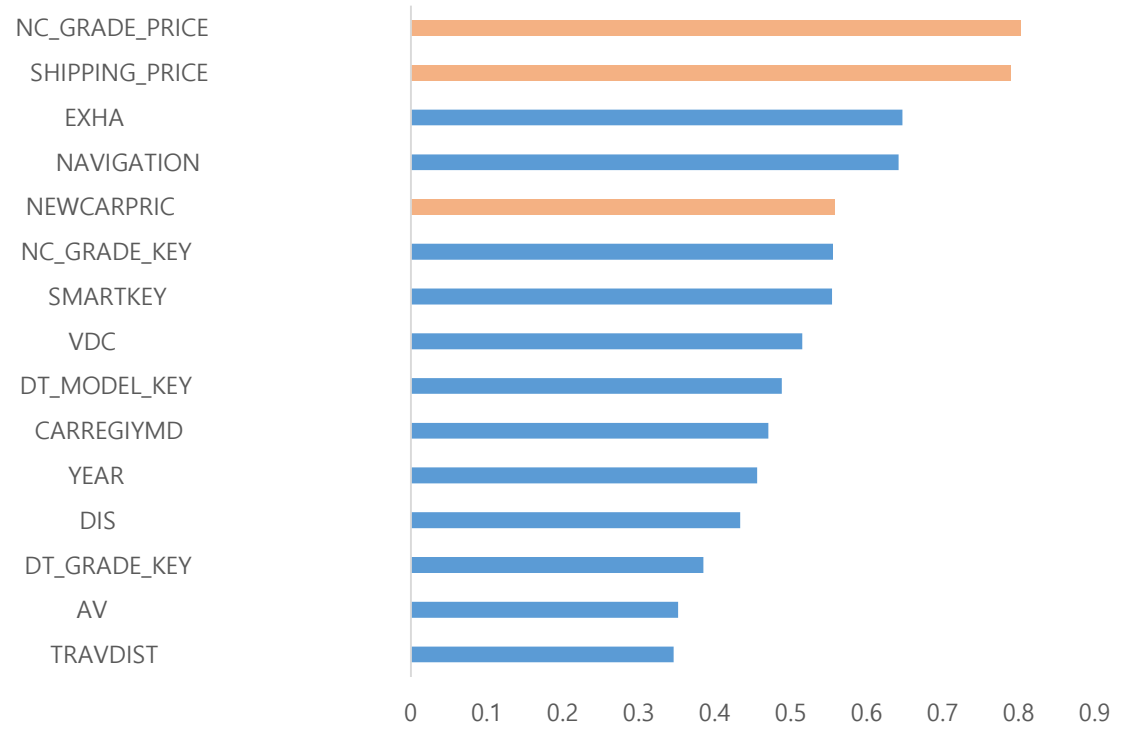
목표

EXHA(배기량),TRAVDIST(주행거리) 등
104개의 col을 가진 데이터 셋 이용

SUCCPRIC(낙찰가) 예측 모델 수립

1. 상관계수 확인

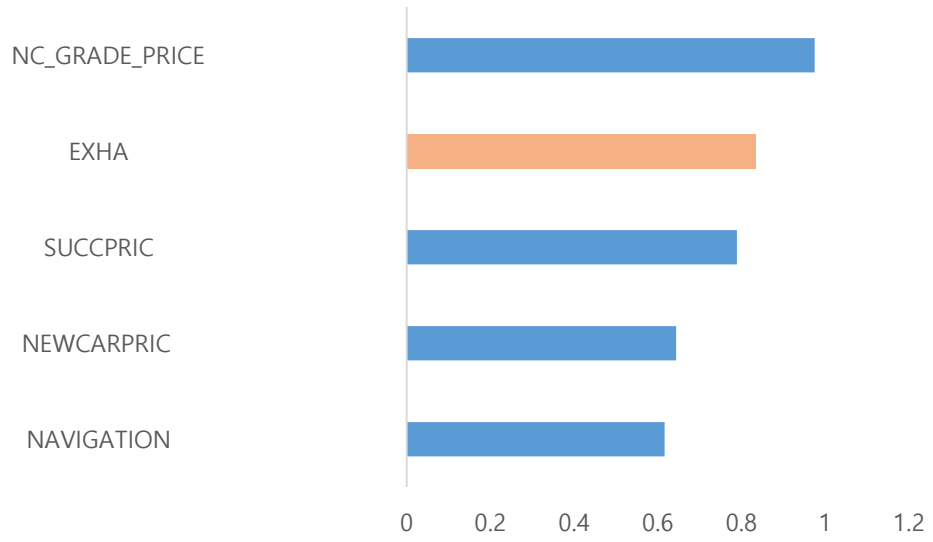
SUCCPRIC과의 상관계수 상위 15개



- 문자형 변수를 제외하고 SUCCPRIC과의 상관계수 상위 15개를 보면 ,PRICE 관련 feature들이 상위에 위치함
- 상관계수만으로는 선형적 관계만 확인할 수 있으므로 각 feature에 대한 설명력을 완전히 보여주지는 못함
- SHPPING_PRICE, NC_GRADE_PRICE, NEWCARPRIC를 합쳐서 PRICE 변수를 만듦

