

Applied Data Analytics

중고차 시세 예측 프로젝트

2025.04.10

Team 3

2020028713 김건우

2020009752 김재엽

2020078759 박현준

2020067356 장윤수

2020051094 정상운



FINANCIAL INNOVATION
& ANALYTICS LAB.

CONTENTS

1. 중고차 시장의 이해
2. 데이터 이해
3. 데이터 전처리
4. 머신러닝 모델 적용 및 평가
5. 보완할 점

발표 흐름

1. 중고차 시장의 이해

- 현대 글로비스
- 중고차 시장 현황

2. 데이터 이해

- 데이터 분류

3. 데이터 전처리

- Drop
- 그대로 사용
- Label Encoding
- 파생 변수 생성
- 가격 변수 처리

4. 머신러닝 모델 적용

- Random Forest
- AdaBoost
- Light GBM

5. 보완점

- 변수 파악의 부족

1. 중고차 시장의 이해

현대글로비스의 오토벨

- 오토벨
 - 현대글로비스의 온라인 중고차 통합 플랫폼



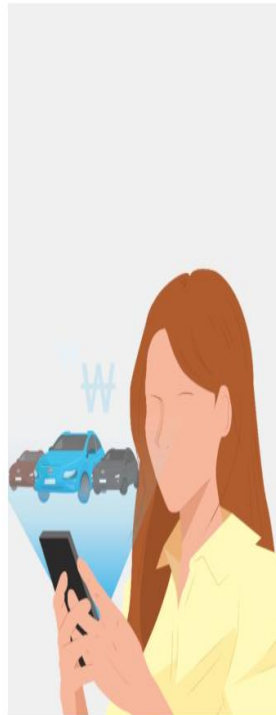
내 차 팔기 서비스

- 전문 차량 평가사 운영
- 편리한 매각 프로세스



내 차 사기 서비스

- 허위매물 필터링 서비스 운영
- 주요 매매 단지·대형업체 제휴 차량 등록



시세조회 서비스

- AI머신러닝 알고리즘 기반 시세 제공
- 내 차 팔기/사기 연계 실용적 시세 제공

- 오토벨 스마트옥션
 - 현대글로비스의 중고차 경매 서비스



국내 최대 규모의 중고차 경매장 운영

- 2,400여 개의 중고차 매매업체 경쟁 입찰
- 투명한 상품정보 제공, 전문 인력 운용

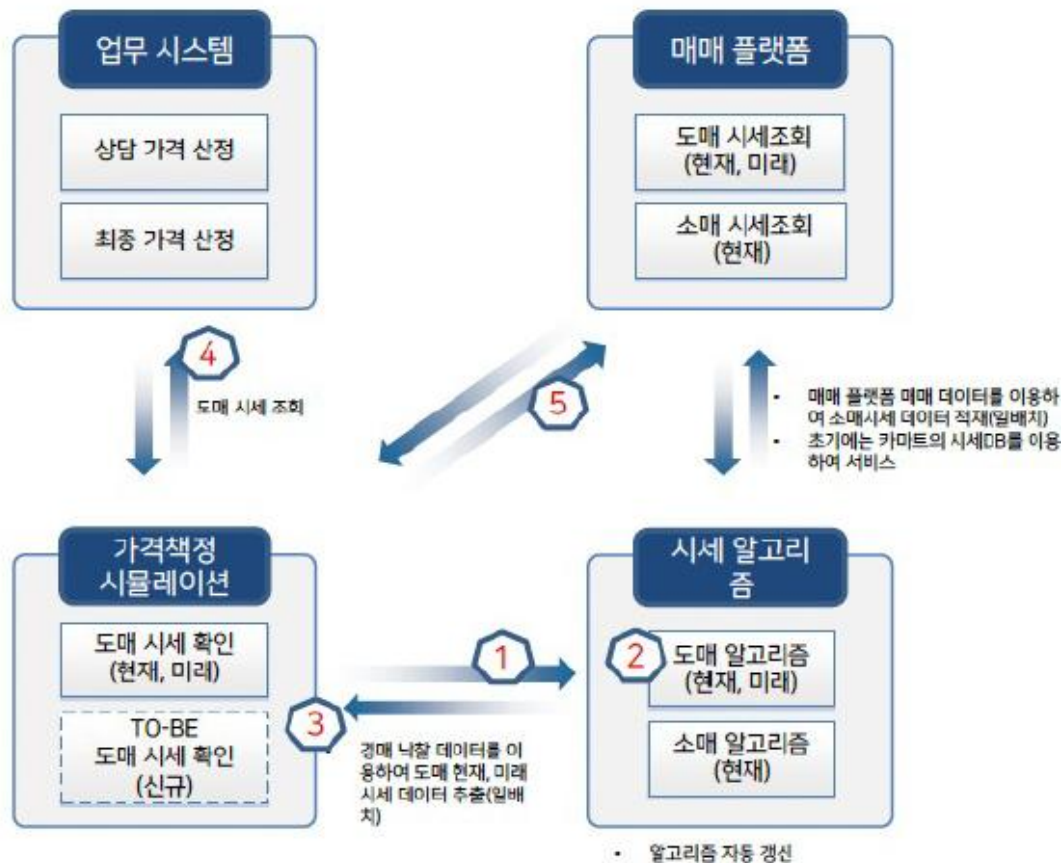


비대면 디지털 경매 시스템 '오토벨 스마트 옥션'

- 클라우드 서비스를 이용한 비대면 중고차 경매시스템
- 증강현실 및 가상현실의 3차원 정보 제공

현대글로비스의 오토비즈

- 거래 시스템



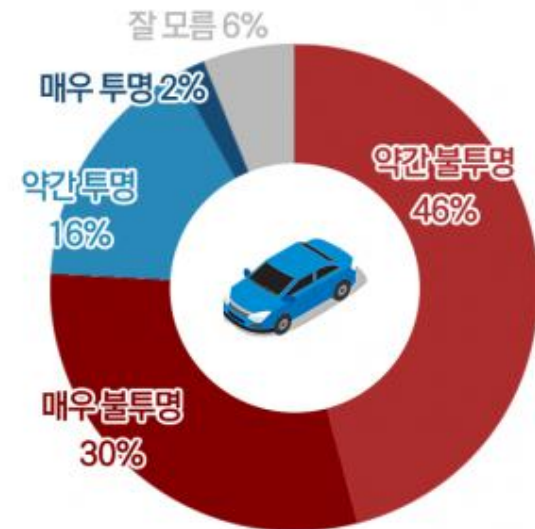
중고차 시장 현황

- 중고차 시장 규모



- 중고차 거래 문제

중고차 시장 인식조사 결과



자료 한국경제연구원/하나증권 ※ 전국 만 19세 이상 1000명 대상

2. 데이터 이해

교환(손상) 여부(60)

