머신러닝을 활용한 브랜드별 국내 중고차 가격 예측 모델에 관한 연구

임승준*, 이정호**, 류춘호***

목 차

요약

국내 중고차 시장은 지속적으로 성장하고 있으며, 이와 동시에 중고차 온라인 플랫폼 서비스 역시 함께 매년 시장 점유율을 확대하고 있다. 중고차 온라인 플랫폼 서비스는 차량의 제원, 점검 이력, 사고 내역, 그리고 세부 옵션 등을 서비스 이용자에게 제공하고 있다. 대부분의 기존 연구는 차량의 제원과 차량의 일부 옵션을 활용한 중고차 가격의 예측이었으며, 중고차 가격과 일부 제원 변수 간 비선형 관계임을 확인하였다. 이에 따라 연구자들은 이러한 비선형 문제를 해결하기 위해 머신러닝(Machine Learning) 모델의 실행을 제안하였으며, 그 결과 회귀(Regression) 기반 머신러닝 모델은 변수의 실질적인 영향력과 방향성을 알 수 있는 장점이 존재하였으나, 트리(Decision Tree) 기반 머신러닝 모델에 비해 비용함수 수치가 저조한 단점이 존재하였다.

본 연구는 국내 브랜드를 대상으로 차량의 제원과 차량의 옵션, 총 70여 개의 변수를 모두 활용하여 회귀 기반 머신러닝 모델과 트리 기반 머신러닝 모델을 순차적으로 실행하여 두 유형의 머신러닝 모델의 장점을 취합하고자 하였다. 이를 통해 브랜드별 변수의 실질적 영향력과 방향성을 확인한 후 브랜드별 가장 우수한 트리 기반 머신러닝 모델을 선정하였다.

본 연구의 시사점은 다음과 같다. 중고차 온라인 플랫폼 서비스를 이용하는 구매자와 판매자가 전반적인 중고차 가격 예측을 지원할 수 있다. 이에 따라 중고차 온라인 플랫폼 서비스 이용자 간 정보의 비대칭으로 인한 문제 해결 역시 지원이 가능할 것으로 기대한다.

표제어: 중고차 온라인 플랫폼 서비스, 중고차 가격, 브랜드, 라쏘 회귀 머신러닝, 트리 기반 머신러닝

접수일(2023년 07월 10일), 수정일(2023년 08월 07일), 게재확정일(2023년 08월 08일)

* 제1저자, 홍익대학교 경영학 박사, yimdgz@hotmail.com
** 교신저자, 건국대학교 경영학과 조교수, jholee@konkuk.ac.kr
*** 공동저자, 홍익대학교 경영학과 교수, ryuch@hongik.ac.kr

1. 서론

국내 중고차 시장은 현재 지속적으로 성장하고 있다. 2020년 기준 국내의 중고차 거래량은 전년도에 비해 약 7.2% 성장하였고, 거래량은 총 380만 대를 돌파하였다. 또한 중고차 온라인 플랫폼 서비스 역시 국내와 해외 모두 지속적으로 성장하고 있으며, 2021년 기준 국내 중고차 시장은 오프라인 거래와 온라인 거래 간 거래량은 비슷한 것으로 확인되었다(Lee, 2022). 중고차 온라인 플랫폼 서비스는 차량 제원뿐만 아니라 차량 옵션, 사고 이력, 그리고 점검 이력까지 자세한 정보를 플랫폼 이용자에게 제공하고 있다.

머신러닝(Machine Learning)은 컴퓨터가 명시적으로 프로그래밍 되지 않고 학습할 수 있는 능력을 주는 것을 의미한다(Samuel, 1959), 최근 들어 정보처리기술 및 분석 알고리즘의 발전과 빅데이터의 발현으로 인해 머신러닝의 활용이 점차 확대되고 있다(Lee, 2022). 또한 머신러닝은 컴퓨터 과학, 통계학, 나아가 사회과학에서도 활용하기 시작하였다(Kim and Lee, 2021; Lee and Shin, 2023).

기존의 중고차 가격 예측을 위한 선행연구는 차량 제원 변수 및 일부 차량 옵션 변수만을 활용하여 차량 가격 예측이 대부분이었다. 특히 차량 제원 변수 중 차량의 운행 거리와 사용기간은 중고차 가격에 미치는 영향력이 타 제원 변수에 비해 상대적으로 크며, 이들 변수와 중고차 가격 간 관계가 비선형 관계임을 확인하였다(Chaudhary et al., 2022; Chen et al., 2017; Lasek and Wyszyński, 2019; Khan, 2022; Yennimar et al., 2022). 이에 따라 차량의 제원과 중고차 가격 간 비선형적 관계를 해결하고자 트리 기반 머신러닝 모델을 활용하여 가장 우수한 모델을 선정하였다. 그러나 이와 같은 연구 결과는 국가마다 변수의 영향력이 차이가 존재할 뿐만 아니라 브랜드별 차이가 존재할 수 있다.

머신러닝 모델을 활용한 연구에서 회귀(Regression) 기반의 머신러닝 모델은 변수의 실질적 영향력과 방향성을 확인할 수 있는 장점이 있으며, 트리(Decision Tree) 기반의 머신러닝 모델은 비용함수(Cost Function) 수치에서 회귀 기반 머신러닝 모델보다 우수하나 변수의 실질적 영향력과 방향성을 알기 어렵다는 한계점이 존재하였다(Venkatasubbu and Ganesh, 2019; Chaudhary et al., 2022). 또한 앞서 언급한 연구들은 모두 차량 제원과 일부 옵션만 활용하여 차량 가격을 예측하였으며, 차량 가격을 범주형 변수로 변환하여 척도 내 오차가 존재할 가능성이 존재하였다.

이에 따라 본 연구는 첫째, 회귀 기반 머신러닝 모델의 장점과 트리 기반 머신러닝 모델의 장점을 취합하여 그 활용성을 최대한 확보한다. 이를 통해 변수의 영향력과 방향성을 확인한 후 가장 우수한 트리 기반 머신러닝 모델을 선정하고자 하였다. 둘째, 차량 제원과 옵션 변수, 총 70여 개를 활용하여 더욱 정확한 중고차의 가격을 예측하고자 하였다. 셋째, 결과값인 차량 가격을 연속형으로 측정하여 척도 내 오차를 해결하고자 하였다. 이와 같은 과정을 통해 중고차 온라인 플랫폼 서비스를 이용하는 구매자와 판매자가 차량의 제원과 옵션을 통해 더욱 정확한 중고차 가격을 유추하는데 도움이 되고자 하였으며, 중고차 온라인 플랫폼 서비스 이용자 간 정보의 비대칭으로 인한 문제 해결에 도움이 되고자 하였다.

본 연구는 우선 온라인 중고차 플랫폼 서비스로부터 차량의 제원과 옵션을 크롤링(Crawling)하여 자료를 획득한다. 다음으로 브랜드별로 Lasso 회귀 머신러닝 모델을 실행하여 결과값에 영향을 미치는 변수의 실질적 영향력과 방향성을 확인하고, 영향력이 0으로 나타난 변수들을 도출한다. 마지막으로 영향력이 0으로 나타난 변수를 소거한 자료를 대상으로 트리 기반 머신러닝 모델을 실행하여 가장 우수한 모델을 선정한다.

2. 선행연구

2.1 국내 중고차 시장

최근 3년간 중고차 거래량은 매년 증가하는 추세에 있다. 2020년도 거래량은 전년 대비 7.2% 증가했을 뿐만 아니라 거래 금액 모두 매년 성장하고 있다. 2020년 기준 거래량 기준 중고차 수는 총 387만 대이며, 총 거래 금액은 30조 원에 근접하고 있다. 이 중 소비자에게 판매된 중고차 수는 총 251만 대이며, 소비자에게 판매된 신차의 수인 190만 대의 1.32배에 달한다(Lee, 2022).

자동차 선진국인 독일의 경우 중고차 판매 대수는 2019년 기준 총 719만 대로 신차 판매 대수 총 360만 대의 두 배에 달한다. 또한 자동차 선진국인 동시에 전 세계에서 가장 큰 자동차 시장을 보유한 미국의 경우 중고차 판매 대수는 2019년 기준 총 4,081만 대로 신차 판매 대수 총 1,706만 대의 2.4배에 달하며, 중고차 거래액은 8,406억 달러로 신차 거래액 6,365억 달러에 비해 높은 것으로 나타났다(Yang, 2020). 이와 같은 결과를 통해 현재 국내 중고차 시장은 지속적으로 성장할 가능성이 높다.

국내 중고차 시장뿐만 아니라 해외 중고차 시장에서 온라인 플랫폼 서비스를 통한 중고차 거래의 규모가 점차 확대하고 있다. 미국의 중고차 온라인 플랫폼 서비스로 Carmax, Vroom, 그리고 Carvana 등이 대표적이며, 국내의 중고차 온라인 플랫폼 서비스로 엔카, K-Car, 그리고 KB차차차 등이 대표적이다. 미국 중고차 시장의 경우 온라인 플랫폼 서비스 중 하나인 Cavana는 2021년도 기준 거래량이 42.5만대로 이전의 3년 동안 거래량 대비 5.5배가 증가했으며, 국내 중고차 시장의 경우 2021년 4분기 기준 중고차 전체 거래량에서 온라인 플랫폼 서비스를 통한 중고차 거래량은 38.3%로 오프라인 중고차 거래량 41.0%에 근접할 정도로 점차 확대되고 있다(Lee, 2022).