2. PRICE feature 생성

SHIPPING_PRICE과의 상관계수 상위 5개



- **PRICE 열을 새로 만들 때, SHIPPING_PRICE>NC_GRADE_PRICE >NEWCARPRIC 순서대로 값을 넣음**
- **위 3개의 PRICE 값이 모두 결측값일 경우 SHIPPING_PRICE와 상관계수가 높은 EXHA를 이용해서 결측치 처리**

3. EXHA 결측치 처리

FUELNM 별 EXHA 평균값

FUELNM	EXHA
Hybrid	2074
LPG	2209
가솔린	1659
겸용	1027
디젤	1676
전기	44

- EXHA를 이용해서 PRICE 결측치를 채우기 전에 EXHA 결측치부터 대체함
- FUELNM별로 EXHA 평균값으로 EXHA 결측치를 대체함
- FUELNM이 결측치일 경우 EXHA 전체 평균으로 대체함

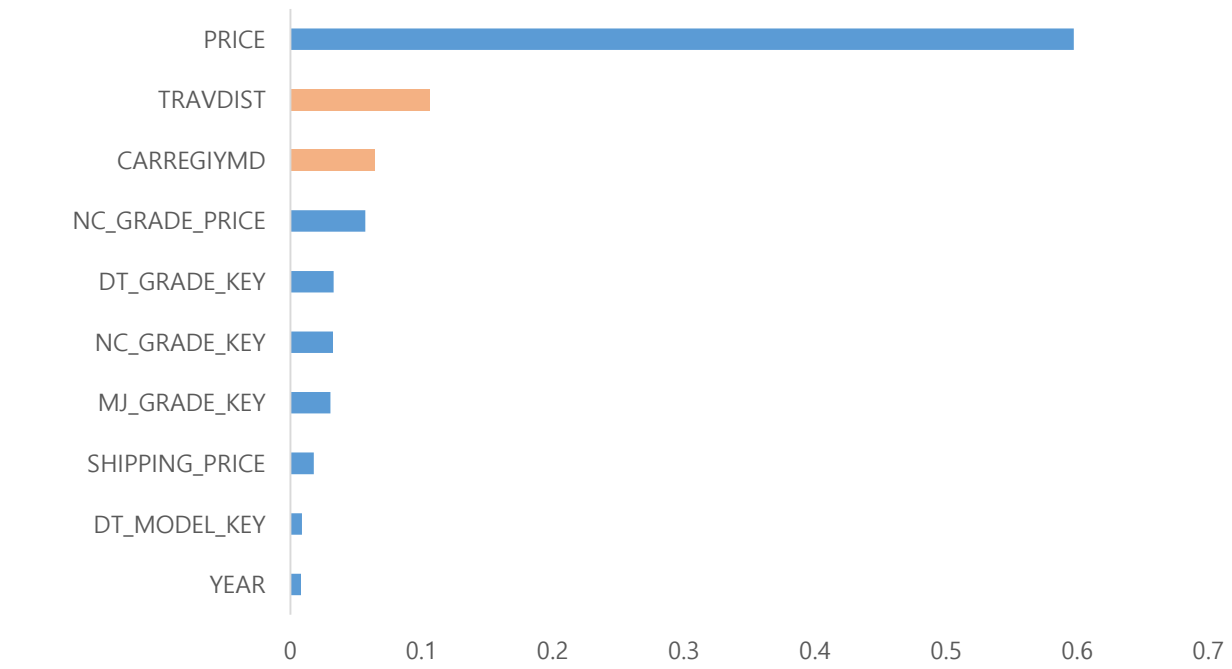
4. PRICE 결측치 처리

EXHA 1000단위 SHIPPING_PRICE 평균값

EXHA	SHIPPING_PRICE
0-999	12,581,245
1000-1999	18,420,902
2000-2999	27,168,002
3000-3999	48,396,963
5000-5999	82,608,486

- **EXHA 1000단위로 SHIPPING_PRICE 평균을 구해서 남은 PRICE 결측치를 대체함**

5. 전체 변수 중요도



- PRICE와 EXHA 결측치를 대체한 뒤, 전체 feature에 대한 랜덤포레스트 중요도 확인
- TRAVDIST(주행거리) 와 CARREGIYMD(차량등록일)가 상위에 위치함. TRAVDIST 결측치 처리와 CARREGIYMD와 관련된 파생변수 생성

