이화여자대학교 대학원

2018학년도

석사학위 청구논문

국산 중고 자동차 가격 예측 및 영향요인 분석

통 계 학 과

조 수 진

2019

국산 중고 자동차 가격 예측 및 영향요인 분석

이 논문을 석사학위 논문으로 제출함

2018 년 12 월

이화여자대학교 대학원

통 계 학 과 조 수 진

조 수 진 의 석사학위 논문을 인준함

지 도 교 수 송 종 우 _____

심 사 위 원 이 외 숙 _____

송 종 우 _____

김 미 정 _____

이화여자대학교 대학원

목 차

표 목 차

그 림 목 차

논 문 개 요

본 연구의 목적은 다양한 데이터마이닝 기법을 이용하여 온라인 시장에서의 국산 중고 자동차 가격을 분석하고, 예측 모형을 통해 중고차 가격 선정에 영향을 미치는 요인을 파악하는 데 있다. 그리고 중고차 가격 예측과 신차 대비 중고차 가격의 비율을 예측하고자 한다. 분석에 사용한 모형으로 단계적 선형 회귀 모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 익스트림 그래디언트 부스팅, 신경망 모형으로 6가지이며, 예측력 평가 지표인 RMSE에 의해 랜덤 포레스트가 최적 모형으로 선택되었다. 공통적으로 사용 기간이 가장 중요한 변수로 도출되었다. 그 외에 중고차 가격 모형에서는 신차 가격과 타이어 너비가 영향을 미치는 변수이며, 신차 대비 중고차 가격의 비율 모형에서는 주행거리가 영향을 미치는 변수로 나타났다.

Ⅰ. 서론

A. 연구의 목적 및 방법

한국자동차산업협회에 따르면 2017년 9월 중고차 등록 거래 수는 314,307건으로 전년 대비 9.2% 증가하였다. 계속해서 중고 자동차 시장이 성장하고 있으나, 정보의 비대칭성에 의한 소비자 피해는 여전히 발생하고 있다. 온라인 중고 자동차 시장에서는 비교적 자동차의 정보를 수집하기 쉽다. 하지만 소비자를 유인하기 위해 판매자가 허위 정보를 제공하는 피해 사례가 많았다. 그럼에도 중고차 시장이 성장할 수 있었던 배경에는 최근 자동차의 품질과 내구성이 좋아졌기 때문이다(한상식 외, 2017).

중고차 가격산정에 핵심요소인 연식, 주행거리, 사고 유무를 중심으로 중고차 거래의 특성을 분석한 윤대권 외(2015)의 선행 연구가 있다. 본 연구는 선행연구를 확장하여 추가적인 자동차의 성능과 특성을 고려해 중고 자동차 가격 예측 모형을 제시하고, 가격 예측에 영향을 미치는 요인을 도출하였다. 분석에 사용한 구체적인 기법들은 stepwise 선형 회귀 모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 익스트림 그래디언트 부스팅, 신경망 모형이며 RMSE(평균 제곱근 오차)를 이용하여 예측력을 평가하였다. 그리고 신차를 구입한 후 3년과 5년이 되었을 때의 신차 대비 중고차 가격 비율을 예측하여 잔존 가치가 높은 차량을 살펴보았다.

본 논문은 다음과 같은 구성으로 있다. 제2장은 분석 자료 수집 과정과 분석에 사용된 주요 변수들에 대해 설명한다. 제3장은 분석 결과를 통한 모형 비교와 최종 모형을 제시한 후 제4장에서 본 연구를 요약하여 결론을 맺는다.

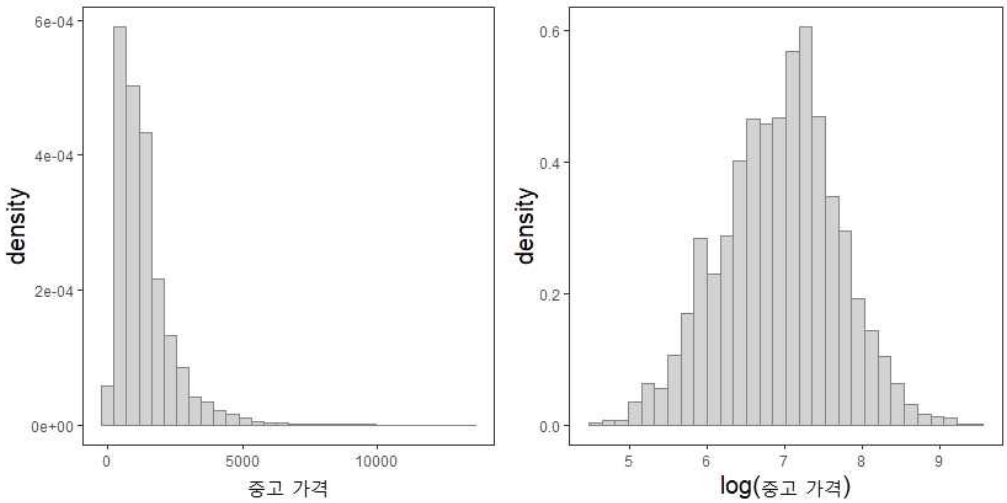
Ⅱ. 분석 자료 설명

A. 자료 수집 과정

본 연구의 연구 대상은 2014년 4월 8일부터 2018년 9월 20일까지 등록된 국산 중고차다. 수입차의 경우 가격 정보 비공개로 인해 자료의 편의가 많아 국산차로 제한하여 분석하였다. 또한 국산 자동차 중 상용차인 경우, 판매 가격 정보가 없는 경우와 자료의 손실이 많은 경우는 분석 대상에서 제외하였다. 중고 가격이 100만원 미만이거나 신차대비 중고차 가격의 비율이 1보다 큰 자료는 잘못된 정보일 가능성이 높아 제외하였다. 그리고 중고 자동차 가격에 영향을 미칠 것으로 기대되는 요인으로 크게 기본 정보(연식, 주행거리, 배기량, 색상 등), 가격 정보(신차 가격, 중고 가격), 차량제원(연비, 최고 출력, 최대 토크 등), 옵션 정보(썬루프, 스마트키, 후방 감지 센서 유무 등), 보험처리 이력 정보, 성능점검 정보(판금 및 교환 횟수, 압류 및 저당 여부 등)를 고려하여 중고 자동차 온라인 중개 사이트인 보배드림(http://www.bobaedream.co.kr/)에서 크롤링하여 수집하였다.

B. 변수 설명

본 연구의 목적은 국산 자동차의 중고 가격에 영향을 미치는 변수들을 이용하여 국산차의 중고 가격 예측과 신차대비 중고차 가격의 비율을 예측하는 것이다. 총 20,422건의 수집된 자료 중에서 중고 가격의 최대값은 1억 3,500만원으로 제네시스 EQ900 5.0 GDi AWD 프레스티지 리무진이다. 신차대비 중고차 가격 비율의 최소값은 0.019로 현대 뉴 제네시스 G330 프리미엄이고, 최대값을 가지는 자동차는 현대 그랜져IG 2.2 e-VGT 프리미엄 스페셜, 르노삼성 SM6 1.6 TCe RE, 쌍용 티볼리 1.6 디젤 2WD LX로 신차 가격과 동일하였다. 중고 가격의 히스토그램을 살펴보면 오른쪽으로 긴 꼬리를 지니는 비대칭 분포로 예측력을 높이기 위하여 로그 변환을 통해 대칭분포로 만들었다.