2.2 회귀 분석을 활용한 중고차 가격 예측

Lasek and Wyszyński (2019)는 폴란드의 중고차 시장에서 중고차의 사용기간, 동력성능, 주행거리, 그리고 자동변속기 여부가 중고차 가격에 영향을 미친다는 것을 확인하였다. BALCE (2016)는 튀르키의 중고차 시장에서 중고차의 엔진 유형, 차량의 유형, 그리고 변속기의 유형이 중고차 가격에 영향을 미치며, 자동차의 손상 수준이 중고차 가격에 조절한다는 것을 검증하였다. Arawomo and Osigwe (2016)는 나이지리아 중고차 시장에서 중고차의 안전 수준, 연비, 그리고 고급 옵션이 중고차 가격을 상승시키는 요인임을 확인하였다. 세부적으로 차량의 안전 수준인 차량 에어백 유무, 차량의 고급 기능인 차량 시트 백, 알로이휠, 에어컨, 가죽 시트의 유무가 중고차 가격에 영향을 미치는 요인으로 나타났다. Meng et al., (2019)은 대만 중고차 시장에서 중고차의 제원 변수, 인테리어 수준, 웨건 여부, 그리고 토요타 여부가 중고차 가격에 영향을 미친다는 것을 확인하였다. 이 중 차량 인테리어 수준이 가장 큰 영향력을 미친다는 것을 확인하였다. 이들의 연구에서 차량 엔진 유형, 변속기 유형, 그리고 브랜드 종류 등이 차량 가격에 미치는 영향은 국가마다 차이가 있을 수 있으므로 국내 중고차 시장에 그대로 적용하기 어렵다는 한계점이 존재한다.

2.3 머신러닝 모델을 활용한 중고차 가격 예측

Khan (2022)은 Kaggle의 차량 제원 변수를 대상으로 선형 회귀 모델, Lasso 회귀 모델, 그리고 Ridge 회귀 모델을 활용하여 중고차 가격을 예측하였다. 그 결과 Lasso 회귀 모델이 가장 우수한 모델임을 확인하였다. Venkatasubbu and Ganesh (2019)는 인도의 중고차 시장을 대상으로 선형 회

귀 모델, Lasso 회귀 모델, Ridge 회귀 모델, 그리고 CART(Classification and Regression Tree) 모델을 활용하여 중고차 가격을 예측하였다. 또한 독립변수로 차량 제원, 크루즈 컨트롤 여부, 그리고 내장의 가죽 여부를 활용하였다. 그 결과 가장 우수한 모델은 Lasso 회귀 모델로 확인되었다. Chaudhary et al., (2022)은 인도 중고차 시장을 대상으로 차량 제원 변수를 통해 선형 회귀 모델, Ridge 회귀 모델, Lasso 회귀 모델, CART 모델, 그리고 RF(Random Forest) 모델을 실행하여 중고차 가격을 예측하였다. 그 결과 RF 모델이 가장 우수한 모델로 선정되었다. Yennimar et al., (2022)은 중고차 가격을 예측하기 위해 선형 회귀 모델과 SVM(Support Vector Machine) 모델을 활용하였다. 그 결과 SVM 모델이 선형 회귀 모델보다 우수한 모델임을 확인하였으며, 비선형 자료를 분석할 시 큰 용이점이 있다고 하였다.

Yadav et al., (2021)은 인도 중고차 시장을 대상으로 선형 회귀 모델, RF 모델, 그리고 ANN(Artificial Neral Network) 모델을 활용하여 중고차 가격을 예측하였다. 그 결과 RF 모델이 가장 우수한 모델임을 검증하였다. Shanti et al., (2021)은 아랍권 중고차 시장을 대상으로 머신러닝 중 ANN 모델, SVM 모델, RF 모델, 그리고 GB(Gradient Boosting) 모델을 활용하여 중고차 가격을 예측하였다. 그 결과 RF 모델과 GB 모델이 가장 우수한 모델로 확인되었다, Gegic et al., (2019)은 보스니아 중고차 시장을 대상으로 SVM, ANN, 그리고 RF 모델을 활용하여 중고차 가격을 예측하였다. 그 결과 RF 모델이 가장 우수한 것으로 나타났다.

Chen et al., (2017)은 중국의 중고차 시장을 대상으로 선형 회귀 모델과 RF 모델을 활용하여 중고차 가격을 예측하였다. 그 결과 대부분의 브랜드에서 RF 모델이 가장 우수한 모델임을 확인함과 동시에 변수와 표본의 크기가 큰 복잡한 모델일 경우 최적의 알고리즘이라는 것을 검증하였다. Nasiboglu and Akdogan (2020)은 중고차 온라인 플랫폼 자료를 대상으로 선형 회귀 모델, Ridge 회귀 모델, Lasso 회귀 모델, Elastic.Net 회귀 모델, KNN(K Nearest Neighbor) 모델, RF 모델, 그리고 XGB(eXtream Gradient Boosting) 모델을 활용하여 브랜드별 중고차 가격을 예측하였다. 이를 통해 브랜드별 가장 우수한 모델을 선정하였다.

앞서 소개한 선행연구는 대부분 다음과 같은 한 계점이 존재하였다. 첫째, 대부분이 차량의 제원 변수와 일부 옵션 변수만을 활용하여 중고차 가격을 예측하였다. 둘째, 회귀 기반 머신러닝 모델을 제외한 머신러닝 모델은 변수의 실질적 영향력과 방향성을 알 수 없었다. 셋째, 결과값인 중고차 가격을 서열척도로 구성하여 척도 내 오차(Error)를 고려하지 못하였다. 또한 Chen et al., (2017), Meng et al., (2019), 그리고 Nasiboglu and Akdogan (2020)의 연구 결과에 따르면 브랜드 간 가장 우수한 머신러닝 모델이 다를 수 있다는 것을 확인하였다.

본 연구는 우선 차량의 제원 변수와 차량의 옵션 변수를 70여 개를 모두 활용하여 더욱 정확한 중고차 가격 예측을 위한 머신러닝 모델을 실행하고자 하였다. 다음으로 Lasso 회귀 머신러닝 모델과 트리 기반 머신러닝 모델을 순차적으로 실행하여 브랜드별 변수의 실질적 영향력과 방향성을 확인한 후 브랜드별 가장 우수한 머신러닝 모델을 선정한다. 또한 결과값인 중고차 가격을 연속형 변수로 구성하여 선행연구들의 척도 내 오차를 해결하고자 하였다.

3. 자료 수집 및 변수 설정

3.1 자료 수집

본 연구는 엔카(http://www.encar.com)에 기재된 중고차의 정보를 활용하였다. 엔카는 서비스 이용자

에게 차량 제원, 세부 옵션, 성능 점검, 사고 이력, 그리고 차량 가격을 제공하고 있다. 또한 국내 시장에서 출시한 현세대와 직전 세대의 차량을 본 연구의 대상으로 하였다. 이를 위해 엔카의 차량 정보(제원 및 옵션)와 개시된 차량 가격을 대상으로 크롤링(Crawling)을 실행하였다. 또한 엔카에 미공개된 연비, 마력, 공차중량, 그리고 토크는 차량 제조사의 공식 제원표를 활용하였다.

국내 중고차 온라인 플랫폼 서비스에 업로드된 국내 브랜드의 현세대와 직전 세대의 상용차의 총 대수는 28,751대로 확인되었다. 제조사별 차량 대수는 기아가 10,008대, 르노가 2,062대, 쉐보레가 2,373대, 쌍용이 1,859대, 제네시스가 3,730대, 그리고 현대가 8,719대로 나타났다.

3.2 변수의 정의와 측정

본 연구에서 중고차 가격 예측을 위한 변수는 크게 두 가지로 구성되었다.

첫 번째 변수인 차량 제원은 세부 변수로 사용연료(1: 디젤, 2: 가솔린), 차종(1: 세단, 2: SUV, 3: Hybrid, 4:경차), 마력, 토크, 주행거리, 사용기간, 연비, 차량 크기(1: 경형, 2: 소형, 3: 준중형, 4: 중형, 5: 준대형, 6: 대형), 공차중량, 차량 구동 방식(1: 전륜, 2: 후륜, 3: 4륜), 배기량, 그리고 색상(브랜드별 빈도순)으로 구성하였다. 또한 사용 연료가 LPG인 차량과 리스 및 렌탈 승계 차량은 장애인 여부 확인과 중고차 가격 측정의 어려움으로 인해 본 연구의 자료에서 제외하였다.

두 번째 변수인 차량 옵션은 Arawomo and Osigwe (2016), Meng et al., (2019), Venkatasubbu and Ganesh (2019), 그리고 Gegic et al., (2019)의 선행연구에서 활용된 변수를 활용하였다. 본 연구에서 이 변수들은 네 가지 유형으로 분류하였다. 첫째, 외관 및 내관 관련 옵션의 유형으로 헤드램프, 썬루프, 전동트렁크, 전동사이드미러, 고스트도어클로징, 알루미늄휠, 열선스티어링, 루프랙, 전동스티어링, 패들시프트, ECM룸미러, 파워도어록, 하이패스, 파워스티어링, 스티어링리모콘, 그리고 파워윈도우의 유무(0: 무, 1: 유)로 구성되어 있다. 둘째, 안전 관련 옵션 유형으로 에어백운전석, 에어백동승석, 에어백측면, 에어백커튼, 브레이크 잠김 장치, 미끄럼방지, 차체자세 제어장치, 타이어 공기압센서, 차선이탈경보, 전자제어 서스펜션, 주차감지센서전방, 주차감지센서후방, 후측방경보, 후방카메라, 그리고 360도 어라운드뷰의 유무(0: 무, 1: 유)로 구성되어 있다. 셋째, 편의 및 멀티미디어 관련 옵션 유형으로 크루즈컨트롤, 헤드업 디스플레이, 전자식 주차브레이크, 자동에어컨, 스마트키, 무선도어잠금, 비센서, 오토라이트, 블라인드뒷좌석, 블라인드후방, 내비게이션, 전면AV모니터, 후면AV모니터, 블루투스, CD플레이어, USB단자, 그리고 AUX단자의 유무(0: 무, 1: 유)로 구성되어 있다. 넷째, 시트 관련 옵션 유형으로 가죽시트, 전동시트운전석, 전동시트동승석, 전동시트뒷좌석,열선시트앞좌석, 열선시트뒷좌석, 메모리시트운전석, 메모리시트동승석, 통풍시트운전석, 통풍시트동승석, 그리고 통풍시트뒷좌석의 유무(0: 무, 1: 유)로 구성되어 있다.