BONET	본넷교환
FRONT_LEFT_FENDER	앞팬더(좌)교환
FRONT_RIGHT_FENDER	앞팬더(우)교환
FRONT_LEFT_DOOR	앞도어(좌)교환
FRONT_RIGHT_DOOR	앞도어(우)교환
BACK_LEFT_DOOR	뒷도어(좌)교환
BACK_RIGHT_DOOR	뒤소어(우)교환
TRUNK	트렁크교환
FRONT_PANNEL	앞패널교환
LEFT_STEP	스텝(좌)교환
RIGHT_STEP	스텝(우)교환
LEFT_FILER_A	A필러(좌)교환
RIGHT_FILER_A	A필러(우)교환
LEFT_FILER_B	B필러(좌)교환
RIGHT_FILER_B	B필러(우)교환
LEFT_FILER_C	C필러(좌)교환
RIGHT_FILER_C	C필러(우)교환
LEFT_REAR_FENDER	리어팬더(좌)교환
RIGHT_REAR_FENDER	리어팬더(우)교환
BACK_PANEL1	뒷패널교환
LEFT_INSIDE_PANEL	인사이드패널(좌)교환
RIGHT_INSIDE_PANEL	인사이드패널(우)교환
LEFT_WHEEL_HOUSE	휠하우스(좌)교환
RIGHT_WHEEL_HOUSE	휠하우스(우)교환
LEFT_INSIDE_WHEEL_HOUSE	리어사이드패널(좌)교환
RIGHT_INSIDE_WHEEL_HOUSE	리어사이드패널(우)교환
LEFT_REAR_WHEEL_HOUSE	리어휠하우스(좌)교환
RIGHT_REAR_WHEEL_HOUSE	리어휠하우스(우)교환
TRUNK_FLOOR	트렁크플로어교환
DASH_PANEL	대시패널교환
SHEET_PANEL	시트백패널교환
SIDE_MEMBER_FRAME	사이드멤버(프레임)교환

LEFT_QUARTER	쿼터패널(좌)교환
RIGHT_QUARTER	쿼터패널(좌)교환
FLOOR_PANEL	플로어패널교환
LEFT_SIDE_PANEL	사이드패널(좌)교환
RIGHT_SIDE_PANEL	사이드패널(우)교환
LEFT_REAR_CORNER_PANEL	리어코너패널(좌)교환
RIGHT_REAR_CORNER_PANEL	리어코너패널(우)교환
BACK_PANEL2	백패널교환
LEFT_CORNER_PANEL	코너패널(좌)교환
RIGHT_CORNER_PANEL	코너패널(우)교환
LEFT_SKIRT_PANEL	스커트패널(좌)교환
RIGHT_SKIRT_PANEL	스커트패널(우)교환
SIDE_MEMBER_FRAME2	사이드멤버(프레임)2교환
LEFT_INSIDE_SHEETING	인사이드패널(좌)판금/용접
RIGHT_INSIDE_SHEETING	인사이드패널(우)판금/용접
LEFT_WHEEL_HOUSE_SHEETING	휠하우스(좌)판금/용접
RIGHT_WHEEL_HOUSE_SHEETING	휠하우스(우)판금/용접
LEFT_REAR_INSIDE_PANEL_SHEETING	리어인사이드패널(좌)판금/용접
RIGHT_REAR_INSIDE_PANEL_SHEETING	리어인사이드패널(우)판금/용접
LEFT_REAR_WHEEL_HOUSE_SHEETING	리어휠하우스(좌)판금/용접
RIGHT_REAR_WHEEL_HOUSE_SHEETING	리어휠하우스(우)판금/용접
TRUNK_FLOOR_SHEETING	트렁크플로어판금/용접
DASH_PANEL_SHEETING	대시패널판금/용접
SHEET_BACK_PANEL_SHEETING	시트백패널판금/용접
SIDE_MEMBER_FRAME_SHEETING	사이드멤버(프레임)판금/용
FLOOR_PANEL_SHEETING	플로어패널판금/용접
LEFT_SIDE_PANEL_SHEETING	사이드패널(좌)판금/용접
RIGHT_SIDE_PANEL_SHEETING	사이드패널(우)판금/용접

옵션 여부(15)

ABS	ABS여부
AB2	AB2여부
NAVIGATION	네비게이션 여부
VDC	VDC 여부
SMARTKEY	스마트키 여부
SUNLOPPANORAMA	파노라마선루프 여부
SUNLOOPCOMMON	일반선루프 여부
SUNLOOPDUAL	듀얼선루프 여부
DIS	DIS 여부
TCS	TCS 여부
AB1	AB1 여부
ETC	ETC 장착 여부
AV	AV 여부
EPS	EPS 여부
ECS	ECS 여부

위험 차량 여부(4)

FLOODING	침수
TOTAL_LOSS	전손
JOINCAR	접합차
NOTAVAILABLE	운행불가

차량 키 (6)

MF_KEY	제조사키
MJ_MODEL_KEY	대표모델키
DT_MODEL_KEY	세부모델키
MJ_GRADE_KEY	대표등급키
DT_GRADE_KEY	세부등급키
NC_GRADE_KEY	신차등급키

기본 차량 정보(15)

GOODNO	차량ID
SUCCYMD	낙찰일자
CARNM	차량명
CHASNO	차대번호
CARREGIYMD	차량등록일
YEAR	년식
MISSNM	미션명
FUELNM	연료명
COLOR	색상
EXHA	배기량
TRAVDIST	주행거리
USEUSENM	용도명
OWNECLASNM	소유명
INNEEXPOCLASCD_YN	내수수출구분
YEARCHK	년식차량구분

가격 정보(4)

NEWCARPRIC	신차금액
SUCCPRIC	낙찰가
SHIPPING_PRICE	출고가
NC_GRADE_PRICE	신차등급가격



SUCCPRIC	낙찰가
----------	-----

3. 데이터 전처리

NaN 값 확인 및 처리

- NaN 값

- CARREGIYMD(차량등록일)
- FUELNМ(연료명)
- USEUSENM(용도명)
- OWNECLSNM(소유명)
- SHIPPING_PRICE(출고가)
- NC_GRADE_PRICE(신차등급 가격)



```

CARREGIYMD      1
FUELNМ          1
USEUSENM        323
OWNECLSNM       13
SHIPPING_PRICE  4325
NC_GRADE_PRICE  3781
dtype: int64

```

- NaN 값 처리

- CARREGIYMD(차량등록일) -> DROP
- FUELNМ(연료명) -> '가솔린'
- USEUSENM(용도명) -> '미상'
- OWNECLSNM(소유명)-> '개인'
- SHIPPING_PRICE(출고가)
- NC_GRADE_PRICE(신차등급 가격)

예측에 영향을 주지 않을 데이터 Drop

- 제거한 변수
 - GOODNO: 차량 ID
 - CHASNO: 차대번호
 - MF_KEY: 제조사키
 - DT_MODEL_KEY: 대표모델키
 - MJ_GRADE_KEY: 대표등급키
 - DT_GRADE_KEY: 세부등급 키
 - DC_GRADE_KEY: 신차등급 키

데이터 값 그대로 사용

- NOTAVAILABLE
- FLOODING
- TOTAL_LOSS
- JOINCAR



```
[NOTAVAILABLE] unique values:
NOTAVAILABLE
0    36793
Name: count, dtype: int64

[FLOODING] unique values:
FLOODING
0    34324
1     2469
Name: count, dtype: int64

[TOTAL_LOSS] unique values:
TOTAL_LOSS
0    36780
1         13
Name: count, dtype: int64

[JOINCAR] unique values:
JOINCAR
0    36793
Name: count, dtype: int64
```

Train Data



```
[NOTAVAILABLE] unique values:
NOTAVAILABLE
0     20
Name: count, dtype: int64

[FLOODING] unique values:
FLOODING
0     20
Name: count, dtype: int64

[TOTAL_LOSS] unique values:
TOTAL_LOSS
0     20
Name: count, dtype: int64

[JOINCAR] unique values:
JOINCAR
0     20
Name: count, dtype: int64
```

Test Data

Label Encoding

- Label Encoding은 범주형 데이터를 숫자로 변환하여 모델이 처리할 수 있도록 만드는 방법
 - 범주형 변수(Categorical Variable)를 처리하기 위해 라벨 인코딩 진행.
 - 모든 데이터를 **수치형 변수**로 변환 (예: -1, 0, 1, 2 ...)