[그림 1] 중고 가격(왼쪽), 로그 중고 가격(오른쪽)의 히스토그램

B.1. 제조사

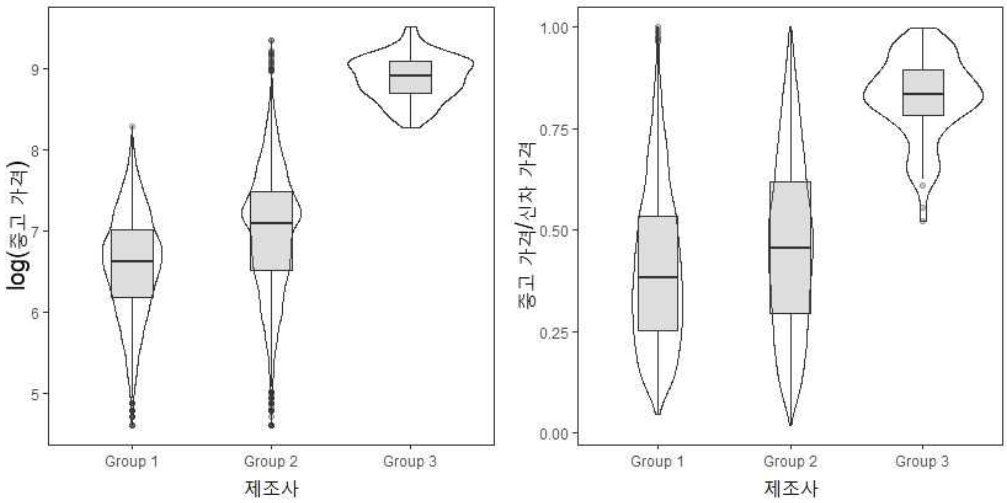
자동차 구매에 있어 자동차 브랜드는 고려 요인이다. 본 연구의 자료에 현대, 쌍용, 기아, 르노삼성 등 8개의 국내 제조사가 있으며 decision tree 모형에 의해 3개의 그룹으로 나누었다. 각 그룹에 해당하는 제조사와 개수는 <표 1>에 있다.

그룹	제조사	N
그룹 1	GM대우, 르노삼성, 쉐보레	3,302
그룹 2	쌍용, 현대, 기아	17,011
그룹 3	제네시스, 어울림모터스	109

<표 1> 그룹별 제조사명

[그림 2]를 살펴보면, 제조사에 따라 중고 가격의 차이가 있음을 알 수 있다. 그룹 3에 해당하는 제네시스와 어울림모터스의 자동차는 가격이 높은 편이며, 그룹 1에 해당하는 GM대우, 르노삼성, 쉐보레의 자동차는 상대적으로 가격이 낮게 나타났다. 신차 가격과 중고 가격을 비교하였을 때 그룹 3 제조사의 자동차는 신차 가격과 차이가 가장 없었으며, 그룹 1 제조사의 자동차는 중고 가격이 신차 가격

과 차이가 컸다.



[그림 2] 제조사에 따른 상자 그림

B.2. 색상

신차 구입 시 색상에 따라 가격이 다르다. 그러므로 중고차 가격 예측 시 색상을 고려해야 한다. 개인마다 선호하는 색상이 다르며, 색상의 인기도에 따라 가격의 차이가 존재할 것이다. 26가지 종류의 색상 중 주로 선호한느 색상인 검정색, 회색, 흰색과 기타로 나누었다. 색상에 따른 자동차 개수는 <표 2>에 있다.

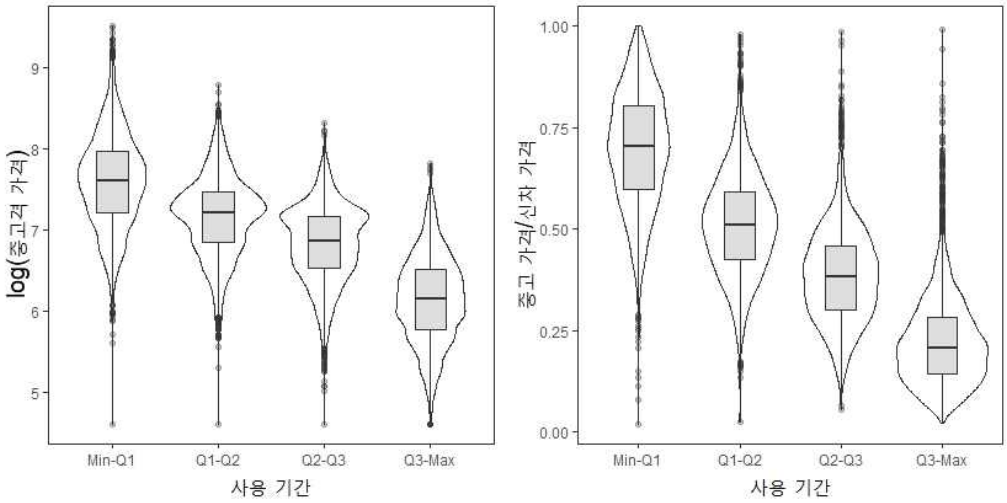
검정색	회색	흰색	기타
6,800	2,069	3,986	7,567

<표 2> 색상에 따른 자동차 개수

B.3. 사용 기간

자동차에는 수명이 존재한다. 운전자의 운전 습관에 따라 자동차의 수명이 짧아

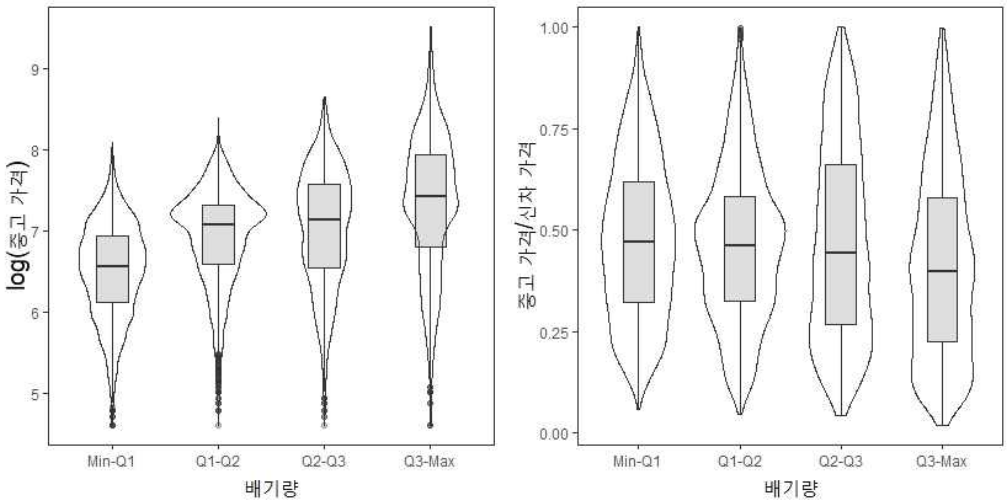
지거나 길어진다. 관리를 잘 받은 자동차여도 오래된 자동차는 고장 날 확률이 높다. 따라서 중고 시장에서 사용 시간은 중요한 변수이다. 본 연구에서 사용 기간은 차량의 연식일부터 최초 판매 등록일까지의 개월 수로 정의하였다. 사용 기간에 따른 상자 그림은 [그림 3]에 있다. 자동차가 오래될수록 중고 가격이 낮으며, 시간이 지날수록 가격은 더 하락하는 것으로 보인다.