차량의 세부 옵션에서 헤드램프는 옵션 없음이 0, HID 램프 옵션이 1, 그리고 LED 램프 옵션을 2로, 크루즈 컨트롤은 옵션 없음이 0, 일반 크루즈 컨트롤 옵션이 1, 아답티브 크루즈 컨트롤 옵션을 2로 코딩하여 가격에 따른 차등을 두었다.

4. 머신러닝 모델의 실행 과정

본 연구의 머신러닝 모델 실행과정은 다음과 같다. 첫째, 수집한 자료를 국내 6개 브랜드로 분리한다. 둘째, 선행연구에서 회귀 기반 머신러닝 모델 중 가장 우수한 Lasso 회귀 모델(Khan, 2022; Venkatasubbu and Ganesh, 2019)을 실행하여 변수의

실질적 영향력과 방향성을 확인한다. 마지막으로 트리(Decision tree) 기반 머신러닝 모델을 실행하여 비용함수 수치가 가장 우수한 모델을 선정한다.

4.1 과대적합과 K-Fold 교차검증

머신러닝 모델의 실행에 앞서 자료를 대상으로 훈련 세트(Training Set)와 시험 세트(Test Set)로 분리하는 샘플링(Sampling) 작업을 실행한다. 머신러닝 모델의 실행 시 훈련 세트의 점수가 시험 세트의 점수에 비해 과도하게 높으면 과대적합(Overfitting), 반대는 과소적합(Underfitting)이라 하며, 과대적합이 과소적합보다 더욱 위험하다. <Fig. 4-1>은 과소·과대적합의 그래프이다(Ramampiandra et al., 2023).

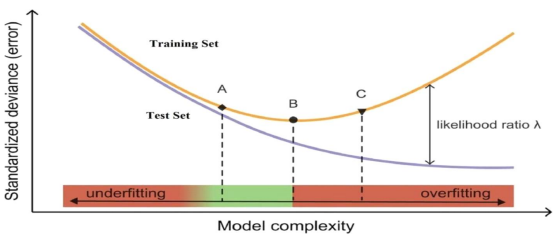

Fig. 4-1 Overfitting and Underfitting

과대적합과 과소적합의 균형점을 찾는 것은 매우 중요한 과정이며, 이를 위해 K-Fold 교차검증을 실행한다. K-Fold 교차검증은 훈련 세트를 검증 세트(Validation Set)으로 분리하여 k번 반복한다.

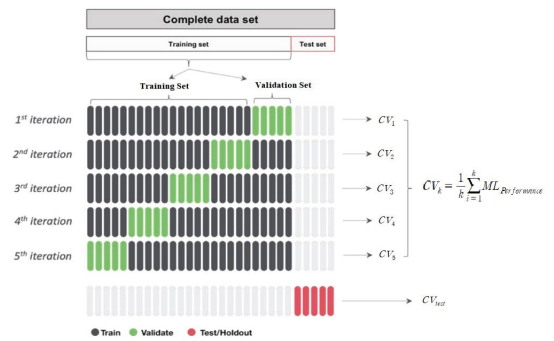

Fig. 4-2 K-Fold Validation Test

<Fig. 4-2>는 k가 5일 때의 K-Fold 교차검증의 원리이다(Staartjes et al., 2022). CV_k는 머신러닝 모델의 Performance의 평균, k는 교차검증의 횟수, 그리고 Perfomance는 훈련 세트를 통한 머신러닝 모델의 검증 세트에 대한 Y(실제값)과 $\hat{Y}$(예측값) 간 MSE(Mean Squared Error) 또는 R^2 수치이다.

4.2 Lasso 회귀 머신러닝 모델

Lasso 회귀 모델은 대표적인 회귀 기반 머신러닝 모델 중 하나이며, 기존의 선형 회귀의 과대적합을 해결하고자 제안된 모델이다. 선형 회귀의 목적함수는 비용함수 중 하나인 MSE의 최소화이다.

Lasso 회귀 모델은 추가적으로 페널티를 부여하여 MSE를 최소화한다. 이러한 페널티를 알파(α)라고 하며, 회귀계수(W_i)의 크기를 제어한다. 알파값이 증가할수록 회귀계수는 감소하며, 이 과정을 L1 규제(Regularization)라고 한다.

$$\begin{aligned} Objective\ Function: \quad & Min(MSE + Penalty) \\ & = Min(MSE + \alpha(L1)norm) \\ & = Min(MSE + \alpha\sum_{j=1}^{m} | W_j |) \end{aligned}$$

Lasso 회귀 모델은 회귀계수의 절댓값에 패널티를 부여하여 영향력이 상대적으로 미미한 회귀계수를 0으로 도출한다. 이를 통해 선형 회귀 모델을 일반화(Generalize)하게 만들며, 과대적합을 방지하는 효과가 있다(Khan, 2022; Venkatasubbu and Ganesh, 2019).

4.3 트리 기반 머신러닝 모델

트리(Decision Tree) 기반 머신러닝 모델은 결과값

의 종류에 따라 두 가지로 구분되며, 이에 따른 종류는 다음 <Tab. 4-1>과 같다. CART(Classification and Regression Tree) 모델은 트리 기반 머신러닝 모델 중에서 가장 기본인 모델이다. 본 연구에서 결과값인 중고차 가격은 연속형 변수이므로 Regression Tree 모델을 실행하며, 다른 트리 기반 머신러닝 모델 역시 모두 Regressor가 추가된 모델을 실행한다,

Tab. 4-1 Decision Tree based ML Model

결과값: 범주형	결과값: 연속형
CART(Classification Tree)	CART(Regression Tree)
Random Forest(RF)	RF Regressor(RFR)
Gradient Boosting(GB)	GB Regressor(GBR)
eXtream GB(XGB)	XGB Regressor(XGBR)
Light GB(LGB)	LGB Regressor(LGBR)

CART(Regression Tree) 모델의 분기는 가중 오차값(E_{weight})이 그 기준이며. 분기점의 목적함수는 가중 오차값의 최소화이다. 분기점에 따른 두 영역($area1$, $area2$)에서 MSE_{area1}는 $area1$ 내 결과값의 MSE이며, m_{area1}은 $area1$ 내 자료의 수를 의미한다 (Chaudhary et al., 2022).

$$Objective\ Function: Min(E_{weight})$$

$$E_{weight} = \frac{m_{area1}}{m} MSE_{area1} + \frac{m_{area2}}{m} MSE_{area2}$$

CART(Regression Tree) 모델에서 $area1$과 $area2$를 나누는 기준점인 가중 오차값(E_{weight})값을 최소화한 점은 노드(Node)라고 불리며, 첫 노드를 루트 노드(Loop Node), 마지막 노드를 리프 노드(Leaf Node)라고 불린다. 이와 같은 노드 생성 기준에 따라 트리 분할이 완성된 후 리프 노드 속 결과값의 평균을 최종 결과값으로 저장한다. 리프 노드는 Hyper Parameter Tuning을 통해 범위를 조정할 수 있다.

RF(Random Forest) 모델과 GB(Gradient Boosting) 계열 모델은 CART 모델의 과대적합 위험성을 해결하기 위해 제안된 모델이다. RFR(Random Forest Regressor) 모델은 CART(Regression tree) 모델을 n개 실행하여 이들의 결과값에 대한 평균을 도출한다. RF 모델은 과대적합을 방지하는 장점이 있으나, n의 크기가 어느 정도 증가했을 때 이후 성능의 향상폭이 작아지는 단점이 있다(Kwon, 2022).

GB(Gradient Boosting) 계열 모델 중 LGB(Light Gradient Boosting) 모델은 리프 노드를 지속적으로 분할하여(리프 중심 트리 분할) 예측 오류를 감소시킨다. 또한 LGB 모델은 실행시간이 상대적으로 짧은 편이나, 과대적합의 위험성이 상대적으로 높다.

GB 모델과 XGB(eXtream Gradient Boosting) 모델은 여러 개의 약한 학습기를 통해 학습하고 예측하면서 잘못 예측된 결과값에 가중치를 부여하는 과정을 통해 오류를 개선해간다. GB 모델과 XGB 모델은 트리의 깊이를 효과적으로 줄임(균형 트리 분할)으로서, 가장 우수한 트리 기반 머신러닝 모델로 평가받으나, 실행시간이 상대적으로 길다는 단점이 존재하였다. XGB 모델은 병렬수행(다중 CPU)을 지원하여 GB 모델에 비해 실행시간이 짧으며, 과대적합을 방지하는 장점이 있다(Kwon, 2022). 이에 따라 본 연구는 비용함수 측면에서 GB 모델과 큰 차이가 없는 XGB 모델을 본 연구에 활용한다.

4.4 Hyper Parameter Tuning

머신러닝 모델의 실행과정에서 K-Fold 교차검증 후 Hyper Parameter Tuning을 실행하여 최적화된 Hyper Parameter를 도출한다.