6. TRAVDIST 결측치 처리 및 파생변수 생성

파생변수 YEAR_adj 생성

YEAR (년식)	CARREGIYMD (차량등록일)	SUCCYMD (낙찰일자)	YEAR_adj
2011	20100616.0	20160105	5
2013	20130207.0	20160105	3
2014	20140128.0	20160105	2

YEAR_adj = SUCCYMD[:4] - YEAR

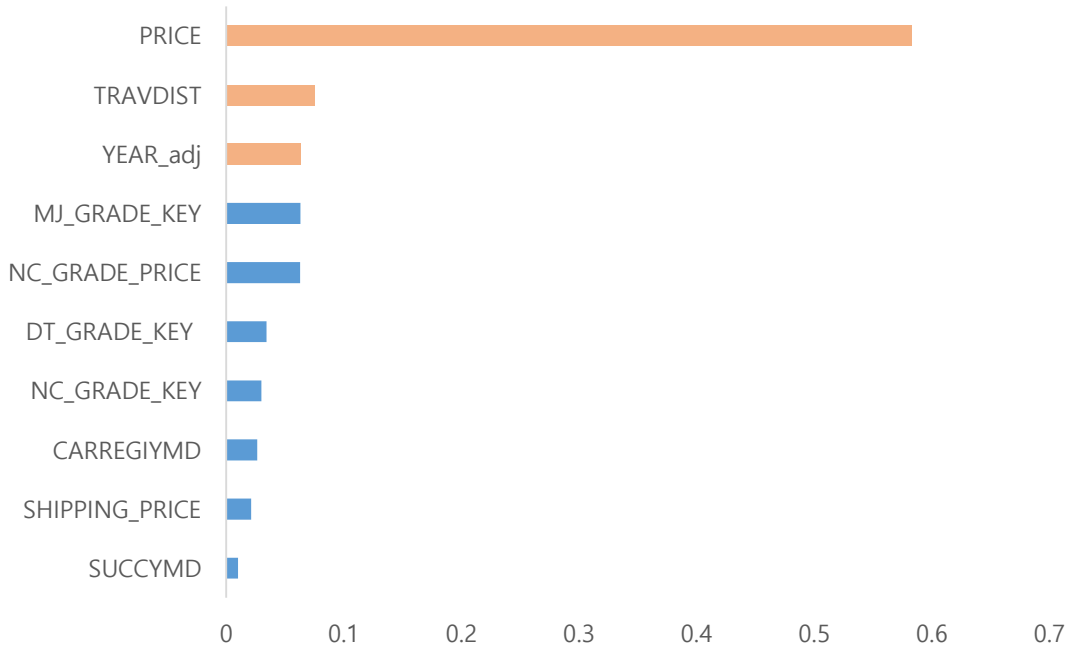
TRAVDIST 결측치 처리

TRAVDIST= 전체 차량 평균주행거리(년)*YEAR_adj

- CARREGIYMD 와 YEAR는 차량등록일과 년식으로 차이가 거의 없음. SUCCYMD-YEAR로 판매가 이루어진 시기까지의 차량 이용 년수를 파생변수로 생성함
- TRAVDIST(주행거리)의 결측치는 전체 차량 년당 평균주행거리를 구한 뒤, 해당 차량 YEAR_adj를 곱하여 TRAVDIST 값으로 대체