[그림 3] 사용 기간에 따른 상자 그림

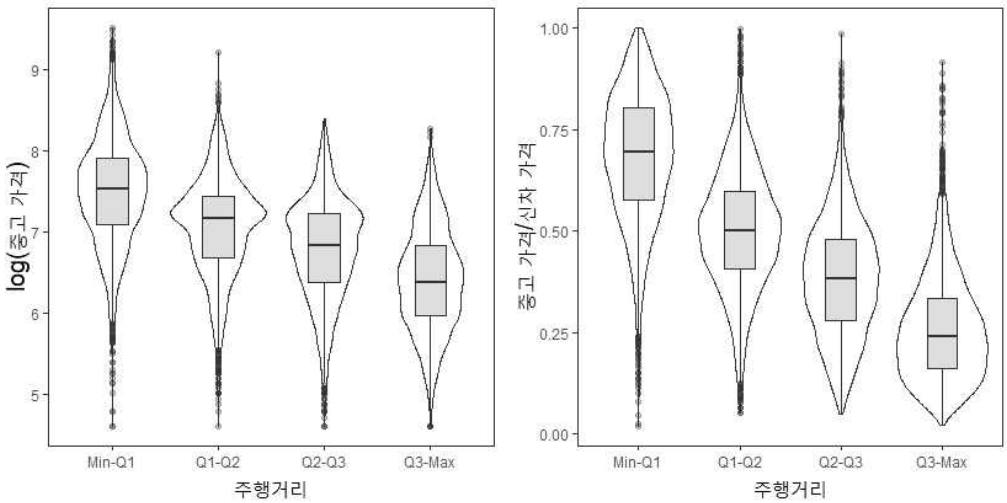
B.4. 배기량

배기량은 자동차 엔진 내부에서 피스톤이 한 번 움직일 때 발생하는 가스의 부피를 뜻한다. 배기량이 클수록 공기와 연료의 혼합을 많이 흡입할 수 있어 출력이 높아진다. 따라서 힘이 좋은 자동차를 찾는다면, 배기량이 높은 자동차를 선택해야 한다. [그림 4]를 살펴보면, 배기량이 클수록 중고 가격은 높지만, 신차 대비 중고차 가격의 비율을 보면, 배기량에 따른 차이가 존재하지 않는다.



[그림 4] 배기량에 따른 상자 그림

B.5. 주행거리



[그림 5] 주행거리에 따른 상자 그림

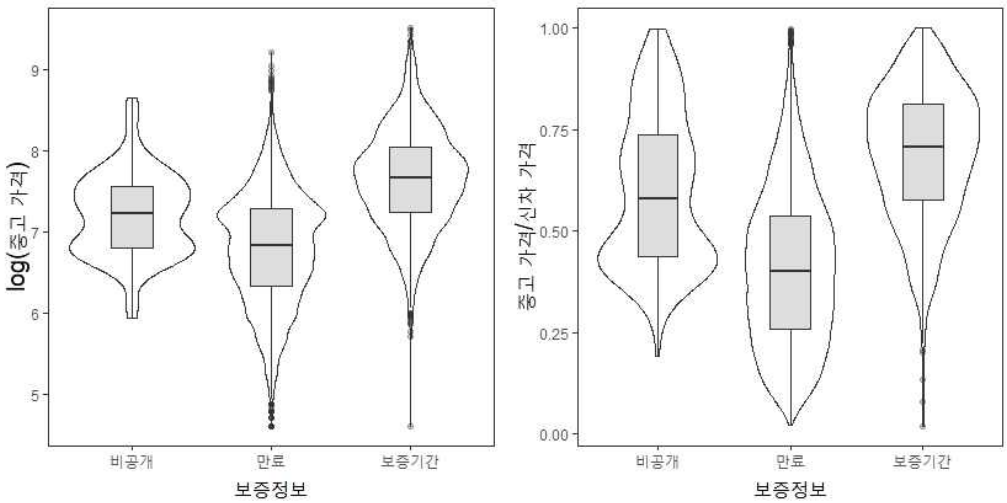
주행거리는 운전자가 자동차를 타고 이동한 총거리를 말한다. 주행거리가 적다는 것은 상대적으로 신차와 가깝다고 볼 수 있기 때문에 중고차 가격 예측에 중요한

요인으로 고려된다. 주행거리에 따른 상자 그림인 [그림 5]를 보면, 주행거리가 길수록 중고차 가격이 낮게 형성되며, 감가가 커진다.

B.6. 변속기

변속기는 속도에 따라 엔진 회전수를 변화시키며 전달해주는 장치로 수동 변속기와 자동 변속기로 나뉜다. 수동 변속기는 연비 향상에 유리한 반면, 자동 변속기는 운전 편의성이 뛰어난다. 따라서 일반적으로 자동 변속기를 선호한다. 본 연구에서도 자동 변속기를 가진 자동차는 18,704건, 수동 변속기를 가진 자동차는 1,718건으로 자동 변속기를 더 선호하는 것을 알 수 있으며, 자동 변속기를 가진 자동차의 중고 가격이 더 높아 예측 변수로 고려하였다.

B.7. 보증정보



[그림 6] 보증정보에 따른 상자그림

보증정보는 제조사 보증 중 엔진 및 동력 부품 기준으로 보증기간, 만료, 비공개로 나누었다. 보증기간은 신차 구입일 기준으로 계산되며 보증기간 또는 주행거리 중 먼저 도달한 것은 만료로 간주되었다. 보증기간 내의 자동차는 고장이 날 확률

이 적다는 것을 뜻하기 때문에 보증기간이 만료되었는지에 대한 정보는 중요하다. 보증정보 비공개는 결측치로 파악할 수 있지만, 판매자가 정보 공개를 원치 않는 것으로 볼 수 있다. 따라서 비공개 또한 보증정보의 하나의 요인으로 고려하였다. [그림 6]을 보면, 보증정보를 공개하지 않은 자동차가 보증 기간이 만료된 자동차보다 중고 가격이 높게 나타났다.

B.8. 연료

가솔린, 디젤, LPG, CNG 등 연료의 종류가 다양하다. 대중적으로 사용되는 가솔린, 디젤, LPG 연료로 제한하였다. <표 3>을 보면, 가솔린을 사용하는 자동차가 13,230건으로 가장 많았다. 가솔린은 디젤보다 연료의 가격이 비싸지만 비교적 진동이 없어 대부분 가솔린 연료를 사용한다. 반면 디젤의 경우, 연료의 가격은 저렴하지만 소음과 진동이 크고 매연을 많이 배출한다. LPG는 가격이 저렴하고 유해물질을 적게 배출한다는 장점이 있지만, 연료 효율이 좋지 않으며 영업용 자동차에만 설치가 가능하고 장애인, 국가유공자만 구입이 가능하다. 연료에 대한 상자 그림을 그리면, LPG, 가솔린, 디젤 순으로 중고차 가격이 높았다.

가솔린	디젤	LPG
13,230	5,490	1,702

<표 3> 연료에 따른 자동차 개수

B.9. 보험처리 이력 정보 및 성능점검 정보

중고 자동차를 구매하기 전 보험처리 이력과 성능점검 확인은 필수이다. 사고이력이 있는 자동차는 자동차 성능에 영향을 미칠 것이다. 그러므로 보험처리 횟수, 판금 및 교환 횟수, 사고 및 침수 여부를 예측의 변수로 고려하였다. 압류 및 저당 여부, 불법 구조 변경 여부 또한 중고 자동차 가격에 영향을 미치는 요인일 것이다. 그리고 보증정보 변수에서와 동일하게 보험처리 이력과 성능점검 정보를 공개하지 않은 경우 하나의 범주로 고려하였다. 따라서 보험처리 횟수와 판금 및 교

환 횟수는 0회, 1-5회, 6-10회, 11회 이상, 비공개로 범주화하였다. 보험처리 횟수와 판금 및 교환 횟수에 따른 상자 그림을 봤을 때 보험처리 횟수와 판금 및 교환 횟수가 11회 이상인 경우, 정보를 공개하지 않았을 때보다 중고차 가격이 낮았다.

B.10. 차량제원

연비, 최고 출력, 최대 토크, 차량 중량, 구동 방식, 타이어 너비, 타이어 편평비 그리고 타이어 휠 크기를 반영하여 자동차의 성능과 특성을 반영하였다. 구동 방식에는 엔진 위치와 구동축에 따라 사륜구동 방식과 이륜구동 방식으로 나뉜다. 사륜구동 방식인 4WD, AWD와 이륜구동 방식인 전륜, 후륜으로 나누어 고려하였다. 사륜구동보다는 이륜구동 방식이 많았으며, 그중에서도 전륜 방식이 가장 많았다.