Tab. 4-2 Hyper Parameter

Hyper Parameter	Definition
α	Ridge/Lasso/Elastic.Net 회귀 시 설정
cv	교차검증 횟수
min_samples_split	노드 분할을 위한 최소한의 샘플 수
min_samples_leaf	리프 노드가 되기 위한 최소 샘플 수
max_features	최적의 분할을 위한 feature 수
max_depth	트리의 최대 깊이
max_leaf_nodes	리프 노드의 최대 개수
n_estimate	트리의 개수
subsample	모델 학습 시 사용할 데이터 비율

Hyper Parameter는 머신러닝 모델 실행 시 사용자가 지정하는 초매개변수라고 정의되며, 도출된 Hyper Parameter 값을 통해 최적화된 머신러닝 모델을 실행할 수 있다. Hyper Parameter는 머신러닝 모델마다 그 종류가 차이가 있으며, <Tab. 4-2>는 일반적으로 사용되는 Hyper Parameter의 목록이다 (kwon, 2022).

4.5 비용함수

본 연구에서 실행할 머신러닝 모델의 목적함수는 비용함수(Cost Function)의 최소화이다. 또한 결과값인 중고차 가격이 연속형 변수임에 따라 비용함수의 종류는 MSE, RMSE, MAE, 그리고 MAPE가 있으며, 상대적으로 작을수록 우수한 모델이다. 또한 비용함수는 아니지만 R^2이 있으며, 상대적으로 클수록 우수한 모델이다. <Tab. 4-3>에서 $\hat{Y}_i$는 머신러닝 모델을 통한 자료의 예측값, Y_i는 자료의 실제값, 그리고 n은 자료의 총개수를 의미한다.

Tab. 4-3 Cost Function & R^2

Cost Function & R^2	Definition
RMSE(Root Mean Squared Error)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}$
MSE (Mean Squared Error)	$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$
MAE (Mean Absolute Error)	$\frac{1}{n}\sum_{i=1}^{n}\lvert(Y_i-\hat{Y}_i)\rvert$
MAPE(Mean Absolute Percentage Error)	$\frac{1}{n}\sum_{i=1}^{n}\lvert\frac{(Y_i-\hat{Y}_i)}{Y_i}\times 100\rvert$
R^2(Coefficient of Determination)	$Var(\hat{Y})/Var(Y)$

5. 머신러닝 모델의 실행 결과

본 연구는 회귀 기반 머신러닝 모델과 트리 기반 머신러닝을 순차적으로 실행한다. 이에 따라 확보한 자료를 브랜드별로 분리한 후 이들 6개 브랜드의 Lasso 회귀 모델을 실행하여 변수의 실질적 영향력과 방향성을 확인하며, 중고차 가격에 미치는 영향력이 0인 변수를 도출한다. 다음으로 영향력이 0인 변수를 소거한 자료를 대상으로 브랜드별 트리(Decision Tree) 기반 머신러닝 모델인 CART(Regression Tree) 모델, RFR(Random Forest Regresssor) 모델, LGBR(Light Gradient Boosting Regressor) 모델, 그리고 XGBR(eXtream Gradient Boosting Regressor) 모델을 실행한다. 이와 같은 과정을 거쳐 비용함수를 비교하여 브랜드별 가장 우수한 모델을 선정한다.

5.1 기아

기아 브랜드의 차량 크기는 평균이 3.502, 표준편차가 1.569로 준중형차와 중형차 사이로 나타났다. 또한 차량의 마력, 배기량, 그리고 중량 역시 차량 크기와 동일한 특성으로 확인되었다.

자료의 특성을 확인한 후 회귀 기반 머신러닝 모델 중 하나인 Lasso 회귀 모델 실행을 위해 변수 중 연료와 차종은 더미변수로 설정하였다. 이후 자료의

훈련 세트의 비율을 20%, 10%, 그리고 1%로 설정한 후 훈련 세트와 시험 세트 간 Score를 비교하였고, 그 결과 두 세트 간 차이가 가장 작게 나타난 비율은 1%로 나타났다. 다음으로 Lasso 회귀 모델에 대한 K-Fold 교차검증을 실행하여 과대적합 여부를 판별하였다. 분석 결과 K-Fold 수치는 k는 6일 때 가장 우수한 것으로 확인되었다.

위와 같은 과정을 거친 후 최종적으로 Lasso 회귀 모델을 실행하여 비용함수, 표준화 회귀계수, 그리고 영향력이 0인 변수를 확인하였다. Lasso 회귀 모델의 비용함수와 R^2은 RMSE가 234.246, MSE가 54866.734, MAE가 173.797, MAPE가 8.439(%), 그리고 R^2이 0.956으로 나타났다. <Fig. 5-1>은 Lasso 회귀의 표준화 회귀계수(절대값 기준, 15위까지) 이다.

Tab. 5-1 K-Fold, R^2, Cost Function of Best Machine Learning Model by Brand

브랜드		KIA	RENAULT	CHEVROLET	SSANGYONG	GENESIS	HYUNDAI
차종		Sedan/SUV/Hybrid/Light Car					
자료 개수		10008	2062	2373	1859	3730	8719
소거된 변수 개수		7	19	20	18	28	14
최우수 머신러닝 모델		LGBR	XGBR	RFR	XGBR	RFR	LGBR
K-fold 교차검증	Test Set	0.01(1%)	0.01(1%)	0.01(1%)	0.01(1%)	0.01(1%)	0.01(1%)
	Score (k = 8)	0.975	0.939	0.957	0.957	0.961	0.955
Hyper Parameter Tuning (k = 8)		max_depth = 30 learning_rate = 0.1 subsample = 0.5 min_samples_leaf = 10 min_samples_split = 10 n_estimate = 600	max_depth = 40 learning_rate = 0.1 subsample = 0.9 n_estimate = 200	max_depth = 20 max_leaf_nodes = None min_samples_leaf = 5 min_samples_split = 10 n_estimate = 100	max_depth = 40 learning_rate = 0.1 subsample = 0.7 n_estimate = 200	max_depth = 30 max_leaf_nodes = None min_samples_leaf = 5 min_samples_split = 10 n_estimate = 200	max_depth = 30 learning_rate = 0.1 subsample = 0.5 min_samples_leaf = 10 min_samples_split = 10 n_estimate = 600
Cost Function & R^2	RMSE	169.195	95.350	58.153	79.830	206.931	112.861
	MSE	28627.082	9091.669	3381.725	6372.880	42820.544	12737.508
	MAE	106.850	71.205	44.481	67.087	158.138	83.491
	MAPE	4.389	4.610	4.528	3.631	3.902	3.742
	R^2	0.973	0.974	0.993	0.985	0.979	0.979

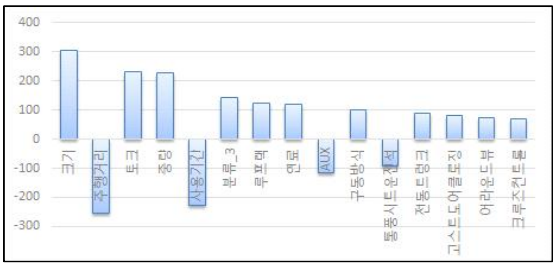

Fig. 5-1 Standardized Regression Coefficient of the Lasso in KIA

차량 제원 변수의 경우 크기(차량), 토크, 중량, 분류_3(Hybrid 여부) 연료(가솔린 여부), 그리도 구동방식 등이 중고차 가격에 양의 방향으로, 주행거리와 사용기간 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다. 차량 옵션 변수의 경우 루프랙, 전동트렁크, 고스트도어클로징, 어라운드뷰의 유무 그리고 크루즈 컨트롤 수준 등이 중고차 가격에 양의 방향으로, AUX와 통풍시트운전석 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다.

Lasso 회귀 모델에서 결과값에 미치는 영향력이 0으로 나타난 변수는 블라인드 후방, 블라인드 뒷자석, 무선도어잠금, 메모리시트동승석, 전자 안전성제어, 하이패스, 그리고 타이어공기압자동제어로 확인되었다. 이에 따라 이들 변수를 소거한 후 트리 기반 머신러닝 모델을 실행하였다.

트리 기반 머신러닝 모델의 실행에 앞서 이들 자료의 시험 세트의 비율을 1%로 설정한 후 CART 모델의 K-Fold 교차검증을 실행하여 과대적합 여부를 판별하였다. 분석 결과 CART 모델의 k는 8일 때 가장 우수한 것으로 확인되었으며, 이에 따라 모든 트리 기반 머신러닝 모델의 k를 8로 고정하였다. 다음으로 Hyper Parameter Tuning을 실행하여 최적의 초매개변수를 확인한 후 이를 활용하여 최적의 머신러닝 모델을 도출하였고, 최종적으로 도출된 머신러닝 모델의 비용함수와 을 확인하였다. 브랜드별 최우수 머신러닝 모델의 K-Fold 수치, 초매개변수, 비용함수, 그리고 R^2은 <Tab. 5-1>과 같다.

기아 브랜드에서 트리 기반 머신러닝 모델 간 비용함수와 R^2이 가장 우수한 모델은 LGBR 모델로 확인되었으며, 비용함수와 R^2은 RMSE가 169.195, MSE가 24627.082, MAE가 106.850, MAPE가 4.389(%), 그리고 R^2이 0.973으로 나타났다.

<Fig. 5-2>는 기아 브랜드에 대한 LGBR 모델의 결과값 예측을 위한 주요 변수(15위)를 보여준다. LGBR 모델은 변수의 중요도를 거리(Distance)로 표현되며, 트리 기반 머신러닝 모델의 주요 변수는 회귀 기반 머신러닝 모델과는 다르게 방향성이 없으며, 오로지 결과값을 도출하기 위한 중요도 순으로 나타난다.