CARNM (차량명)

→ CAR_NAME_ENC

COLOR (색상)

→ COLOR_ENC

YEARCHK (년식 차량 구분)

→ YEARCHK_ENC

INNEEXPOCLASCD_YN (내수수출구분)

→ INNEEXPOCLASCD_YN_ENC

USEUSENM (용도명)

→ USEUSENM_ENC

OWNECLASNM (소유명)

→ OWNECLASNM_ENC

MISSNM (미션명 / 변속기)

→ MISSNM_ENC

FUELNM (연료명)

→ FUELNM_ENC

년식 차량 구분 및 내수 수출 구분

- YEARCHK -> YEARCHK_ENC
- INNEEXPOCLASCD_YN -> INNEEXPOCLASCD_YN_ENC
 - 상기 변수들에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [Y, N]
 - 인코딩: [0, 1]



✓ df['YEARCHK'] 라벨 인코딩 완료

📄 YEARCHK 인코딩 매핑:

YEARCHK	YEARCHK_ENC
0	N
1	Y

✓ df['INNEEXPOCLASCD_YN'] 라벨 인코딩 완료

📄 인코딩 매핑 확인:


INNEEXPOCLASCD_YN	INNEEXPOCLASCD_YN_ENC
6	0
0	X

YEARCHK	YEARCHK_ENC
N	0
Y	1

INNEEXPOCLASCD_YN	INNEEXPOCLASCD_YN_ENC
O	0
X	1

차량명

- CARNM -> CAR_NAME -> CAR_NAME_ENC
 - 차량명의 앞 글자만 추출 후 CAR_NAME으로 저장
 - 차량명에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [모닝, K3, K5, ...] 49개
 - 인코딩: [29, 2, 3, ...]



	CAR_NAME	CAR_NAME_ENC
0	모닝	29
1	K3	2
2	K3	2
3	K5	3
4	K5	3
...
36789	더뉴모닝	21
36790	더뉴K9	19
36791	더뉴K9	19
36792	더뉴K5	17
36793	더뉴K5	17

[36793 rows x 2 columns]

CAR_NAME	CAR_NAME_ENC
모닝	29
K3	2
K5	3
더뉴모닝	21
더뉴K5	19
더뉴K9	17

색상

- COLOR -> COLOR_ENC
 - 색상에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [A, B, C, D, E]
 - 인코딩: [0, 1, 2, 3, 4]




	COLOR	COLOR_ENC
0	C	2
1	A	0
2	A	0
3	B	1
4	D	3
...
36789	A	0
36790	D	3
36791	B	1
36792	A	0
36793	A	0

[36793 rows x 2 columns]

COLOR	COLOR_ENC
A	0
B	1
C	2
D	3
E	4

용도명

- USEUSENM -> USEUSENM_ENC
 - 용도명에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [렌트, 리스, 미상, 사업, 업무, 자가]
 - 인코딩: [0, 1, 2, 3, 4, 5]




	USEUSENM	USEUSENM_ENC
3	렌트	0
18	리스	1
1327	미상	2
184	사업	3
14	업무	4
0	자가	5

USEUSENM	USEUSENM_ENC
렌트	0
리스	1
미상	2
사업	3
업무	4
자가	5

소유명

- OWNECLASNM -> OWNECLASNM_ENC
 - 소유명에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [개인, 개인사업, 법인, 법인상품, 상품용, 재외국인, 종교단체]
 - 인코딩: [0, 1, 2, 3, 4, 5, 6]



	OWNECLASNM	OWNECLASNM_ENC
7	개인	0
3437	개인사업	1
0	법인	2
13	법인상품	3
39	상품용	4
1905	재외국인	5
25782	종교단체	6

OWNECLASNM	OWNECLASNM_ENC
개인	0
개인사업	1
법인	2
법인상품	3
상품용	4
재외국인	5
종교단체	6

미션명(변속기)

- MISSNM -> MISSNM_ENC
 - 미션명에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [A/T, CVT, M/T]
 - 인코딩: [0, 1, 2]



	MISSNM	MISSNM_ENC
0	A/T	0
20	CVT	1
39	M/T	2

MISSNM	MISSNM_ENC
A/T	0
CVT	1
M/T	2

연료명

- FUELNAM -> FUELNAM_ENC
 - 연료명에 대해 고유한 숫자를 부여
 - » Train과 test 모두 일치한 숫자가 부여되도록
 - » Train에 없고, test에 존재하는 변수의 경우 '기타' 처리 후 -1로 인코딩
 - 고유값: [Hybrid, LPG, 가솔린, 경용, 디젤, 전기]
 - 인코딩: [0, 1, 2, 3, 4, 5]

	FUELNAM	FUELNAM_ENC
8	Hybrid	0
0	LPG	1
1	가솔린	2
19	경용	3
15	디젤	4
165	전기	5

FUELNAM	FUELNAM_ENC
Hybrid	0
LPG	1
가솔린	2
경용	3
디젤	4
전기	5

COUNT 변수 생성

- OPTION_COUNT
 - 옵션 관련 15개의 변수를 하나로 통합
 - **탑재 옵션의 개수**를 나타내는 OPTION_COUNT 변수 생성

✓ OPTION 여부 통합해서 1개 변수로 저장

```
[ ] 1 option_cols = [  
2     'ABS', 'AB2', 'NAVIGATION', 'VDC', 'SMARTKEY',  
3     'SUNLOOPPANORAMA', 'SUNLOOPCOMMON', 'SUNLOOPDUAL',  
4     'DIS', 'TCS', 'AB1', 'ETC', 'AV', 'EPS', 'ECS'  
5 ]
```

```
[ ] 1 # Train  
2 df['OPTION_COUNT'] = df[option_cols].fillna(0).sum(axis=1)  
3
```


COUNT 변수 생성

- DAMAGE_COUNT

- 교환 이력은 해당 부위에 손상이 있었던 것으로 판단
- 차량 손상 관련 60개의 변수를 하나로 통합
- 차량 손상 부위의 개수를 나타내는 DAMAGE_COUNT 변수 생성