4WD	AWD	전륜	후륜
1,192	398	15,005	3,827

<표 4> 구동 방식에 따른 자동차 개수

B.11. 옵션 정보

자동차에 옵션을 추가했는지에 따라 가격이 다르게 형성된다. 외관, 내장, 안전, 편의, 멀티미디어 기준으로 대표적인 옵션을 이용하였다. 분석에 사용된 옵션 정보는 썬루프, HID/제논 램프 혹은 LED 램프, 열선시트, 통풍시트, 측면 에어백, 후방 감지 센서, 후방 카메라, 스마트키, 네비게이션 옵션이다.

B.12. 장애인용 차량

장애인 전용 자동차는 일반 자동차보다 낮은 가격대에 형성되어 있으며 복지카드를 가고 있는 사람만 구입이 가능하다. 하지만 5년이 경과하면 비장애인도 구

입이 가능하여, 고려 요인으로 선택하였다.

B.12. 신차 가격

중고 자동차 가격 예측 시 신차 가격에 영향을 받는다. 신차 가격에 대한 중고 가격의 산점도를 그리면, 신차 가격이 높을수록 중고 가격도 높은 모습을 보여준다. 따라서 정확한 가격 예측을 위해 중고차 가격 예측 모형에서만 신차 가격을 예측 변수로 고려하였다. 반응 변수인 중고 가격을 로그 변환하였으므로 신차 가격 또한 로그 변환하여 반응 변수와의 단위를 맞추었다.

<표 5>는 분석에 사용된 변수를 정리하였다.

변수 이름	변수 설명	자료형
설명변수		
new_price	신차 가격 (원) *	수치형
age	사용 기간 (월) **	
disp	배기량 (cc)	
distance	주행거리 (km) **	
mileage	연비(km/l)	
power	최고 출력 (마력)	
max_torque	최대 토크 (kg.m)	
weight	차량 중량 (kg)	
tire_width	타이어 너비 (mm)	
tire_flat	타이어 편평비 (%)	
tire_wheel	타이어 휠 크기 (inch)	
brand	제조사	범주형
color	색상	
auto	변속기	
warranty	보증정보	
fuel	연료	
insurance	보험처리 회수	
change	판금 및 교환 회수	
dist_secu	압류 및 저당 여부	
acci_flood	사고 및 침수 여부	

illegal_struc	불법 구조 여부	범주형
drives_sys	구동 방식	
sunroof	썬루프 유무	
smartkey	스마트키 유무	
HID_LED_lamp	HID, 제논, LED 램프 유무	
heat_sheet	열선시트 유무	
wind_sheet	통풍시트 유무	
back_censor	후방 감지 센서 유무	
back_camera	후방 카메라 유무	
side_air	측면 에어백 유무	
navi	네비게이션 유무	
handicap	장애인용 차량 여부	
반응변수		
used_price	중고 가격 (원)	수치형
prop_used	신차 가격 대비 중고 가격 비율	

* 중고차 가격 예측 모형만 이용

** 중고차 가격 예측 모형과 신차 대비 중고차 가격 비율 모형만 이용

〈표 5〉 변수 설명

Ⅲ. 분석 결과

이 장에서는 중고 자동차 가격 예측 모형(Model 1)과 신차 대비 중고차 가격 비율 예측 모형(Model 2)의 최적 모형을 찾고자 한다. 2014년 4월부터 2017년 9월까지 약 84% 자료를 train 데이터로 사용하였고, 2017년 10월부터 2018년 9월까지 나머지 자료를 test 데이터로 사용하였다. 최적 모형을 찾기 위하여 train 데이터를 이용해 교차 검증을 실시하였다. 그 후, test 데이터를 통해 최적 모형의 예측력을 평가하고, 예측에 영향을 미치는 요인을 알아보았다. 마지막으로 사용 기간이 3년과 5년 된 자동차를 이용하여 각 연식에서의 가격 특징을 살펴보았다.

A. Train 데이터 내 교차 검증 비교

2014년 4월부터 2017년 9월까지의 train 데이터를 이용하여 교차 검증을 100번 반복하였다. 매 교차 검증마다 자료의 70%는 train 데이터, 나머지 30%는 test 데이터로 나누었다. Train 데이터로 학습시킨 모형을 이용하여 test 데이터를 통해 RMSE를 계산한 후, 모형을 비교하였다. 분석에 사용한 통계 모형으로 stepwise 회귀 모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 익스트림 그래디언트 부스팅, 신경망 모형이다.

A.1. 중고 자동차 가격 예측 모형 (Model 1)

중고 자동차 가격 예측 모형(Model 1)에서의 교차 검증 결과는 <표 6>에 있다. 6개의 후보 모형 중에서 랜덤 포레스트 모형의 test RMSE가 0.1904로 가장 작았다. 이는 랜덤 포레스트 모형을 사용하는 것이 예측력이 가장 좋은 것을 말한다. 따라서 랜덤 포레스트 모형을 중고 자동차 가격 예측 모형(Model 1)의 최적 모형으로 선택하였다.

	train RMSE	test RMSE
Stepwise Linear	0.2463	0.2471
Random Forest	0.0902	0.1904
Gradient Boosting	0.3061	0.3071
Support Vector Machine	0.1944	0.2024
Extreme Gradient Boosting	0.1999	0.2132
Neural Network	0.2104	0.2200

<표 6> 교차 검증 에러 결과 (Model 1)

A.2. 신차 대비 중고차 가격 비율 예측 모형 (Model 2)

신차 대비 중고차 가격 비율 예측 모형(Model 2)의 교차 검증 결과는 <표 7>과 같다. 랜덤 포레스트 모형이 6개의 후보 모형 중에서 0.0738로 test RMSE 값이 가장 작게 나타났다. 따라서 신차 대비 중고차 가격 비율 예측 모형(Model 2)의 최적 모형은 랜덤 포레스트 모형으로 결정되었다.

	train RMSE	test RMSE
Stepwise Linear	0.0987	0.0989
Random Forest	0.0352	0.0738
Gradient Boosting	0.1033	0.1037
Support Vector Machine	0.0767	0.0792
Extreme Gradient Boosting	0.0773	0.0812
Neural Network	0.0888	0.0900

<표 7> 교차 검증 에러 결과 (Model 2)

B. 최종 모형 결과 및 해석

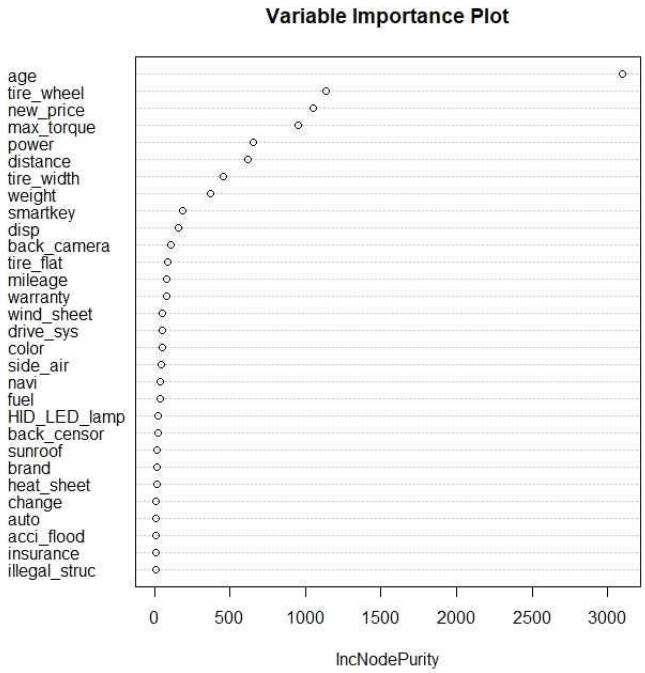
앞 절에 의해 결정된 최종 모형은 모두 랜덤 포레스트 모형이다. 최종 모형을 바탕으로 2017년 10월부터 2018년 9월까지 test 데이터의 중고 자동차 가격과 신차 대비 중고차 가격 비율을 예측하였다. 그리고 각 모형에서 예측에 중요한 변수를 파악하였다.