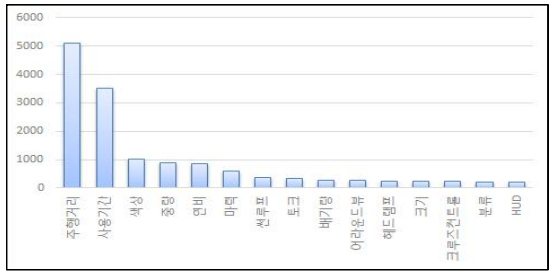

Fig. 5-2 LGBR Variable Ranking in KIA

<Fig. 5-1>과 <Fig. 5-2>를 통해 회귀 기반 머신러닝 모델과 트리 기반 머신러닝 모델 간 주요 변수의 순위와 영향력은 차이가 존재하며, 트리 기반 머신러닝 모델 간에도 주요 변수의 순위가 다른 것을 확인하였다. 트리 기반 머신러닝 모델은 CART 모델 알고리즘이 그 기반이다. 연구자는 트리의 깊이, 트리의 개수, 그리고 node 수 등을 Hyper Parameter Tuning의 실행을 통해 결정하며, 그 밖의 모든 연산은 컴퓨터 스스로가 처리한다. 이러한 과정으로 인해 회귀 기반 머신러닝 모델과 트리 기반 머신러닝 모델 간 결과의 차이가 존재하는 것으로 판단된다. 즉 트리 기반 머신러닝 모델은 회귀 기반 머신러닝 모델보다 비용함수 측면에서 상대적으로 우수하나,

변수의 실질적 영향력과 방향성을 알기 어렵다.

<Fig. 5-3>은 기아 브랜드에 대한 LGBR 모델의 실제값과 예측값 간 오차에 대해 시각화한 것이다. 가로축 Y는 시험 세트의 실제값, 세로축 $\hat{Y}$는 시험 세트의 예측값이다. 기아 브랜드에 대한 LGBR 모델의 실제값(선)과 예측값(점) 간 오차는 비교적 고르게 분포하며, 오차가 크지 않다고 판단된다.

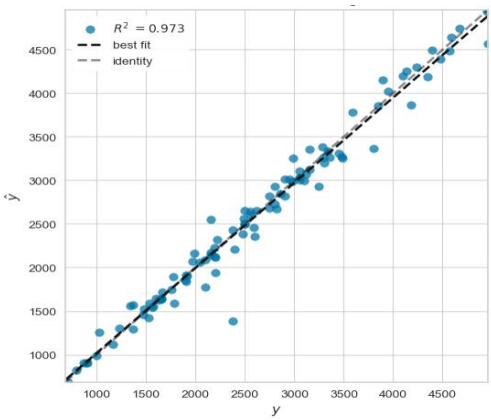

Fig. 5-3 Visualization of LGBR in KIA

5.2 르노

르노 브랜드의 차량 크기는 평균이 3.600, 표준편차가 0.709로 준중형차와 중형차 사이로 나타났다. 또한 차량의 마력, 배기량, 그리고 중량 역시 차량 크기와 동일한 특성으로 확인되었다.

자료의 특성을 확인한 후 훈련 세트의 적정 비율을 확인하고자 훈련 세트와 시험 세트 간 Score를 비교하였고, 그 결과 두 세트 간 차이가 가장 작게 나타난 비율은 1%인 것으로 나타났다. 다음으로 Lasso 회귀 모델에 대한 K-Fold 교차검증을 실행하여 k는 4일 때 가장 우수한 것으로 확인되었다.

위와 같은 과정을 거쳐 최종적으로 Lasso 회귀 모델을 실행하였다. 그 결과 RMSE가 163.819, MSE가 26836.733, MAE가 121.414, MAPE가 7.961, 그리고 이 0.923으로 나타났다.

차량 제원 변수의 경우 중량, 연료(가솔린 여부), 연비, 그리고 크기 등이 중고차 가격에 양의 방향으로, 사용기간, 차종(세단 여부), 그리고 주행거리 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다. 차량 옵션 변수의 경우 마사지시트, 주차감지센서전방, 메모리시트운전석, 크루즈컨트롤 수준, 루프랙, 열선스테어링의 유무 등이 중고차 가격에 양의 방향으로, 패들시프트와 하이패스 유무 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다.

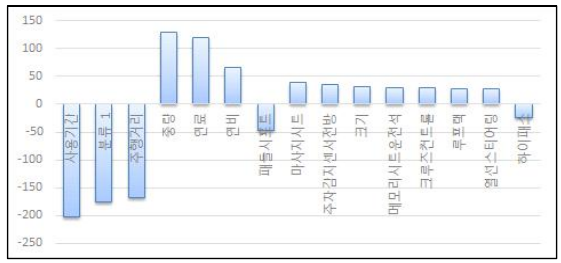

Fig. 5-4 Standardized Regression Coefficient of the Lasso in RENAULT

Lasso 회귀 모델에서 결과값에 미치는 영향력이 0으로 나타난 변수는 무선도어잠금, 블루투스, 비센서, 오토라이트, 통풍시트운전석, 통풍시트뒷좌석, 후측방경보, 전면AV모니터, 후방카메라, AUX, 스티어링리모콘, 에어백운전석, 마력, 구동방식, 고스트도어클로징, ECM, 파워스티어링, ESC, 그리고 에어백측면의 유무로 확인되었다. 이에 따라 이들 변수를 소거하여 트리 기반 머신러닝 모델을 실행하였다.

트리 기반 머신러닝 모델의 실행에 앞서 이들 자료의 시험 세트의 비율을 1%로 설정한 후 CART 모델의 K-Fold 교차검증을 실행하여 k가 6에서 8일 때 가장 우수한 것을 확인하였으며, 이에 따라 모든 트리 기반 머신러닝 모델의 k를 8로 고정하였다. 다음으로 Hyper Parameter Tuning을 실행하여 최적의 초매개변수를 확인한 후 이를 활용하여 최적의 머신러닝 모델을 도출하였다.

르노 브랜드에서 트리 기반 머신러닝 모델 간 비용함수와 R^2이 가장 우수한 모델은 XGBR 모델로 확인되었으며, RMSE가 95.350, MSE가 9091.669, MAE가 71.205, MAPE가 4.610, 그리고 R^2이 0.974로 나타났다.

<Fig. 5-5>를 통해 르노 브랜드에 대한 XGBR 모델의 실제값(선)과 예측값(점) 간 오차가 비교적 고르게 분포하며, 오차가 크지 않다고 판단된다.

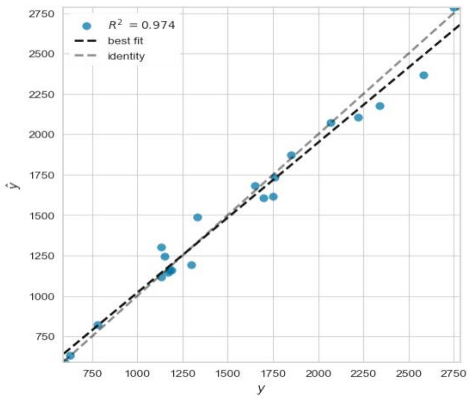

Fig. 5-5 Visualization of XGBR in RENAULT

5.3 쉐보레

쉐보레 브랜드의 차량 크기는 평균이 1.448, 표준편차가 0.797로 경차와 소형차 사이로 나타났다. 또한 차량의 마력, 배기량, 그리고 중량 역시 차량 크기와 동일한 특성으로 확인되었다.

자료의 특성을 확인한 후 훈련 세트의 적정 비율을 확인하고자 훈련 세트와 시험 세트 간 Score를 비교하였고, 그 결과 두 세트 간 차이가 가장 작게 나타난 비율은 1%인 것으로 나타났다. 다음으로 Lasso 회귀 모델에 대한 K-Fold 교차검증을 실행하여 k는 4일 때 가장 우수한 것으로 확인되었다.

위와 같은 과정을 거쳐 최종적으로 Lasso 회귀 모델을 실행하였다. 그 결과 RMSE가 79.064, MSE가 6251.148, MAE가 61.033, MAPE가 6.965, 그리고 R^2이 0.970으로 나타났다.

차량 제원 변수의 경우 마력, 배기량, 연료(가솔린 여부), 그리고 구동방식 등이 중고차 가격에 양의 방향으로, 사용기간, 주행거리, 차종(세단 여부) 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다. 차량 옵션 변수의 경우 전자식파킹브레이크, 열선시트뒷자석, 어라운드뷰, 메모리시트동승석, 패들시프트, 전동스티어링, HUD의 유무 등이 중고차 가격에 양의 방향으로 영향을 미치는 것으로 확인되었다.

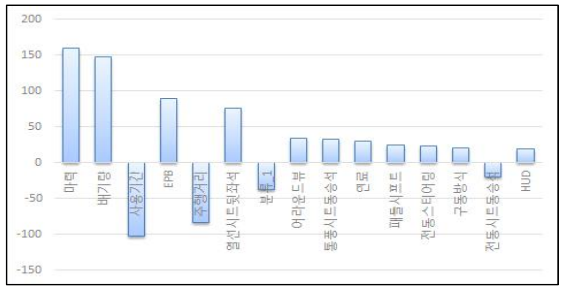

Fig. 5-6 Standardized Regression Coefficient of the Lasso in CHEVROLET

Lasso 회귀 모델에서 결과값에 미치는 영향력이 0으로 나타난 변수는 무선도어잠금, 블라인드후방, 통풍시트뒷좌석, 통풍시트운전석, 전동시트뒷좌석, 메모리시트동승석, 마사지시트, 후측방경보, 중량, 크기, 헤드램프, 주차감지센서후방, 고스트도어클로징, 알루미늄휠, ECM, 토크, 에어백운전석, 에어백동승석, ABS, 그리고 트랙션컨트롤시스템의 유무로 확인되었다. 이에 따라 이들 변수를 소거하여 트리 기반 머신러닝 모델을 실행하였다.