✓ 교환 여부 통합해서 1개 변수로 저장

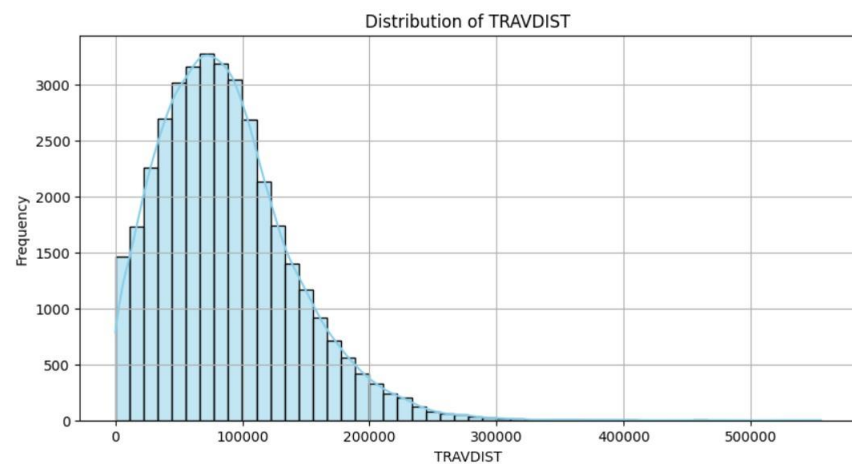
```
[ ] 1 # 1. 60개 변수 이름 리스트
    2 damage_cols = [
    3     'BONET', 'FRONT_LEFT_FENDER', 'FRONT_RIGHT_FENDER',
    4     'FRONT_LEFT_DOOR', 'FRONT_RIGHT_DOOR', 'BACK_LEFT_DOOR', 'BACK_RIGHT_DOOR',
    5     'TRUNK', 'FRONT_PANNEL', 'LEFT_STEP', 'RIGHT_STEP',
    6     'LEFT_FILER_A', 'RIGHT_FILER_A', 'LEFT_FILER_B', 'RIGHT_FILER_B',
    7     'LEFT_FILER_C', 'RIGHT_FILER_C', 'LEFT_REAR_FENDER', 'RIGHT_REAR_FENDER',
    8     'BACK_PANEL1', 'LEFT_INSIDE_PANEL', 'RIGHT_INSIDE_PANEL',
    9     'LEFT_WHEEL_HOUSE', 'RIGHT_WHEEL_HOUSE', 'LEFT_INSIDE_WHEEL_HOUSE',
   10     'RIGHT_INSIDE_WHEEL_HOUSE', 'LEFT_REAR_WHEEL_HOUSE', 'RIGHT_REAR_WHEEL_HOUSE',
   11     'TRUNK_FLOOR', 'DASH_PANEL', 'SHEET_PANEL', 'SIDE_MEMBER_FRAME',
   12     'LEFT_QUARTER', 'RIGHT_QUARTER', 'FLOOR_PANEL', 'LEFT_SIDE_PANEL',
   13     'RIGHT_SIDE_PANEL', 'LEFT_REAR_CORNER_PANEL', 'RIGHT_REAR_CORNER_PANEL',
   14     'BACK_PANEL2', 'LEFT_CORNER_PANEL', 'RIGHT_CORNER_PANEL',
   15     'LEFT_SKIRT_PANEL', 'RIGHT_SKIRT_PANEL', 'SIDE_MEMBER_FRAME2',
   16     'LEFT_INSIDE_SHEETING', 'RIGHT_INSIDE_SHEETING',
   17     'LEFT_WHEEL_HOUSE_SHEETING', 'RIGHT_WHEEL_HOUSE_SHEETING',
   18     'LEFT_REAR_INSIDE_PANEL_SHEETING', 'RIGHT_REAR_INSIDE_PANEL_SHEETING',
   19     'LEFT_REAR_WHEEL_HOUSE_SHEETING', 'RIGHT_REAR_WHEEL_HOUSE_SHEETING',
   20     'TRUNK_FLOOR_SHEETING', 'DASH_PANEL_SHEETING',
   21     'SHEET_BACK_PANEL_SHEETING', 'SIDE_MEMBER_FRAME_SHEETING',
   22     'FLOOR_PANEL_SHEETING', 'LEFT_SIDE_PANEL_SHEETING', 'RIGHT_SIDE_PANEL_SHEETING'
   23 ]
```

```
[ ] 1 df['DAMAGE_COUNT'] = df[damage_cols].fillna(0).sum(axis=1)
```

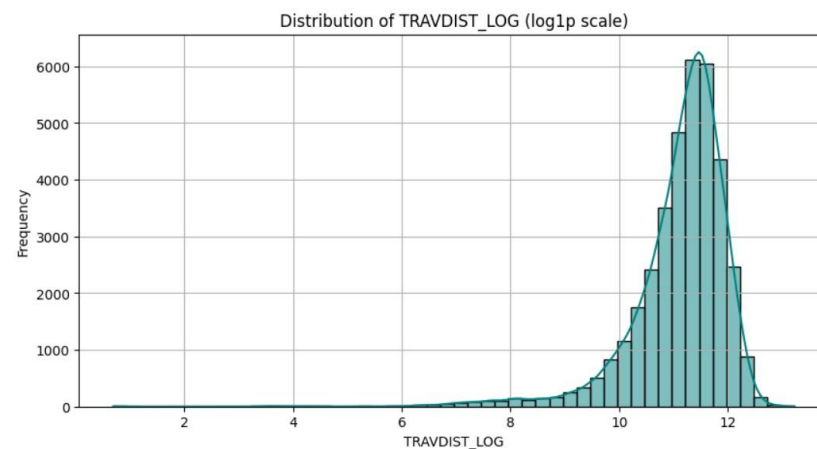
주행거리

- TRAVDIST
 - 주행거리 변수
 - Skewness 비교
 - » TRAVDIST: 0.9345
 - » TRAVDIST_LOG: -2.3562

41

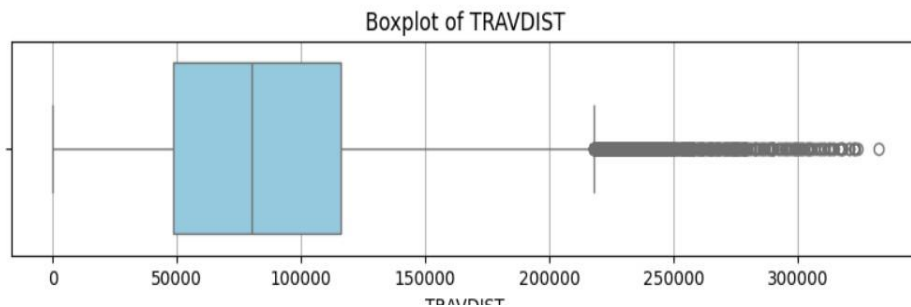
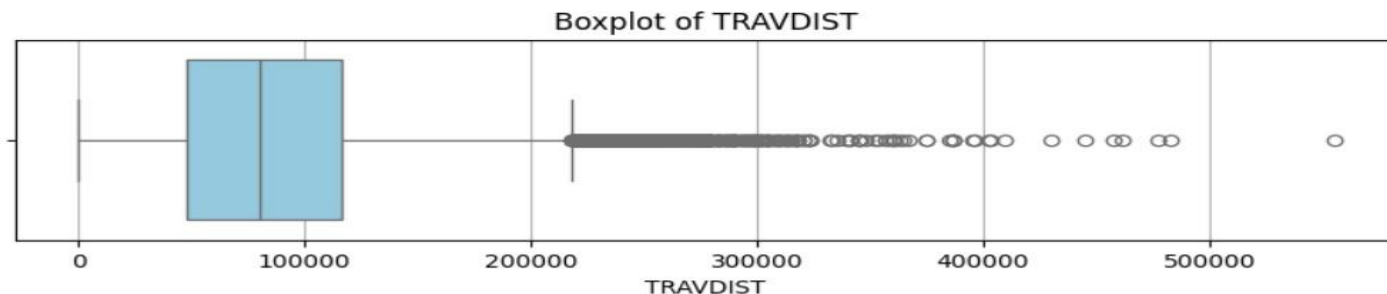


42



주행거리

- TRAVDIST
 - 주행거리 변수
 - BOXPLOT 확인
 - » 상위 0.01% 극단적 이상치(332544.6) 제거 -> Test set에는 이상치 존재하지 않음.
 - » Median 값 대체 (80270.5)



연간 주행거리 변수 생성

- ANNUAL_TRAVDIST
 - 차량 사용 년도= 낙찰년도 - 차량등록년도 (최소 값은 1)
 - » $CAR_AGE = SUCCYMD_YEAR - CARREGIYMD_YEAR$
 - 연간 주행 거리= 주행거리 / 차량 사용 년도
 - » $ANNUAL_TRAVDIST = TRAVDIST / CAR_AGE$



	ANNUAL_TRAVDIST
0	6413.333333
1	20746.666667
2	18963.000000
3	27537.250000
4	16335.000000
...	...