B.1. 중고 자동차 가격 예측 모형 (Model 1)

최종 모형에 대한 중고 자동차 가격 예측 모형(Model 1)의 RMSE 결과는 <표 8>에 있다. 랜덤 포레스트 모형으로 예측 시 test RMSE 값은 0.2148이다. 이를 지수 변환하였을 때 중고 자동차 가격의 오차는 약 165만원이다. 이는 평균 모형으로 예측했을 때 중고 자동차 가격의 오차가 약 790만원이므로 약 4.8배의 정확성이 향상되었다.

	train RMSE	test RMSE
Random Forest	0.0894	0.2148

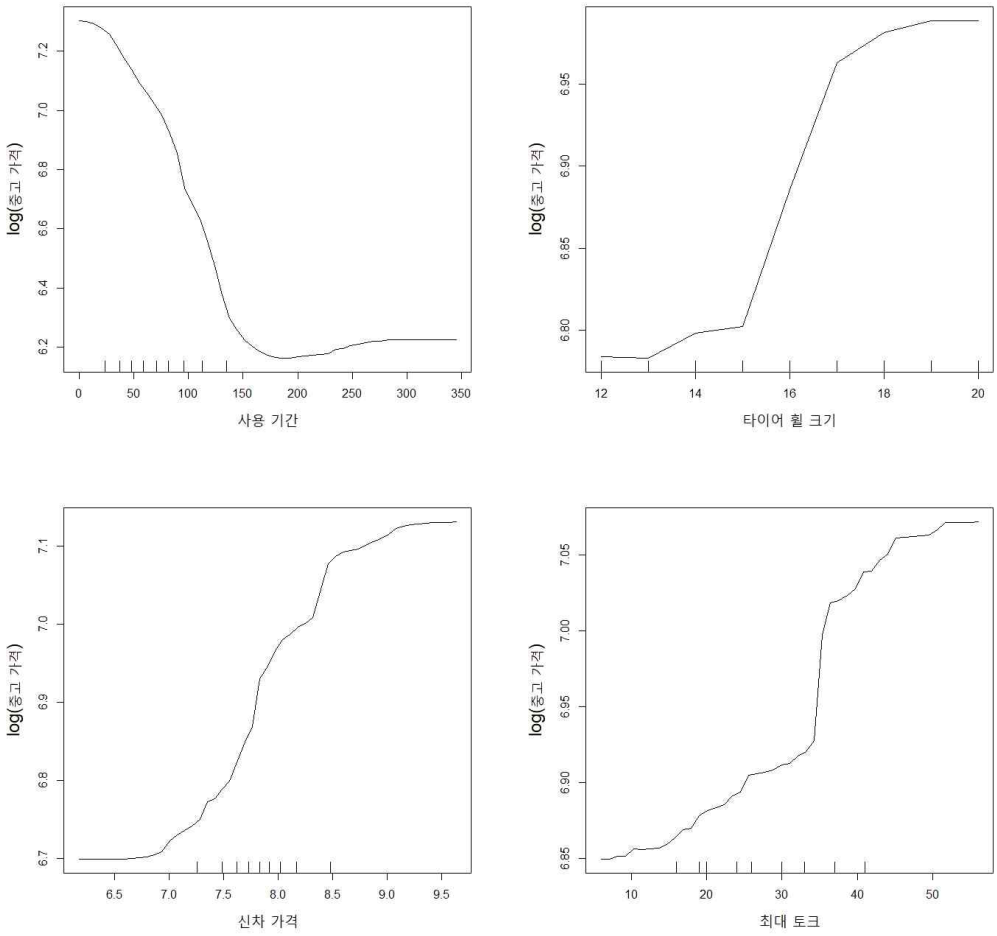
<표 8> 최종 모형의 RMSE (Model 1)



[그림 7] 변수 중요도 그림 (Model 1)

중고 자동차 가격 예측 모형(Model 1)에 대한 변수 중요도 그림은 [그림 7]과 같다. 사용 기간(age)이 중고차 가격 예측에 가장 영향을 미치는 요인으로 나타났다.

타이어 휠 크기(tire_wheel), 신차 가격(new_price), 최대 토크(max_torque), 최고 출력(power), 주행거리(distance), 타이어 너비(tire_width), 차량 중량(weight) 순으로 중요하였다. 상대적으로 중요한 변수 4개의 부분 의존도 그림은 [그림 8]에 있다.



[그림 8] 부분 의존도 그림 (Model 1)

오래 사용(age)할수록 중고 가격은 감소하지만, 사용 기간이 약 180개월(15년) 이상인 자동차는 희소가치로 인하여 중고 가격이 소폭 상승한다. 타이어 휠 크기가 클수록, 신차 가격(new_price)이 높을수록, 최대 토크가 클수록 중고 가격은

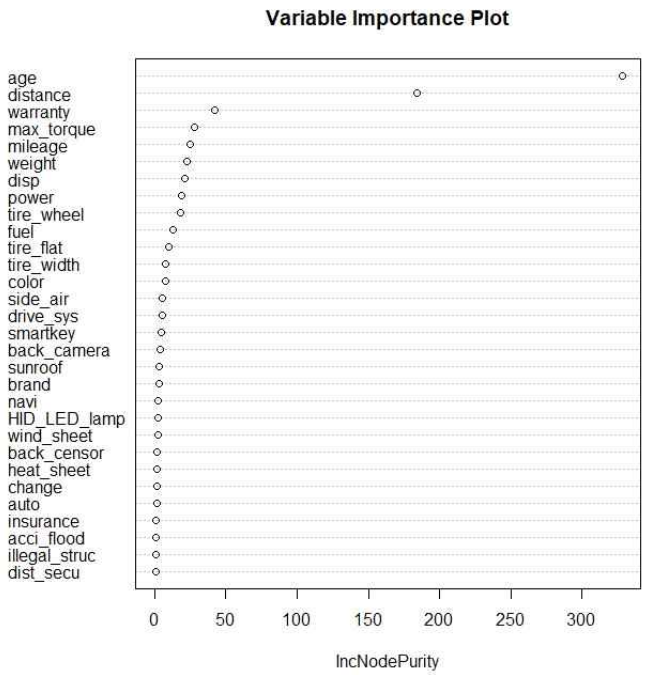
높아진다. 특히, 타이어의 휠 크기가 15inch 이상인 자동차와 최대 토크가 약 34kg.m 이상이면 중고 가격은 급격히 높아진다.

B.2. 신차 대비 중고차 가격 비율 예측 모형 (Model 2)

신차 대비 중고차 가격 비율 예측 모형(Model 2)의 예측력 평가 결과는 <표 9>과 같다. 최종 모형인 랜덤 포레스트 모형으로 적합시켰을 때 신차 가격에 대한 중고차 가격의 비율 오차는 0.0790이다. 신차 대비 중고차 가격 비율의 정보만 가지고 평균 모형으로 예측한 오차는 0.1760으로 약 2.2배 개선되었다.