트리 기반 머신러닝 모델의 실행에 앞서 이들 자료의 시험 세트의 비율을 1%로 설정한 후 CART 모델의 K-Fold 교차검증을 실행하여 k가 6에서 8일 때 가장 우수한 것을 확인하였으며, 이에 따라 모든 트리 기반 머신러닝 모델의 k를 8로 고정하였다. 다음으로 Hyper Parameter Tuning을 실행하여 최적의

초매개변수를 확인한 후 이를 활용하여 최적의 머신러닝 모델을 도출하였다.

쉐보레 브랜드에서 트리 기반 머신러닝 모델 간 비용함수와 R^2이 가장 우수한 모델은 RFR 모델로 확인되었으며, RMSE가 58.152, MSE가 3381.725, MAE가 44.481, MAPE가 4.528(%), 그리고 R^2이 0.993으로 나타났다.

<Fig. 5-7>을 통해 쉐보레 브랜드에 대한 RFR 모델의 실제값(선)과 예측값(점) 간 오차가 비교적 고르게 분포하며, 오차가 크지 않다고 판단된다.

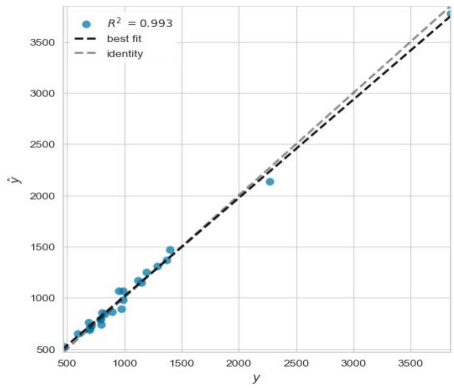

Fig. 5-7 Visualization of RFR in CHEVROLET

5.4 쌍용

쌍용 브랜드의 차량 크기는 평균이 2.870, 표준편차가 1.260으로 준중형차와 중형차 사이로 나타났다. 또한 차량의 마력, 배기량, 그리고 중량 역시 차량 크기와 동일한 특성으로 확인되었다.

자료의 특성을 확인한 후 훈련 세트의 적정 비율을 확인하고자 훈련 세트와 시험 세트 간 Score를 비교하였고, 그 결과 두 세트 간 차이가 가장 작게 나타난 비율은 1%인 것으로 나타났다. 다음으로 Lasso 회귀 모델에 대한 K-Fold 교차검증을 실행하여 k는 4일 때 가장 우수한 것으로 확인되었다.

위와 같은 과정을 거쳐 최종적으로 Lasso 회귀 모델을 실행하였다. 그 결과 RMSE가 111.058, MSE가 12333.903, MAE가 75.292, MAPE가 4.501(%), 그리고 R^2이 0.939로 나타났다.

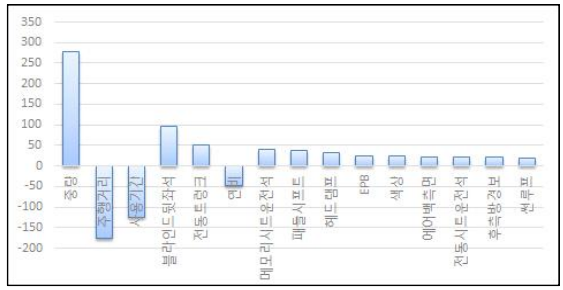

Fig. 5-8 Standardized Regression Coefficient of the Lasso in SSANGYONG

차량 제원 변수의 경우 중량과 색상 등이 중고차 가격에 양의 방향으로, 주행거리, 사용기간, 그리고 연비가 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다. 차량 옵션 변수의 경우 블라인드뒷자석, 전동트렁크, 메모리시트운전석, 패들시프트, 헤드램프의 수준, 전자식파킹브레이크, 에어백측면, 전동시트운전석, 후측방경보, 그리고 썬루프 유무 등이 중고차 가격에 양의 방향으로 영향을 미치는 것으로 확인되었다.

Lasso 회귀 모델에서 결과값에 미치는 영향력이 0으로 나타난 변수는 배기량, 전동시트동승석, 무선도어잠금, 메모리시트동승석, ,블라인드후방, 전면AV모니터, 통풍시트운전석, 전동시트뒷좌석, 통풍시트뒷좌석, 블루투스, USB단자, ECS, 마사지시트, 에어백운전석, 토크, 마력, 루프랙, 구동방식, 파워도어록, 에어백커튼, 그리고 TPMS의 유무로 확인되었다. 이에 따라 이들 변수를 소거하여 트리 기반 머신러닝 모델을 실행하였다.

트리 기반 머신러닝 모델의 실행에 앞서 이들 자료의 시험 세트의 비율을 1%로 설정한 후 CART 모델의 K-Fold 교차검증을 실행하여 k가 6에서 8일 때 가장 우수한 것을 확인하였으며, 이에 따라 모든 트

리 기반 머신러닝 모델의 k를 8로 고정하였다. 다음으로 Hyper Parameter Tuning을 실행하여 최적의 초매개변수를 확인한 후 이를 활용하여 최적의 머신러닝 모델을 도출하였다.

쌍용 브랜드에서 트리 기반 머신러닝 모델 간 비용함수와 R^2이 가장 우수한 모델은 XGBR 모델로 확인되었으며, RMSE가 79.830, MSE가 6372.880, MAE가 67.087, MAPE가 3.631(%), 그리고 R^2이 0.985로 나타났다.

<Fig. 5-9>를 통해 쌍용 브랜드에 대한 XGBR 모델의 실제값(선)과 예측값(점) 간 오차가 비교적 고르게 분포하며, 오차가 크지 않다고 판단된다.

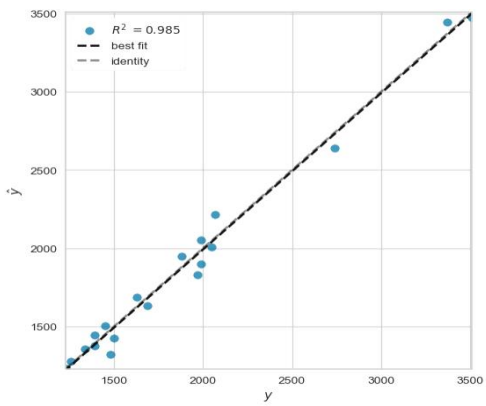

Fig. 5-9 Visualization of XGBR in SSANGYONG

5.5 제네시스

제네시스 브랜드의 차량 크기는 평균이 4.901, 표준편차가 0.571로 준대형차에 가까운 것으로 나타났다. 또한 차량의 마력, 배기량, 그리고 중량 역시 차량 크기와 동일한 특성으로 확인되었다.

자료의 특성을 확인한 후 훈련 세트의 적정 비율을 확인하고자 훈련 세트와 시험 세트 간 Score를 비교하였고, 그 결과 두 세트 간 차이가 가장 작게 나타난 비율은 1%인 것으로 나타났다. 다음으로 Lasso 회귀 모델에 대한 K-Fold 교차검증을 실행하여 k는 4일 때 가장 우수한 것으로 확인되었다.

위와 같은 과정을 거쳐 최종적으로 Lasso 회귀 모델을 실행하였다. 그 결과 RMSE가 382.341, MSE가 146184.311, MAE가 301.287, MAPE가 7.295, 그리고 R^2이 0.950으로 나타났다.

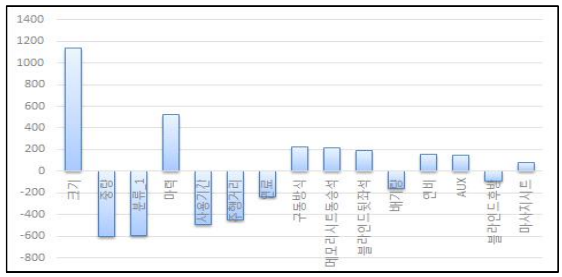

Fig. 5-10 Standardized Regression Coefficient of the Lasso in GENESIS

차량 제원 변수의 경우 크기, 마력, 구동방식, 그리고 연비 등이 중고차 가격에 양의 방향으로, 중량, 차종(세단 여부), 사용기간, 주행거리, 연료(디젤 여부), 그리고 배기량 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다. 차량 옵션 변수의 경우 메모리시트동승석, 블라인드 뒷자석, AUX, 그리고 마사지 시트의 유무 등이 중고차 가격에 양의 방향으로, 블라인드후방 유무 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다.

Lasso 회귀 모델에서 결과값에 미치는 영향력이 0으로 나타난 변수는 usb단자, 전동시트동승석, 열선시트앞좌석, 무선도어잠금, 주차감지센서후방, 후방카메라, Smartkey, Autoair, 오토라이트, 전동시트운전석, 가죽시트, 블루투스, ESC, TPMS, 알루미늄휠, 스티어링리모콘, ECM, 하이패스, 파워도어록, 파워스티어링, 파워윈도우, 에어백운전석, 에어백동승석, 에어백측면, 에어백커튼, ABS, TCS, 그리고 전동사이드미러의 유무로 확인되었다. 이에 따라 이들 변수를 소거하여 트리 기반 머신러닝 모델을 실행하였다.