사용일수 변수 생성

- USED_DAY
 - 사용일수 = 낙찰일자 - 차량 등록일
 - » USED_DAY = SUCCYMD - CARREGIYMD
 - 날짜(일)로 계산



	USED_DAYS
0	2029
1	1062
2	707
3	1113
4	1713
...	...

Season 변수 생성

- AUCTION_QUARTER_ENC
 - 낙찰일자에서 4개의 기간 구분 후 Label Encoding
 - » Q1: 1~3월, 비수기라 가격 낮을 가능성 존재
 - » Q2: 4~6월
 - » Q3: 7~9월, 여름방학 및 휴가철은 성수기라 가격 높을 가능성 존재
 - » Q4: 10~12월, 연식 바뀌기 직전으로 할인이 많아 가격 낮을 가능성 존재

↔ AUCTION_QUARTER SUCCYMD

Q1	20160105
Q1	20160106
Q1	20160107
Q2	20160401
Q2	20160402
Q2	20160405
Q3	20160701
Q3	20160702
Q3	20160705
Q4	20161001
Q4	20161004
Q4	20161005

📄 AUCTION_QUARTER ↔ AUCTION_QUARTER_ENC 매핑

AUCTION_QUARTER	AUCTION_QUARTER_ENC
0	Q1
2510	Q2
5319	Q3
7739	Q4

배기량

- EXHA
 - 배기량에 따른 차종 구분 [국내 자동차 관리법 참조]
 - » 700~1100: 경차
 - » 1101~1600: 소형/준중형
 - » 1601~2200: 중형
 - » 2201~3000: 대형
 - » 30001~5000: 고성능/수입차
 - » 700미만, 5000초과 : 이상치
 - 이상치는 중앙값(1600.0)으로 대체 -> Test set에는 이상치 존재하지 않음.



🚗 df - EXHA 분포 (최종 기준)

- 경차 (700~1100): 12121
- 소형/준중형 (1101~1600): 6301
- 중형 (1601~2200): 10557
- 대형 (2201~3000): 5525
- 고성능/수입차 (3001~5000): 2209
- ! 이상치 (<700 or >5000): 80

Train Data



🚗 df - EXHA 분포 (최종 기준)

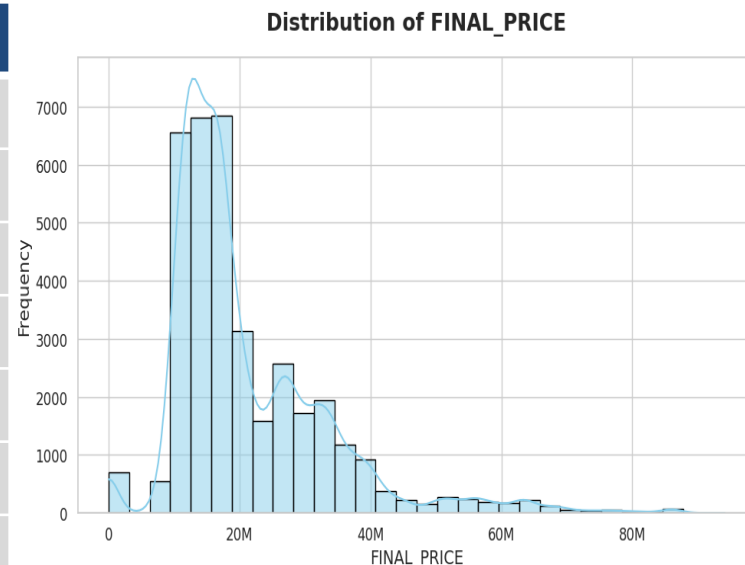
- 경차 (700~1100): 10
- 소형/준중형 (1101~1600): 2
- 중형 (1601~2200): 7
- 대형 (2201~3000): 0
- 고성능/수입차 (3001~5000): 1
- ! 이상치 (<700 or >5000): 0

Test Data

가격 변수 처리

- FINAL_PRICE
 - SHIPPING_PRICE(출고가), NC_GRADE_PRICE(신차등급가격), NEWCARPRIC(신차금액) 이용
 - 우선 순위에 따라 결정
 - » 1. SHIPPING_PRICE 사용
 - » 2. SHIPPING_PRICE가 NaN일 경우, NC_GRADE_PRICE 사용
 - » 3. NC_GRADE_PRICE가 NaN일 경우, NEWCARPRIC 사용
- FINAL_PRICE 분포 확인

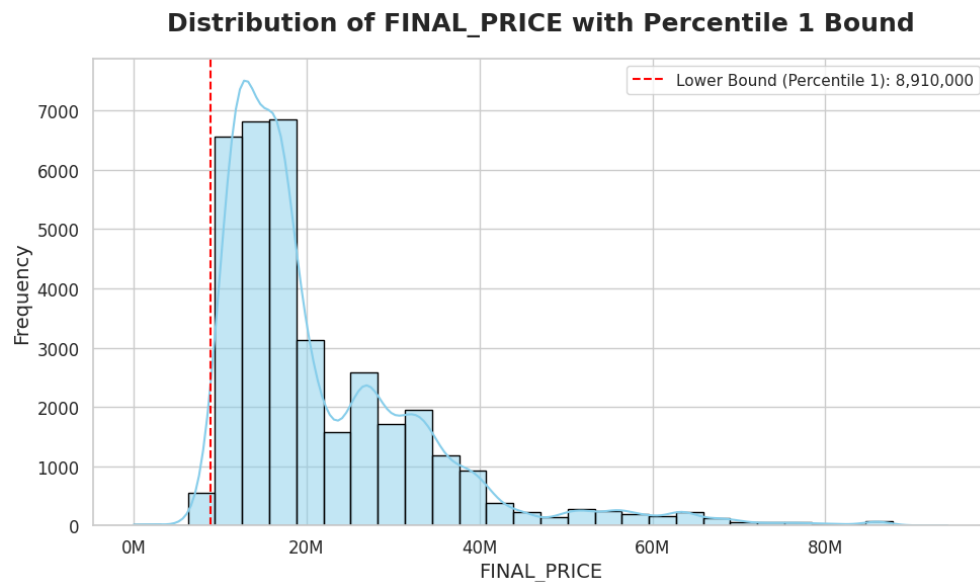
금액	빈도 수	금액	빈도 수
0원	412	5,000만원~6,000만원	762
1원	277	6,000만원~7,000만원	507
2~1,000만원 이하	1,052	7,000만원~8,000만원	141
1,000만원~2,000만원	21,386	8,000만원~9,000만원	105
2,000만원~3,000만원	6,634	9,000만원~1억원	1
3,000만원~4,000만원	4,617	1억원 초과	0
4,000만원~5,000만원	500	NaN	0