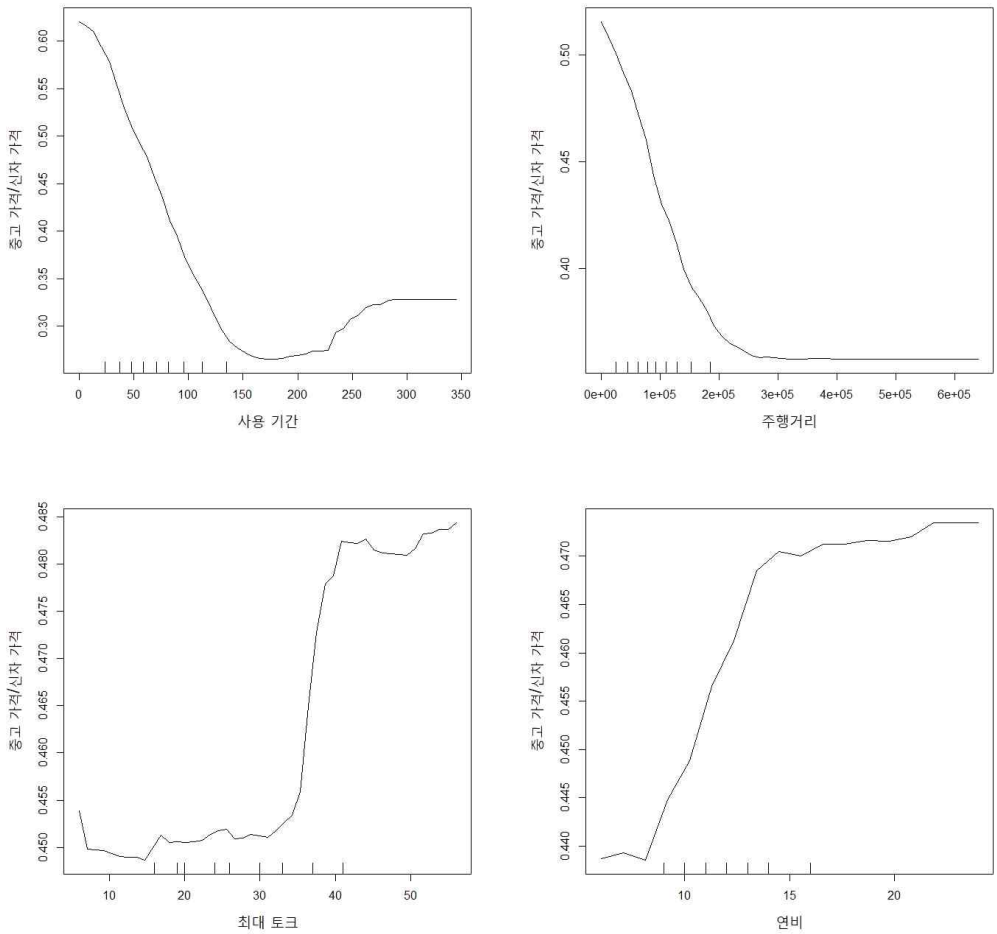
	train RMSE	test RMSE
Random Forest	0.0348	0.0790

<표 9> 최종 모형의 RMSE (Model 2)



[그림 9] 변수 중요도 그림 (Model 2)

[그림 9]는 신차 대비 중고차 가격 비율 예측 모형(Model 2)의 변수 중요도 그림이다. 사용 기간(age)과 주행거리(distance)는 가격 비율 예측에 많은 영향을 받는다. 그 외에 보증정보(warranty), 최대 토크(max_torque), 연비(mileage) 순으로 중요함을 알 수 있다. [그림 10]은 중요한 수치형 변수 상위 4개에 대한 부분 의존도 그림이다.



[그림 10] 부분 의존도 그림 (Model 2)

부분 의존도 그림에 따르면 자동차를 사용(age)할수록 가격 비율은 감소하여 중고 자동차의 잔존 가치는 하락하지만, 약 180개월(15년)이 지난 자동차는 희소성

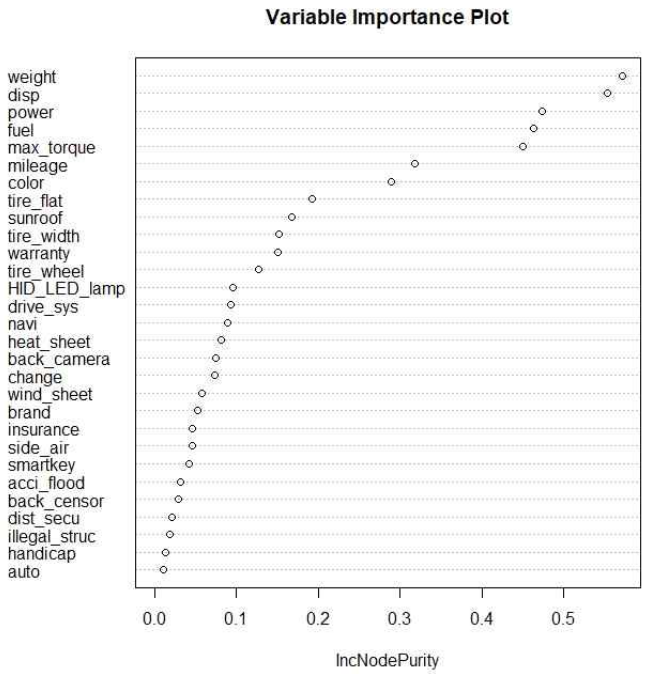
에 의해 가격 비율이 증가하여 감가가 줄어든다. 주행거리(distance)는 총 이동 거리가 길수록 신차에 대한 중고차 가격 비율이 감소하는 모습이고, 주행거리가 약 270,000km 이상이면 주행거리에 따른 가격 비율의 영향이 없다. 최대 토크(max_torque)의 경우, 최대 토크가 약 30kg.m 이하인 자동차는 가격 비율에 영향을 거의 주지 않는다. 하지만 최대 토크가 30kg.m 이상인 경우, 신차 가격에 대한 중고 가격 비율이 급격히 증가하고 40kg.m 이상인 자동차는 가격 비율 예측에 영향을 받지 않는다. 연비(mileage)는 연비가 클수록 가격 비율이 커지나, 약 15km/l 이상이면 비율에 영향을 받지 않는다.

C. 사용 기간에 따른 미래의 가격 비율 예측

앞의 A와 B절에 따르면 중고차 가격과 신차 대비 중고차 가격의 비율 예측 시 사용 기간이 가장 중요하였으며, 오래된 자동차일수록 가격이 하락하여 잔존 가치가 떨어지고 있는 모습을 볼 수 있었다. 그리고 자동차의 사용 기간이 3년, 5년이 되었을 때의 자동차 가격을 살펴보았다. 이때, 미래의 주행거리는 알 수 없어 분석 변수에서 제외하였다. 신차 구입 이후, 각 시점에서 최적 예측 모형은 랜덤 포레스트 모형이었다. 이를 바탕으로 최근 1년 사이에 등록된 자동차 중에서 사용 기간이 1년 미만인 자동차가 3년과 5년이 되었을 때 신차 대비 중고차 가격의 비율을 예측하였다.

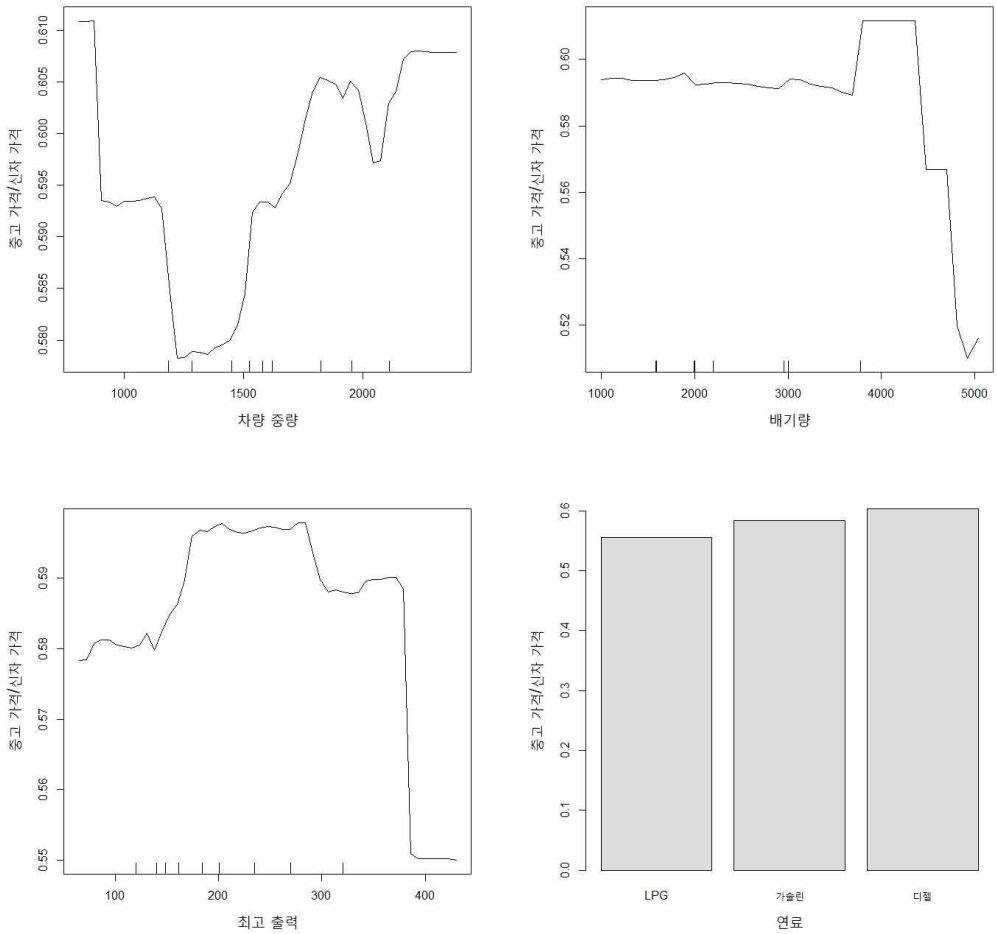
C.1. 3년 된 자동차의 가격 비율 예측 모형 (Model 3)