트리 기반 머신러닝 모델의 실행에 앞서 이들 자료의 시험 세트의 비율을 1%로 설정한 후 CART 모델의 K-Fold 교차검증을 실행하여 k(CV)가 8일 때 가장 우수한 것을 확인하였으며, 이에 따라 모든 트리 기반 머신러닝 모델의 k를 8로 고정하였다. 다음으로 Hyper Parameter Tuning을 실행하여 최적의 초매개변수를 확인한 후 이를 활용하여 최적의 머신러닝 모델을 도출하였다.

제네시스 브랜드에서 트리 기반 머신러닝 모델 간 비용함수와 R^2이 가장 우수한 모델은 RFR 모델로 확인되었으며, RMSE가 206.931, MSE가 42820.544, MAE가 158.138, MAPE가 3.902(%), 그리고 R^2이 0.979로 나타났다.

<Fig. 5-11>을 통해 제네시스 브랜드에 대한 RFR 모델의 실제값(선)과 예측값(점) 간 오차가 비교적 고르게 분포하며, 오차가 크지 않다고 판단된다.

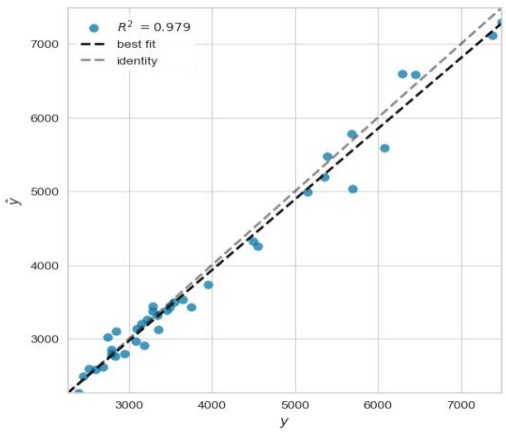

Fig. 5-11 Visualization of RFR in GENESIS

5.6 현대

현대 브랜드의 차량 크기는 평균이 3.901, 표준편차가 1.338로 중형차에 가까운 것으로 나타났다. 또한 차량의 마력, 배기량, 그리고 중량 역시 차량 크기와 동일한 특성으로 확인되었다.

자료의 특성을 확인한 후 훈련 세트의 적정 비율을 확인하고자 훈련 세트와 시험 세트 간 Score를 비교하였고, 그 결과 두 세트 간 차이가 가장 작게 나타난 비율은 1%인 것으로 나타났다. 다음으로 Lasso 회귀 모델에 대한 K-Fold 교차검증을 실행하여 k는 4일 때 가장 우수한 것으로 확인되었다.

위와 같은 과정을 거쳐 최종적으로 Lasso 회귀 모델을 실행하였다. 그 결과 RMSE가 169.429, MSE가 28706.274, MAE가 134.888, MAPE가 6.539, 그리고 R^2이 0.943으로 나타났다.

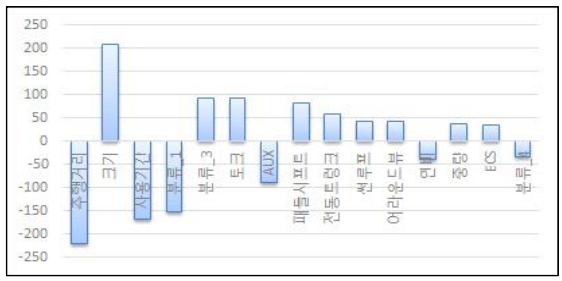

Fig. 5-12 Standardized Regression Coefficient of the Lasso in HYUNDAI

차량 제원 변수의 경우 크기(차량), 마력, 차종(Hybrid 여부), 토크, 그리고 중량 등이 중고차 가격에 양의 방향으로, 주행거리, 사용기간, 분류_1(세단 여부), 연비, 그리고 분류_4(경차 여부) 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다. 차량 옵션 변수의 경우 패들시프트, 전동트렁크, 썬루프, 어라운드뷰, 그리고 전자제어현가장치의 유무 등이 중고차 가격에 양의 방향으로, AUX 유무 등이 중고차 가격에 음의 방향으로 영향을 미치는 것으로 확인되었다.

Lasso 회귀 모델에서 결과값에 미치는 영향력이 0으로 나타난 변수는 오토라이트, 블루투스, 통풍시트 운전석, 마사지시트, 전동사이드미러, 파워도어록, 마력, TCS, 에어백측면, 열선스티어링, 에어백동승석, 에어백운전석, 주차감지센서후방, 그리고 파워윈도우

의 유무로 확인되었다. 이에 따라 이들 변수를 소거하여 트리 기반 머신러닝 모델을 실행하였다.

트리 기반 머신러닝 모델의 실행에 앞서 이들 자료의 시험 세트의 비율을 1%로 설정한 후 CART 모델의 K-Fold 교차검증을 실행하여 k가 8일 때 가장 우수한 것을 확인하였으며, 이에 따라 모든 트리 기반 머신러닝 모델의 k를 8로 고정하였다. 다음으로 Hyper Parameter Tuning을 실행하여 최적의 초매개변수를 확인한 후 이를 활용하여 최적의 머신러닝 모델을 도출하였다.

현대 브랜드에서 트리 기반 머신러닝 모델 간 비용함수와 R^2이 가장 우수한 모델은 LGBR 모델로 확인되었으며, RMSE가 112.861, MSE가 12737.508, MAE가 83.491, MAPE가 3.742(%), 그리고 R^2이 0.979로 나타났다.

<Fig. 5-13>을 통해 현대 브랜드에 대한 LGBR 모델의 실제값(선)과 예측값(점) 간 오차가 비교적 고르게 분포하며, 오차가 크지 않다고 판단된다.

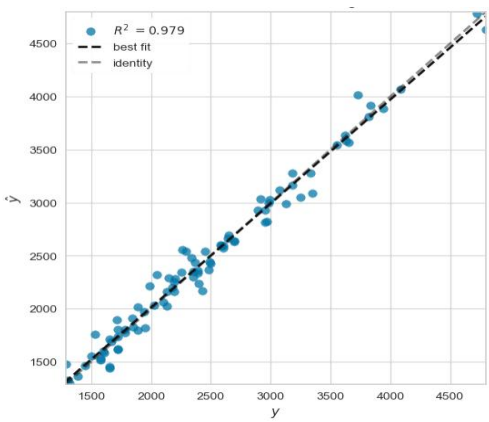

Fig. 5-13 Visualization of LGBR in HYUNDAI

6. 결론

본 연구는 국내 온라인 중고차 플랫폼 서비스를 통해 국내 6개 브랜드의 차량 정보를 획득하였다. 또한 획득한 자료를 기반으로 두 유형의 머신러닝 모델을 순차적으로 실행한 방법론을 활용하였다. 우선 Lasso 회귀 머신러닝 모델을 실행하여 중고차 가격 예측을 위한 변수들의 실질적 영향력과 방향성을 확인하고 영향력이 0인 변수들을 도출하였다. 다음으로 영향력이 0인 변수를 소거한 자료를 대상으로 트리(Decision Tree) 기반 머신러닝 모델을 실행하여 가장 우수한 모델을 선정하였다.

Lasso 회귀 모델의 실행 결과 브랜드별 중고차 가격에 미치는 변수의 영향력과 순위는 차이가 존재하였으며, 기아, 제네시스, 그리고 현대 브랜드는 차종과 차량 크기가 중고차 가격에 미치는 영향력이 타 브랜드보다 더욱 큰 것으로 나타났다.

트리 기반 머신러닝 모델의 실행 결과 브랜드별 중고차 가격 예측을 위한 최우수 모델은 기아의 경우 LGBR 모델, 르노의 경우 XGBR 모델, 쉐보레의 경우 RFR 모델, 쌍용의 경우 XGBR 모델, 제네시스의 경우 RFR 모델, 그리고 현대의 경우 LGBR 모델로 확인되었다. 특히 쌍용, 제네시스, 그리고 현대의 MAPE는 3%대로 나타나 본 연구의 목적 중 하나인 브랜드별 트리 기반 머신러닝 모델 실행 결과의 우수성을 확인하였다. 또한 머신러닝 모델을 활용한 중고차 가격 예측의 선행연구는 대부분 중고차 가격을 서열척도로 구성하였으나, 본 연구는 중고차 가격을 연속형으로 측정하여 척도 내 오차를 해결하고자 하였다.

본 연구의 시사점으로 첫째, 회귀 기반 머신러닝 모델인 Lasso 회귀 모델을 통해 변수의 실질적인 영향력과 방향성을 확인하고 영향력이 0인 변수를 소거하였다. 이후 회귀 기반 머신러닝 모델보다 비용함수 수치가 우수한 트리 기반 머신러닝 모델을 실행하여 가장 우수한 모델을 선정하였다. 이를 통해 두 유형의 머신러닝 모델의 장점을 최대한 활용하고자 하였다. 둘째, 브랜드별 특성을 확인할 수 있었다. 대표적으로 제네시스의 경우 고급 브랜드의 특성상 차량의 옵션 변수 중 28개의 옵션이 기본 제공

됨으로써 Lasso 회귀 모델의 분석 결과 이들 변수의 영향력이 0으로 나타났다. 또한 기아와 현대의 경우 차종(세단, SUV, Hybrid, 경차)과 차량 크기(경형 ~ 대형)가 타 브랜드에 비해 다양했으며, 이들 변수는 중고차 가격에 미치는 영향력이 타 브랜드에 비해 상대적으로 컸다. 이와 같은 과정을 통해 중고차 온라인 플랫폼 서비스의 고객인 구매자와 판매자가 차량 제원과 차량 옵션 여부를 기반으로 전반적인 중고차 가격 예측을 지원하고자 하였다. 또한 중고차 가격 예측 모델을 통해 중고차 매매관계자 간 정보의 비대칭성으로 인한 문제 해결에 도움이 되길 희망한다.