가격 변수 처리

- FINAL_PRICE
 - 0원, 1원을 제외 후 Lower Bound와 Median 값 판단
 - » 하위 1%를 Lower Bound로 판단 (891만원)
 - » Median= 1,713만원

하한	Value
Lower Bound(하위 1%)	891만원
Lower Bound 미만 개수	348



가격 변수 처리

- FINAL_PRICE
 - Test set에는 하한 값(891만) 미만인 4개 존재
 - Train set의 Median 값(1,713만)으로 대체

FINAL_PRICE			
0	17130000.0	10	11890000.0
1	17130000.0	11	11890000.0
2	17130000.0	12	11890000.0
3	24050000.0	13	11890000.0
4	58670000.0	14	11890000.0
5	19460000.0	15	11890000.0
6	16050000.0	16	12110000.0
7	26130000.0	17	15700000.0
8	11890000.0	18	17670000.0
9	19460000.0	19	17130000.0

전처리

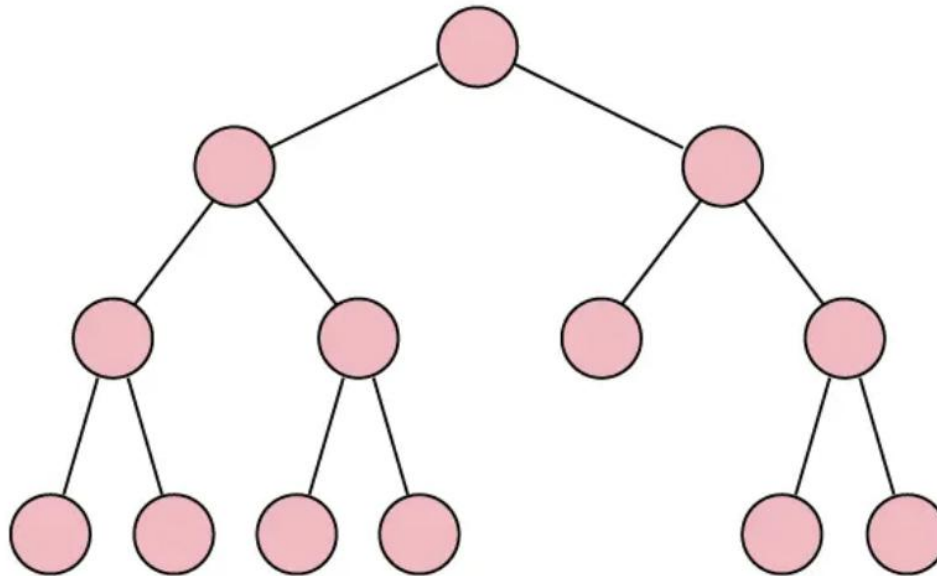
- 총 20개 변수 사용
 - EXHA, TRAVDIST, FLOODING, TOTAL_LOSS, JOINCAR, NOTAVAILABLE, CAR_NAME_ENC, DAMAGE_COUNT, OPTION_COUNT, COLOR_ENC, MISSNM_ENC, USEUSENM_ENC, OWNECLASNM_ENC, FUELMN_ENC, YEARCHK_ENC, INNEEXPOCLASCD_ENC, ANNUAL_TRAVDIST, AUCTION_QUARTER_ENC, USED_DAYS, FINAL_PRICE

	EXHA	TRAVDIST	SUCCPRIC	FLOODING	TOTAL_LOSS	JOINCAR	NOTAVAILABLE	CAR_NAME_ENC	DAMAGE_COUNT	OPTION_COUNT	...	MISSNM_ENC	USEUSENM_ENC	OWNECLASNM_ENC	FUELMN_ENC	YEARCHK_ENC	INNEEXPOCLASCD_YN_ENC	ANNUAL_TRAVDIST	AUCTION_QUARTER_ENC	USED_DAYS	FINAL_PRICE
0	1000.0	38480.0	4300000	0	0	0	0	29	0	1	...	0	5	2	1	0	1	6413.333333	0	2029	11310000.0
1	1600.0	62240.0	11650000	0	0	0	0	2	0	6	...	0	5	2	2	1	1	20746.666667	0	1062	19750000.0
2	1591.0	37926.0	12350000	0	0	0	0	2	0	6	...	0	5	2	2	1	1	18963.000000	0	707	19340000.0
3	2000.0	110149.0	5900000	0	0	0	0	3	0	2	...	0	0	2	1	0	1	27537.250000	0	1113	17680000.0
4	2000.0	81675.0	4730000	0	0	0	0	3	6	2	...	0	0	2	1	1	1	16335.000000	0	1713	15800000.0
...
36788	998.0	62180.0	5910000	0	0	0	0	21	0	3	...	0	5	3	2	0	1	15545.000000	1	1557	12500000.0
36789	3778.0	97801.0	19200000	0	0	0	0	19	3	5	...	0	5	3	2	0	1	19560.200000	1	1617	56800000.0
36790	3342.0	153601.0	18200000	0	0	0	0	19	2	6	...	0	0	3	2	1	1	38400.250000	1	1478	49089202.0
36791	1999.0	140058.0	5800000	0	0	0	0	17	3	3	...	0	0	3	1	0	1	28011.600000	1	1770	17130000.0
36792	1999.0	159467.0	5700000	0	0	0	0	17	0	3	...	0	0	3	1	0	1	31893.400000	1	1820	17130000.0

4. 머신러닝 모델 적용 및 평가

Machine Learning Model

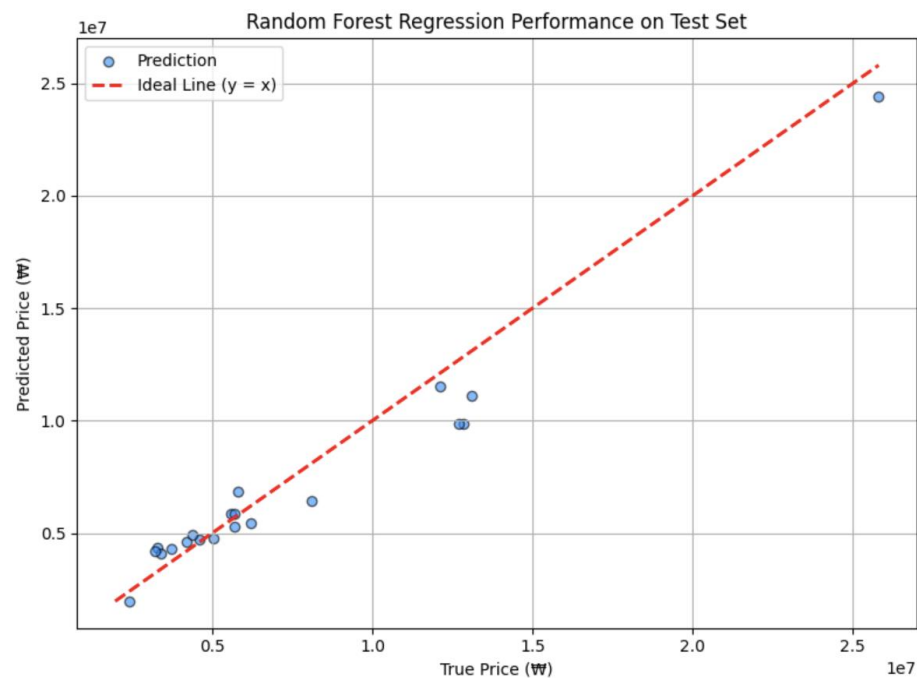
- Random Forest
 - 여러 개의 Decision Tree를 만들어 각각 예측한 후, 그 결과를 모아 최종 예측을 수행함
 - 각 나무는 서로 다른 데이터 샘플과 특성으로 학습되어 모델의 다양성을 높임
 - 과적합을 줄이고, 하나의 모델보다 더 안정적이고 정확한 예측이 가능함