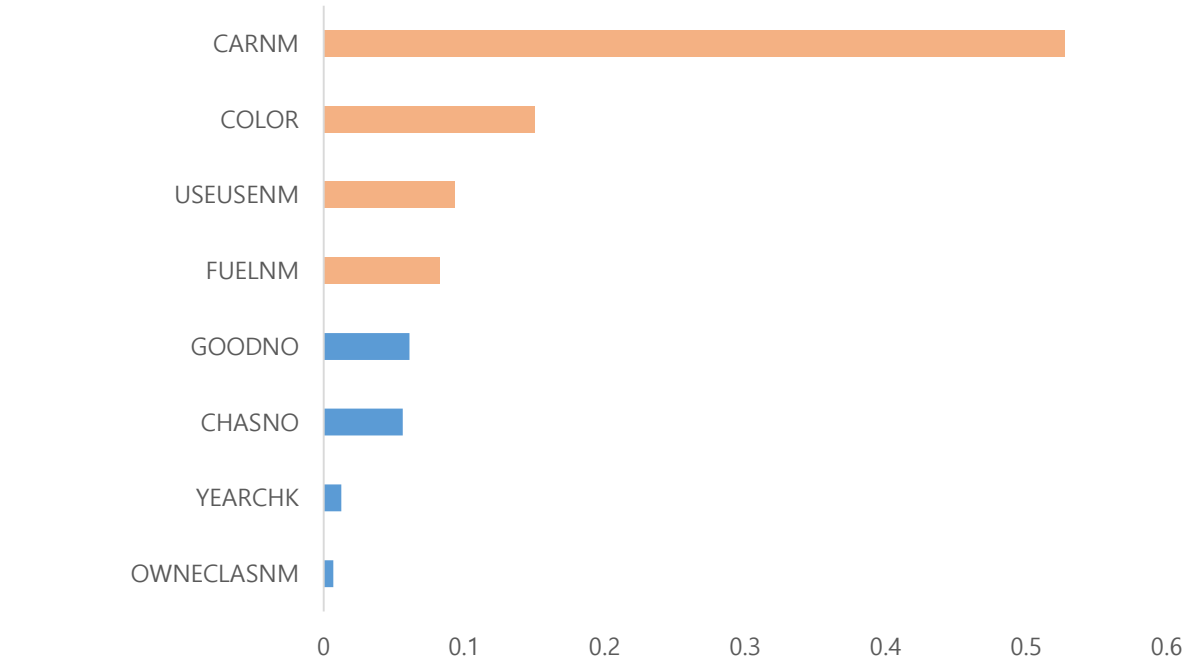
7. 숫자형 변수 중요도

숫자형 변수 랜덤포레스트 중요도



- **결측치 처리 후, 숫자형 변수 랜덤포레스트 중요도를 살펴보면, PRICE, TRAVDIST, YEAR_adj가 상위권에 위치하는 것을 알 수 있음.**

8. 문자형 변수 중요도



- 문자형 변수 랜덤포레스트 중요도를 살펴보면, CARNM, COLOR, USEUSENM, FUELNM 이 상위권에 위치함.
- GOODNO(차량ID) 과 CHASNO(차대번호)은 각 행마다 unique한 값으로 인덱스와 비슷한 역할을 한다. 큰 의미가 없을 것으로 생각함.