본 연구는 회귀 기반 머신러닝 모델을 활용하여 변수의 영향력과 방향성을 확인한 후 영향력이 0으로 나타난 변수를 소거하여 트리 기반 머신러닝 모델을 실행하였다. 그러나 이러한 과정은 두 유형의 머신러닝 모델 간 서로 연계되지 않는다는 문제점이 제기될 수 있다. 이에 따라 향후 연구 방향으로 트리 기반 머신러닝 모델을 실행한 후 SHAP Value의 시각화를 통해 변수의 실질적 영향력을 확인할 필요성이 있다.

본 연구의 한계점은 다음과 같다. 첫째, 판매자와 구매자 간 실제 거래가격은 중고차 온라인 플랫폼 서비스에 개시된 가격과 다를 수 있으나, 구매자의 입장에서 개시된 가격은 거래를 위한 시작점인 동시에 거래를 위한 기준이 될 수 있다. 이에 따라 개시된 가격은 실질적 유통가격으로 판단할 수 있으나, 실제 거래가격은 아니라는 한계점이 존재한다. 둘째, 정보의 비대칭으로 인해 매매관계자 간 신뢰수준이 낮다는 것이다. 관련 보고서에 따르면 차량 상태에 대한 소비자의 불신이 이에 대한 대표적인 원인으로 지목되었다(Kim, 2022). 이에 따라 차량 상태의 투명성이 확보되어 차량의 정비 내역과 사고 내역을 분석에 활용할 수 있다면 기존보다 비용함수 수치가 더욱 우수한 중고차 가격 예측을 위한 머신러닝 모델의 선정을 기대할 수 있다. 셋째, 일부 브랜드에서 차종 및 차량 크기가 타 변수에 비해 영향력이 큰 것으로 확인되었다. 이에 따라 차종 및 차량 크기에 따른 머신러닝 모델의 실행 시 기존보다 더욱 우수한 머신러닝 모델이 선정될 가능성이 있다.

[References]

[1] Arawomo, D. F., and Osigwe, A. C. (2016), Nexus of fuel consumption, car features and car prices: Evidence from major institutions in Ibadan, *Renewable and Sustainable Energy Reviews*, 59, 1220-1228

[2] BALCE, A. O. (2016), Factors Affecting Prices in an Used Car E-Market, *Journal of Internet Applications and Management*, 7(2), 5-20

[3] Chaudhary, L., Sharma, S., and Sajwan, M. (2022), Comparative Analysis of Supervised Machine Learning Algorithm, *Available at SSRN 4143890*.

[4] Chen, C., Hao, L., and Xu, C. (2017), Comparative analysis of used car price evaluation models, *In AIP Conference Proceedings*, 1839(1), 020165

[5] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., and Kevric, J. (2019), Car price prediction using machine learning techniques, TEM Journal. 8(1), 113.

[6] Khan, Z. (2022), Used Car Price Evaluation using three Different Variants of Linear

Regression, *International Journal of Computational and Innovative Sciences*, 1(1)

[7] Kim, C. (2022), The Problems of the Used Car Market in Korea and the Entry of Large Enterprises, *KIET Industrial Economic Policy and Issues*, 65-69 (김주홍 (2022), 우리나라 중고차 시장의 문제점과 대기업 진출, *월간 KIET 산업경제 정책과 이슈*, 65-69)

[8] Kim, H., and Lee, M. (2021), Research Trends in Machine Learning Applications in Marketing, *Korean Journal of Marketing*, Vol.36, No.1, pp.1-25 (김혜진, 이명구 (2021), 마케팅 분야의 머신러닝 연구 동향 분석, *마케팅연구*, 36(1), 1-25)

[9] Kwon, C. (2022), Python Machine Learning Complete Guide, Wikibooks (권철민 (2022), 파이썬 머신러닝 완벽 가이드, 위키북스.)

[10] Lasek, A., and Wyszyński, K. (2019), Determinants of used car prices-The Black Volkswagen Golf case, *University of Warsaw Faculty of Economic Science*, Warsaw, January 2019, 1-16

[11] Lee, J. (2022), Discover the potential of the domestic used car market, *Eugene Investment & Securities Automotive Issues*, 1-36 (이재일 (2022), 국내 중고차 시장의 잠재력 알아보기, *유진투자증권 자동차 이슈*, 1-35)

[12] Lee, D., and Shin, M. (2023), Prediction of Dormant Customer in the Card Industry, *Journal of Service Research and Studies,* 13(2), 99-113 (이동규, 신민수(2023), 카드 산업에서 휴면 고객 예측, *서비스 연구*, 13(2), 99-113)

[13] Lee, S. (2022), 2022 Copyright industry issues in the next-generation digital enviroment, Trade and Industry Statistics Team KOREA COPYRIGHT COMMISSION, 14-37 (이수현 (2022), 2022 차세대 디지털 환경에서의 저작권 산업 이슈, 한국저작권위원회 통상산업통계팀, 14-37)

[14] Meng, S. M., Liu, L. J., Kuritsyn, M., and Pechnikov, V. (2019), Price Determinants on Used Car Auction in Taiwan, I*nternational Journal of Asian Social Science*, 9(1), 48-58

[15] Nasiboglu, R., and Akdogan, A. (2020), Estimation of the second hand car prices from data eXtreamcted via web scraping techniques, *Journal of Modern Technology and Engineering,* 5(2), 157-166

[16] Ramampiandra, E. C., Scheidegger, A., Wydler, J., and Schuwirth, N. (2023), A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation, *Ecological Modeling*, 481, 110353

[17] Samuel, A. L. (1959), Machine learning, T*he Technology Review*, 62(1), 42-45

[18] Shanti, N., Assi, A., Shakhshir, H., and Salman, A. (2021), Machine Learning

Powered Mobile App for Predicting Used Car Prices, *In 2021 3rd International. Conference on Big-data Service and Intelligent Computation*, Xiamen, China, November 19-21, 2021, 52-60

[19] Staartjes, V. E., Regli, L., and Serra, C. (Eds.). (2022), Machine learning in clinical neuroscience: Foundations and application, Springer.

[20] Venkatasubbu, P., and Ganesh, M. (2019), Used Cars Price Prediction using Supervised Learning Techniques, *Int. J. Eng. Adv. Technol(IJEAT)*, 9(1S3)

[21] Yadav, A., Kumar, E., and Yadav, P. K. (2021), Object detection and used car price predicting analysis system (UCPAS) using machine learning technique, Linguistics and Culture Review. 5(S2), pp.1131-1147.

[22] Yang, J. (2020), Current Status and Challenges of the Domestic Market for Used Cars, *Industrial trends*, 49 (양재완 (2020), 중고차 내수 시장의 현황과 과제, *산업동향*, 49)

[23] Yennimar, Y., Kelvin, K., Suwandi, S., and Amir, A. (2022), Comparison Analysis of SVM Algorithm with Linear Regression in Predicting used Car Prices, *Journal Mantik.* 5(4), 2720-2028

Yim, Seungjun (yimdgz@hotmail.com)

Seungjun Yim received his Ph.D from the Department of Business Administration(Operations Research & Operations Management) at Hongik University. His current research interests include Service Quality, Product Service-System Design(PSS), and Big Data/Machine Learning.

Lee, Joungho (jholee@konkuk.ac.kr)

Joungho Lee is a professor of the School of Business at Konkuk University. He received his Ph.D. from the Department of Business Administration at Hongik University. His current research interests include Supply Chain Management and Business Analytics.

Ryu, Choonho (ryuch@hongik.ac.kr)

Choonho Ryu is a professor at the Business School of Hongik University. He received his Ph.D. from the Operations and Information Management Department of the Wharton School, University of Pennsylvania. His current research interests include operations research(combinatorial optimization), productivity and teacher evaluation.

A Study on the Prediction Models of Used Car Prices for Domestic Brands Using Machine Learning

Seungjun Yim*, Joungho Lee**, Choonho Ryu***

ABSTRACT

The domestic used car market continues to grow along with the used car online platform service. The used car online platform service discloses vehicle specifications, accident history, inspection history, and detailed options to service consumers.

Most of the preceding studies were predictions of used car prices using vehicle specifications and some options for vehicles. As a result of the study, it was confirmed that there was a nonlinear relationship between used car prices and some specification variables. Accordingly, the researchers tried to solve the nonlinear problem by executing a Machine Learning model. In common, the Regression based Machine Learning model had the advantage of knowing the actual influence and direction of variables, but there was a disadvantage of low Cost Function figures compared to the Decision Tree based Machine Learning model.

This study attempted to predict used car prices of six domestic brands by utilizing both vehicle specifications and vehicle options. Through this, we tried to collect the advantages of the two types of Machine Learning models. To this end, we sequentially conducted a regression based Machine Learning model and a decision tree based Machine Learning model. As a result of the analysis, the practical influence and direction of each brand variable, and the best tree based Machine Learning model were selected.

The implications of this study are as follows. It will help buyers and sellers who use used car online platform services to predict approximate used car prices. And it is hoped that it will help solve the problem caused by information inequality among users of the used car online platform service.

Keywords: Used Car Online Platform Service, Used Car Price, Brand, Lasso Regression Machine Learning, Decision Tree Based Machine Learning

* First Author, Ph.D, Department of Business Administration, Hong-ik University
** Corresponding Author, Assistant Professor, School of Business, Konkuk University
*** Co-Author, Professor, Department of Business Administration, Hong-Ik University