Random Forest

Seed	Train MSE	Valid MSE	Valid MAPE
101	5.116915e+11	1.076974e+12	10.938143
202	4.992836e+11	1.122869e+12	10.912910
303	4.968027e+11	1.188919e+12	12.992526
404	4.964748e+11	1.181315e+12	13.048108
505	5.030496e+11	1.098009e+12	11.475160
평균	5.014604e+11	1.133617e+12	11.87%

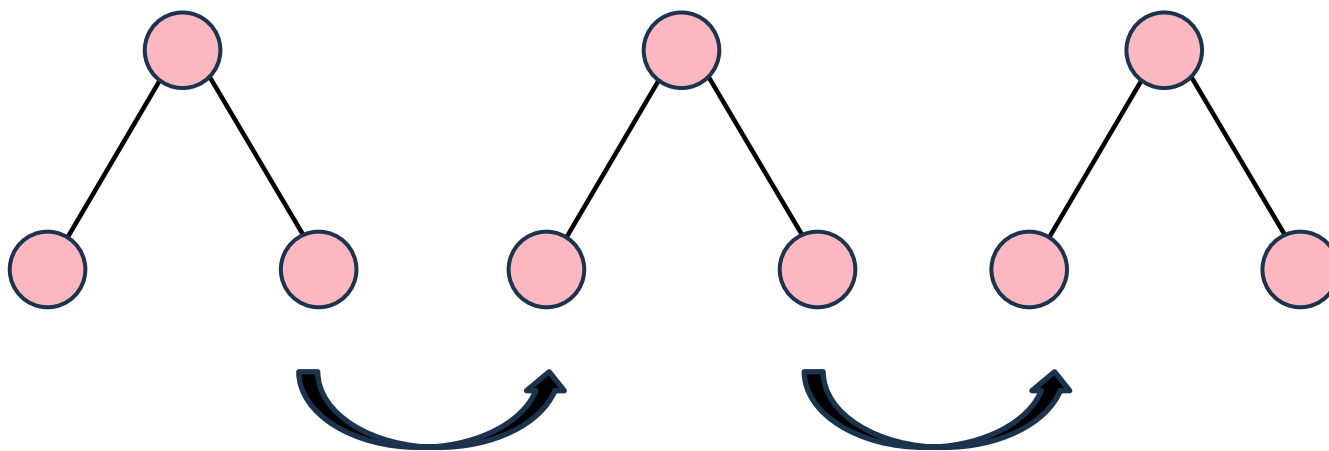
- ▶ n_estimators : 150
- ▶ min_samples_split : 5
- ▶ Max_depth: 10



- ▶ TEST MSE : 1.575605e+12
- ▶ TEST MAE : 958,652.90 (95만 8천원)
- ▶ TEST MAPE : 14.17%

Machine Learning Model

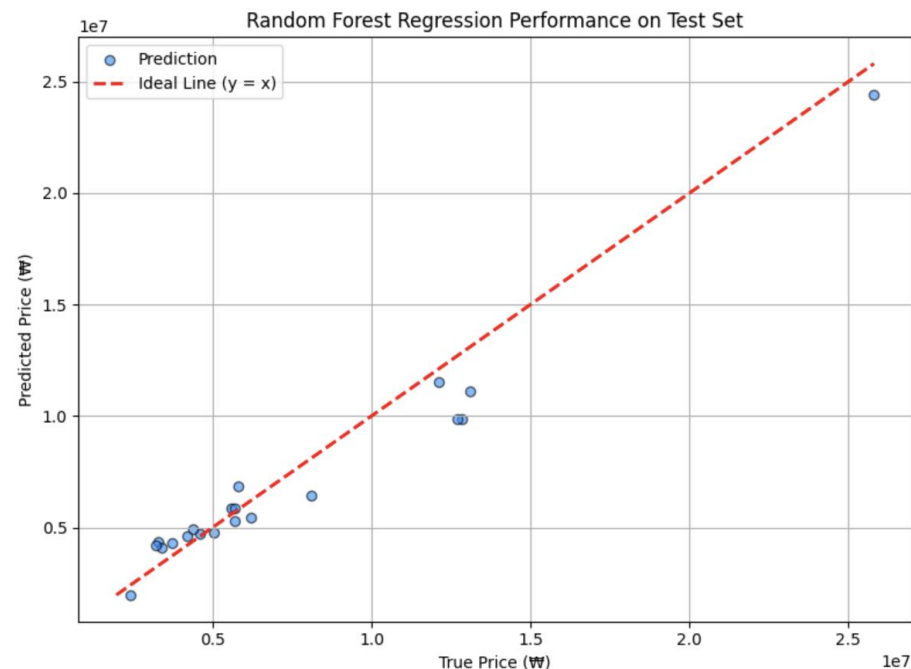
- Adaboost
 - 단순한 모델부터 시작해 오차가 난 데이터를 다음 모델이 보완하는 방식으로 학습함
 - 틀린 예측에 가중치를 높여, 다음 모델이 어려운 데이터에 더 집중할 수 있도록 설계됨
 - 약한 학습기를 순차적으로 연결해 강력한 하나의 앙상블 모델을 구성함



AdaBoost

Seed	Train MSE	Valid MSE	Valid MAPE
101	5.116915e+11	9.732999e+11	10.538922
202	4.992836e+11	9.987109e+11	10.485088
303	4.968027e+11	1.036915e+12	12.427623
404	4.964748e+11	1.057465e+12	12.547614
505	5.030496e+11	9.570505e+11	11.081605
평균	5.014604e+11	1.004688e+12	11.42%

- ▶ n_estimators : 130
- ▶ Learning_Rate : 0.1
- ▶ Max_depth: 10

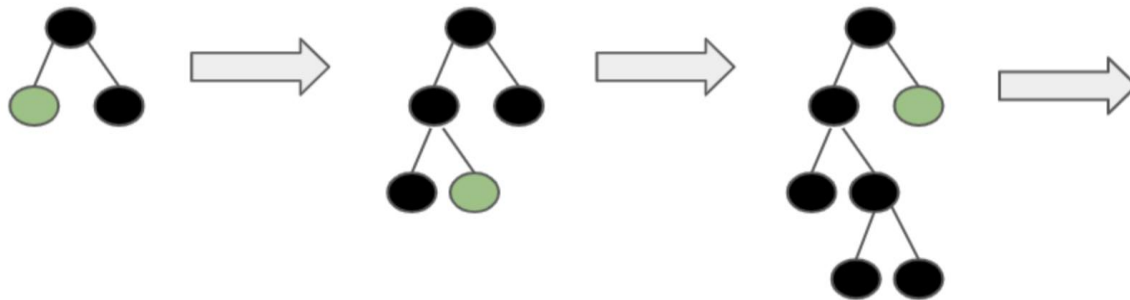


- ▶ TEST MSE : 1.022938e+12
- ▶ TEST MAE : 811,087.80 (81만 1천원)
- ▶ TEST MAPE : 12.59%

Light GBM

- Light GBM (Light Gradient Boosting Machine)
 - Gradient Boosting 기반의 트리 모델로, 속도와 성능을 모두 개선한 모델
 - 여러 개의 약한 학습기(=작은 트리)를 **순차적으로** 학습하여 오차를 보완
 - » **정보 이득이 가장 큰 leaf만** 깊게 자라는 방식