9. 최종 FEATURE 선정

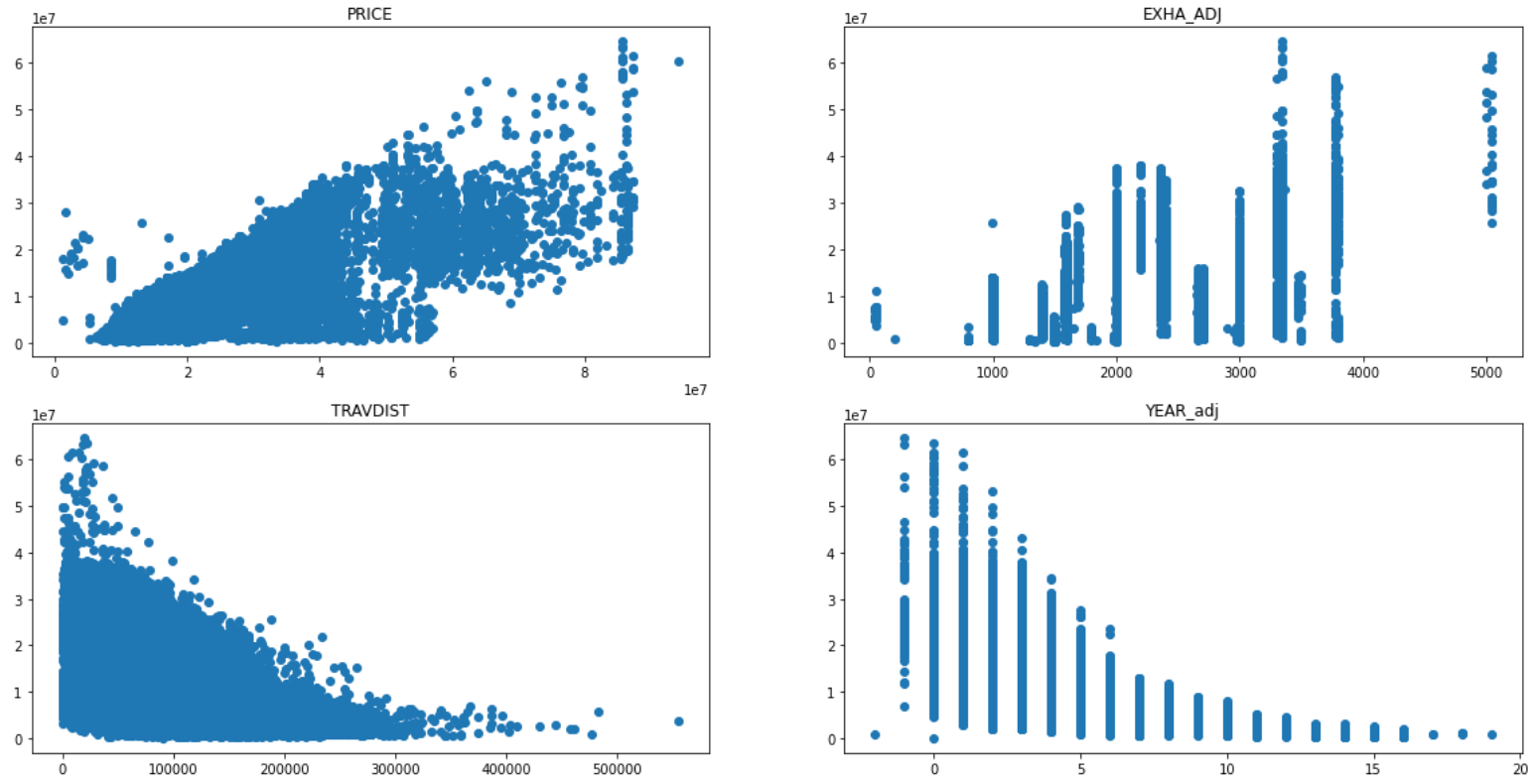
최종 FEATURE

PRICE	EXHA_adj	TRAVDIST	YEAR_adj
11310000.0	1000.0	38480	5
19750000.0	1600.0	62240	3
19340000.0	1591.0	37926	2

CARNM	COLOR	USEUSENM	FUELNM
1(모닝)	1(C)	1(자가)	1(LPG)
2(K3 Nobless)	2(A)	1(자가)	2(가솔린)
3(K3 Trendy)	2(A)	1(자가)	2(가솔린)

10. 이상치 확인

이상치 확인(Y축: SUCCPRIC)



1. SPLIT , NORMALIZE, GRID SEARCH

1)SPLIT DATA

Train set: Validation set=4:1

2)Z_NORMALIZE

2-1)Train set >>Z_normalize

2-2)Validation set>> Z_normalize
(Train set의 mean, std 이용)

3)GRID SEARCH(랜덤 포레스트, 기준:MSE)

max_depths = [5, 50, 100]
n_estimators = [5, 50, 100]

max_depth / n_estimators	5	50	100
5	4,324	1,423	1,423
50	3,957	1,222	1,222
100	3,926	1,204	1,204

($x10^9$)

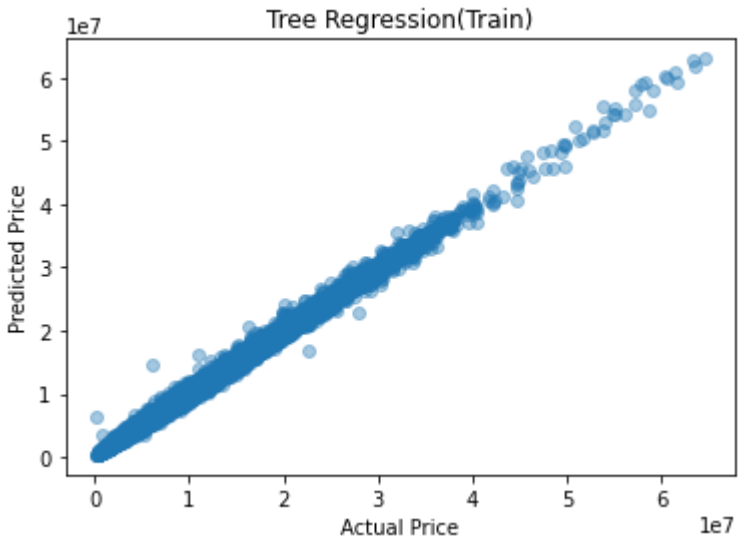
{'max_depth': 50, 'n_estimator': 100}

결과 확인

1. RMSE, MAPE

최종 결과

Train RMSE: 413,263
Train MAPE: 4.22



Validation RMSE: 1,097,712
Validation MAPE: 10.92

End of Document