3년 된 자동차의 신차 대비 중고차 가격의 비율 예측 모형(Model 3)에 대한 변수 중요도 그림은 [그림 11]과 같다. 차량 중량(weight), 배기량(disp), 최고 출력(power), 연료(fuel), 최대 토크(max_torque) 순으로 중요하였다. 이는 앞 절에 있는 [그림 9]와 차이가 있었다. 특히, 사용 기간(age)을 하나의 요인으로 고려하였을 때 연료는 상대적으로 덜 중요하였으나, 3년 된 자동차 내에서 가격 비율을 예측하였을 때 중요한 요인에 속하였다.



[그림 11] 변수 중요도 그림 (Model 3)

상위 4개의 변수에 대한 부분 의존도 그림은 [그림 12]에 있다. 1.2톤 이하의 자동차는 차량이 무거울수록 신차 대비 중고차의 가격 비율이 낮아졌으나, 1.2톤 이상의 자동차는 무거울수록 가격 비율이 높아졌다. 연식이 3년 된 차량은 배기량이 약 3,800cc 이하인 자동차는 신차 대비 중고차의 가격 비율에 영향을 주지 않으나, 배기량이 약 4,000cc 이상인 자동차는 배기량이 높을수록 가격 비율이 낮아진다. 최고 출력이 약 270마력 이하인 자동차는 최고 출력이 클수록 신차 대비 중고차의 가격 비율은 상승하는 모습이다. 하지만 최고 출력이 약 270마력 이상인 자동차는 최고 출력이 높을수록 가격의 비율이 하락하고 있으며, 특히 약 350마력 이상인 경우 급격히 감소한다. 연료의 경우, LPG을 이용하는 자동차가 상대적으로 가격 비율이 낮았으며, 가솔린, 디젤 순으로 신차 대비 중고차 가격의 비율이 높았다.



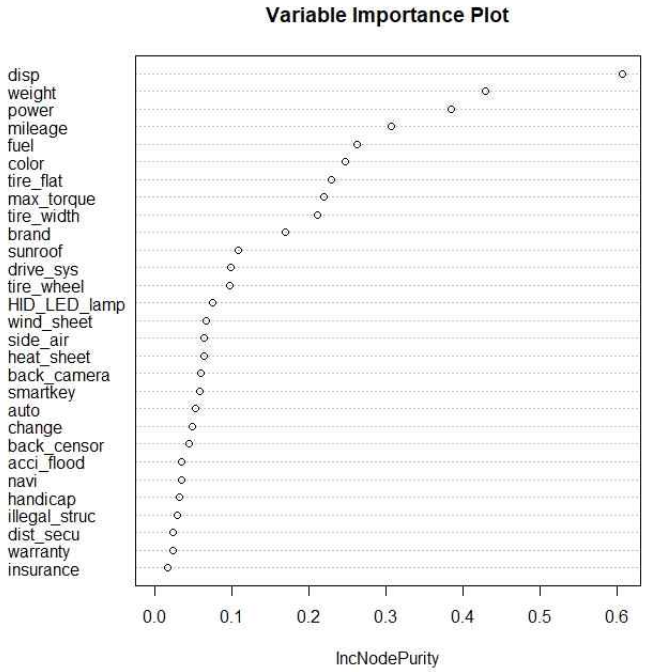
[그림 12] 부분 의존도 그림 (Model 3)

나이가 1년 미만인 자동차를 이용하여 신차가 출고된지 3년이 되었을 때 신차 대비 중고차 가격의 비율을 예측하였다. <표 10>은 69개의 모델 중에서 연식이 3년 되었을 때 신차 대비 중고차 가격의 비율이 높은 상위 5개의 자동차를 나타낸 결과이다. 비율이 가장 높았던 자동차의 모델은 기아 올 뉴 카니발 2.2 디젤 9인승 하이리무진 노블레스로 예상된다. 대체적으로 기아 차종이 3년이 지나도 잔존 가치가 높게 평가되었다.

모델명	중고 가격/신차 가격
기아 올 뉴 카니발 2.2 디젤 9인승 하이리무진 노블레스	0.7301
현대 싼타페 더 프라임 2.2 디젤 2WD 1 밀리언 얼티밋	0.7233
기아 더 뉴 카니발 2.2 디젤 9인승 노블레스 스페셜	0.7186
기아 더 뉴 카니발 2.2 디젤 9인승 프레스티지	0.7074
기아 더 뉴 쏘렌토 2.0 디젤 2WD 프레스티지	0.6939

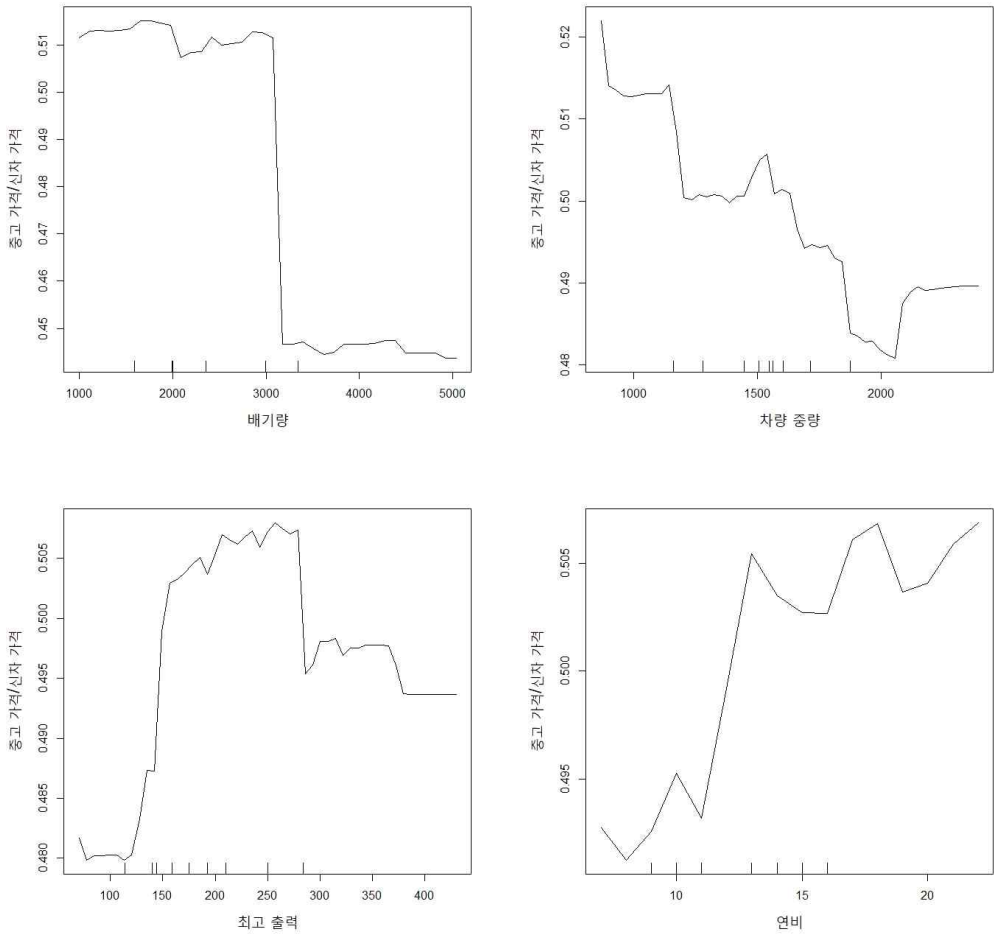
<표 10> 3년차 되었을 때 신차 대비 중고차 가격 비율 예측 상위 5개