LightGBM leaf-wise

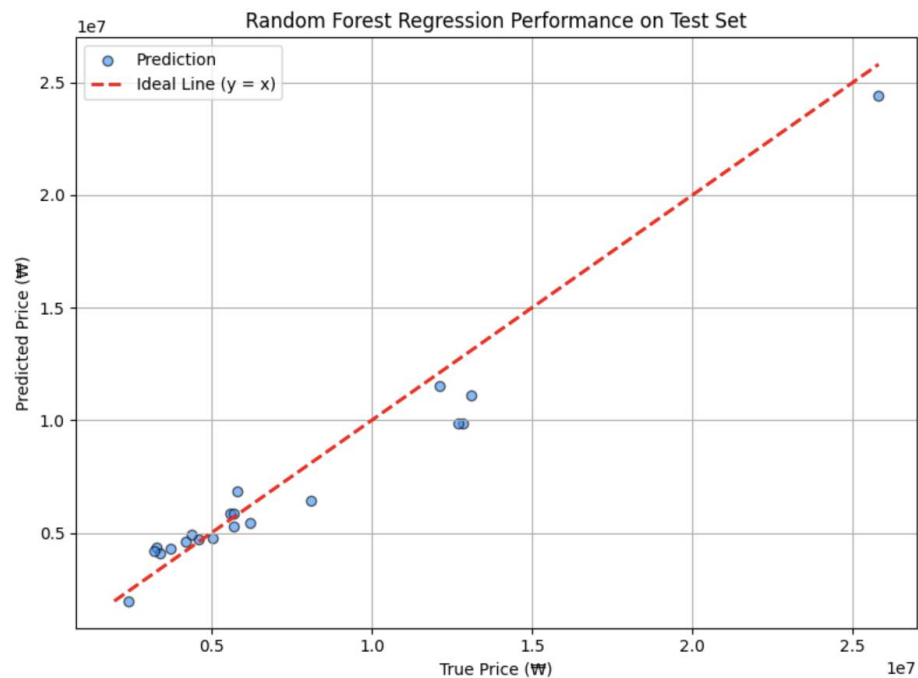


- 불필요한 계산을 줄이고 빠르게 더 정확한 예측이 가능함
- 중고차 가격처럼 복합적인 요소가 영향을 주는 예측 문제에 최적화된 모델
 - » **가격에 가장 큰 영향을 주는 특성에 집중하여** 정확한 시세 예측 가능
- 기존 Tree 모형 대비 속도와 성능 향상을 기대하며 추가 적용

Light GBM

Seed	Train MSE	Valid MSE	Valid MAPE
101	5.571358e+11	7.199594e+11	9.363708
202	5.477281e+11	7.338003e+11	9.182130
303	5.510093e+11	7.624074e+11	10.650787
404	5.480023e+11	7.449433e+11	10.918523
505	5.604986e+11	7.116264e+11	9.550685
평균	5.528748e+11	7.345473e+11	9.93%

- ▶ n_estimators : 150
- ▶ Learning_Rate : 0.1
- ▶ Max_depth: 10



- ▶ TEST MSE : $1.177773e+12$
- ▶ TEST MAE : 773,343.04 (77만 3천원)
- ▶ TEST MAPE : 11.71%

평가

- 모델 간 비교
 - Light GBM이 가장 우수

모델	MAE	MAPE
Random Forest	958,652.90 (95만 8천원)	14.17%
AdaBoost	811,087.80 (81만 1천원)	12.59%
Light GBM	773,343.04 (77만 3천원)	11.71%

5. 보완점

값 그대로 사용한 데이터

- NOTAVAILABLE, JOINCAR
 - train 데이터 내에서 모두 0으로 구성되어 있어, 모델이 이 변수에 의미 있는 판단 기준을 학습하지 못하는 구조적 한계
 - 향후 Test 데이터에 1이 존재할 경우, 무반응 혹은 잘못 학습할 가능성 존재
 - Train 데이터 내에 1이 존재하는 데이터를 소량이라도 포함시켜 모델이 해당 클래스 존재 가능성을 학습할 수 있도록 함

```

[NOTAVAILABLE] unique values:
NOTAVAILABLE
0    36793
Name: count, dtype: int64

[FLOODING] unique values:
FLOODING
0    34324
1     2469
Name: count, dtype: int64

[TOTAL_LOSS] unique values:
TOTAL_LOSS
0    36780
1      13
Name: count, dtype: int64

[JOINCAR] unique values:
JOINCAR
0    36793
Name: count, dtype: int64

```

Train Data

```

[NOTAVAILABLE] unique values:
NOTAVAILABLE
0     20
Name: count, dtype: int64

[FLOODING] unique values:
FLOODING
0     20
Name: count, dtype: int64

[TOTAL_LOSS] unique values:
TOTAL_LOSS
0     20
Name: count, dtype: int64

[JOINCAR] unique values:
JOINCAR
0     20
Name: count, dtype: int64

```

Test Data

OPTION_COUNT

- 가중치 고려
 - 현재 방식은 모든 옵션을 동등한 중요도로 판단
 - » 내비게이션, 스마트키, 에어백처럼 영향력이 높은 옵션 존재
 - » 가중합 또는 중요도별 점수화 고려 가능
- 중복 설치 불가능한 단일 옵션 고려
 - 선루프 옵션
 - » 파노라마 선루프, 일반 선루프, 듀얼 선루프는 상호 배타적인 선택지로, 차량 한 대에는 한 종류만 선택
 - » 셋 중 하나라도 존재하면 1로 고려

인용 및 출처

- “오토비즈 - 유통사업.” 현대글로비스, <https://www.glovis.net/kr/home/business/distribution/autobiz>.
- “중고차 시장 진입 2라운드...완성차 업계 움직인다.” 매일경제, 13 Apr. 2022, <https://www.mk.co.kr/economy/view.php?sc=50000001&year=2022&no=272120>.
- “중고차 시장 진출 허용한 정부, 불신 해소가 우선이다.” 더스쿠프, 12 May 2022, <https://www.thescoop.co.kr/news/articleView.html?idxno=38099>.
- “중고차 거래동향.” 현대자동차 인증중고차, <https://certified.hyundai.com/p/hilab/stat/getTradeTrend.do>.
- “중고차 진출 현대차, '신뢰·품질' 무기로 승부수.” 뉴스웨이, 16 Oct. 2023, <https://www.newsway.co.kr/news/view?ud=2023101617292550366>.
- “중고차 판매가이드.” 엔카, http://www.encar.com/sg/sg_sellguide.do.
- “엔카 보도자료 및 뉴스.” 엔카 뉴스, <https://fem.encar.com/company/encar-news>.
- “Hi-Lab 콘텐츠.” 현대자동차 인증중고차, <https://certified.hyundai.com/p/hilab/contents/getHiLabContentsMain.do>.
- 중고차 사고이력, 어떻게 확인할까? 밀알자동차, <https://milalcar.co.kr/blogPost/20>.



FINANCIAL INNOVATION
& ANALYTICS LAB.

Q&A