C.2. 5년 된 자동차의 가격 비율 예측 모형 (Model 4)



[그림 13] 변수 중요도 그림 (Model 4)

연식이 5년 된 자동차에서 변수 중요도 그림은 [그림 13]과 같다. 신차 대비 중고차 가격 비율 예측 시 배기량(disp), 차량 중량(weight), 최고 출력(power), 연비(mileage), 연료(fuel) 순으로 중요하였다. 마찬가지로 사용 기간을 구분 없이 분석하였을 때와 차이가 있었으며, 모형 2에서 덜 중요했던 연료(fuel) 종류는 5년 된 자동차의 가격 비율 예측 시에도 중요 요인으로 나타났다.



[그림 14] 부분 의존도 그림 (Model 4)

상위 4개의 중요한 변수에 대한 부분 의존도 그림은 [그림 14]와 같다. 배기량의 경우, 3,000cc 이상과 이하로 구분할 수 있다. 3,000cc 이하와 3,000cc 이상의 자동차 내에서는 각각 신차 대비 중고차 가격의 비율에 영향을 미치지 않으나, 3,000cc 이상의 자동차의 가격 비율이 더 낮다. 차량이 무거울수록 가격 비율이 하락하며, 연비의 경우, 연비가 좋을수록 비율이 높아진다. 270마력 이하인 자동차는 최고 출력이 높을수록 신차 대비 중고차 가격 비율이 증가하나, 270마력 이상인 자동차는 가격 비율이 하락하였다. 이는 3년 된 자동차의 가격 비율 예측

모형 (Model 3)과 다른 양상을 보이고 있다.

모델명	중고 가격/신차 가격
기아 더 뉴 모하비 3.0 디젤 상시4WD 프레지던트	0.5798
현대 싼타페 더 프라임 2.2 디젤 2WD 1 밀리언 얼티밋	0.5543
기아 더 뉴 카니발 2.2 디젤 9인승 노블레스 스페셜	0.5538
기아 더 뉴 카니발 2.2 디젤 9인승 프레스티지	0.5536
현대 그랜져IG 3.0 GDI 익스클루시브 스페셜	0.5470

<표 11> 5년차 되었을 때 신차 대비 중고차 가격 비율 예측 상위 5개

<표 11>은 나이가 1년 미만인 신차가 5년이 되었을 때 신차 대비 중고차 가격의 비율을 예측하여 비율이 높은 상위 5개의 자동차이다. 비율이 가장 높았던 자동차의 모델은 기아 더 뉴 모하비 3.0 디젤 상시4WD 프레지던트로 예상된다. 또한 연식이 3년이 되었을 때 신차 대비 중고차 가격 비율이 높았던 차종인 현대 싼타페 더 프라임 2.2 디젤 2WD 1 밀리언 얼티밋, 기아 더 뉴 카니발 2.2 디젤 9인승 노블레스 스페셜과 프레스티지는 5년차가 되었을 때도 높게 추정되었다.

Ⅳ. 결론

본 연구는 자동차의 성능과 특성 등의 정보를 이용하여 국산 중고 자동차 가격 예측 모형(Model 1)과 신차 대비 중고차 가격의 비율 예측 모형(Model 2)을 연구하였다. 다양한 데이터마이닝 기법 중에서 교차 검증을 통해 최적 예측 모형을 제시하였고, 가격 예측에 영향을 미치는 중요 요인을 도출하였다.

분석 결과, 두 모형 모두 랜덤 포레스트 모형이 나머지 통계 모형보다 성능이 가장 우수하였다. 최종 모형을 바탕으로 중요한 변수를 살펴본 결과, 자동차의 사용기간, 주행거리와 최대 토크가 공통적으로 중요하였다. 자동차는 오래 사용할수록 가격이 낮아지지만, 일정 기간(약 15년)이 지나가면 희소가치에 의한 프리미엄으로 가격이 높아진다. 최대 토크의 경우, 최대 토크가 클수록 가격이 높아지는 경향이나 두 모형에서 다른 모습을 보이고 있다. 약 30kg.m 이하인 자동차는 신차 대비 중고차 가격의 비율 예측에 영향을 미치지 않았다. 그 외에 중고 가격 예측에 영향을 미치는 요인은 타이어 휠 크기, 신차 가격, 최고 출력이고, 가격 비율 예측에 영향을 미치는 요인으로 보증정보, 연비, 차량 중량이 도출되었다. 두 개의 모형에서 보험처리 이력, 성능점검 정보와 옵션 정보는 가격 예측에 상대적으로 덜 중요하였다. 그리고 연식을 3년과 5년으로 지정하여 모형을 비교하였을 때 중요 요인은 비슷했으나, 중요 변수에 따른 가격 비율의 경향은 달랐다.

수입차의 경우 자료의 편의가 많아 국산차로 제한하여 분석하였다. 수입차의 자료를 수집할 수 있다면 동일한 방법으로 수입 중고 자동차를 연구할 수 있을 것이다. 본 연구에서 제안한 최종 모형을 활용하면 향후 판매자와 소비자에게 도움이 될 것이다. 판매자는 합리적인 가격을 제시할 수 있으며, 소비자는 판매되고 있는 자동차가 허위 매물 판단 자료로 활용될 수 있다.

참 고 문 헌

류석일 (2011). 중고차 온라인 중개사이트 문제점 및 개선방안 조사 연구. 조사보고서, , 1-38.

윤대권, 김용현, 이해택, 하성용 (2015). 중고자동차 가격산정을 위한 평가요인 연구. 한국자동차공학회 춘계학술대회, 1095-1100.

한상식, 전종근 (2017). 온라인-오프라인 연계 판매 플랫폼에서 수입중고차의 가격 프리미엄과 재고기간에 대한 영향 요인. e-비즈니스연구, 18(4), 37-50.

Breiman, L. (2001). Random forests, *Machine Learning*, 45, 5-32.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning*, 20, 273-297.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York, USA.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Springer, New York, USA.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm packages, https://cran.r-project.org/web/packages/gbm.

Tianqi C. and Carlos G. (2016). XGBoost: A Scalable Tree Boosting System, KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Venables. W. N. and Ripley. B. D. (2002). *Modern Applied Statistics with S*, 4th Edition. Springer, New York.

ABSTRACT

A study on the domestic used car price

Cho, Soojin
Department of Statistics
The Graduate School
Ewha Womans University

The purpose of this study is to analyze the price of domestic used car in the online market by using various data mining techniques and to identify the factors influencing used car price through the prediction model. We implement two models to predict used car price and the ratio of used car price to new car price. We used 6 methods for the analysis such as stepwise linear regression, random forest, gradient boosting, support vector machine, extreme gradient boosting, and neural network model. Optimal model by RMSE, which is a prediction evaluation index, is random forest. Period of use came out to be the most important variable in both models. In addition, new car price and tire width significantly affect the used car price model and mileage is an influential variable in the ratio of used car price to new car model.