앙상블 모델과 SHAP Value를 활용한 국내 중고차 가격 예측 모델에 관한 연구: 차종 특성을 중심으로

임 승 준*, 이 정 호**, 류 춘 호***

목 차

요약

중고차 시장에서 온라인 플랫폼 서비스의 시장 점유율은 지속적으로 증가하고 있다. 또한 중고차 온라인 플랫폼 서비스는 서비스 이용자에게 차량의 제원, 사고 이력, 점검 내역, 세부 옵션, 그리고 중고차의 가격 등을 공개하고 있다. 2023년 현재 국내 자동차 시장에서 SUV 차종의 신차 점유율은 50% 이상으로 확대되었으며, 하이브리드 차종은 신차 판매량이 지난해에 비해 두 배 이상 증가하였다. 이에 따라 이들 차종은 국내 중고차 시장에서도 인기를 끌고 있다. 기존 연구는 전체 차량 또는 브랜드별 차량을 대상으로 머신러닝 모델을 실행하여 중고차 가격 예측 모델을 제안하였다. 반면 국내 자동차 시장에서 SUV와 하이브리드 차종의 인기는 매년 상승하고 있으나, 이들 차종을 대상으로 중고차 가격 예측 모델을 제안한 연구는 찾기 어려웠다.

본 연구는 국내 시장에서 자국 브랜드가 생산한 세단, SUV, 그리고 하이브리드 차종을 대상으로 차량 제원과 옵션, 총 72개의 특성을 활용하여 이들 차종별 가장 우수한 중고차 가격 예측 모델을 선정하였다. 이를 위해 특성 선택으로 Lasso 회귀 모델을 활용하여 특성을 선별한 후 동일 샘플링으로 앙상블 모델을 실행하였다. 그 결과 모든 차종에서 최우수 모델은 CBR 모델로 선정되었으며, 차종별 최우수 모델을 대상으로 Tree SHAP Value의 시각화를 실행하여 특성의 기여도 및 방향성을 확인하였다. 본 연구의 시사점으로 온라인 플랫폼 서비스를 이용하는 매매관계자에게 차종별 중고차 가격 예측 모델을 제안하고 특성의 기여 수준과 방향성을 확인함으로써 이들 간 정보의 비대칭으로 야기된 문제 해결에 지원이 될 것으로 기대한다.

표제어: 중고차 온라인 플랫폼 서비스, 중고차 가격, 차종, 앙상블 모델, SHAP Value

접수일(2024년 01월 09일), 수정일(2024년 02월 27일), 게재확정일(2024년 03월 07일)

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

1. 서론

중고차 온라인 플랫폼 서비스는 국내와 해외 중고차 시장을 막론하고 지속적으로 성장하고 있다. 2021년을 기준으로 국내 중고차 온라인 플랫폼 서비스를 통한 중고차 거래량은 오프라인 중고차 거래량에 근접하는 수준으로 확인되었다(Lee, 2022). 중고차 온라인 플랫폼 서비스는 서비스 이용자에게 차량의 제원, 사고 이력, 정비 이력, 그리고 차량 세부 옵션까지 자세한 정보를 제공하고 있다.

머신러닝(Machine Learning)은 컴퓨터가 명시적으로 프로그래밍 되지 않고 학습할 수 있는 능력을 주는 것으로 정의된다(Samuel, 1959). 머신러닝은 2010년 이후 빅데이터의 출현과 이를 분석하기 위한 알고리즘, 이를 실현하기 위한 컴퓨터 성능의 향상으로 지속적으로 발전하고 있다(Lee, 2022). 현재 머신러닝은 공학에 그치지 않고 사회과학에서도 활용하기 시작하였다(Lee and Shin, 2023; Yim et al., 2023).

2020년대 이후 국내·외 연구에서 중고차 가격 예측 모델의 제안을 위해 머신러닝 모델을 사용하기 시작하였다. 이 연구들은 공통적으로 차량의 제원 특성(Feature)과 차량의 일부 옵션 특성을 토대로 다양한 머신러닝 및 딥러닝 모델을 실행하여 중고차 가격 예측 모델을 제안하였다. 그 결과 대부분의 연구에서 앙상블 모델이 타 모델에 비해 비용함수 수치가 상대적으로 우수하다는 것을 확인하였다(Chaudhary et al., 2022; Gegic et al., 2019; Shanti et al., 2021; Das Adhikary et al., 2022).

국내 자동차 시장에서 차종 중 SUV 차량과 하이브리드 차량의 성장세는 지속되고 있다. SUV 차량은 국내 시장에서 점유율이 매년 증가하고 있으며, 2021년 기준 54%의 점유율을 기록하여 세단의 점유율을 제쳤다(Kim, 2023). 또한 하이브리드 차량은 국내 시장에서 2023년 1분기 판매량이 109,371대로 지난해 동 분기 판매량 50,363대 대비 두 배 이상 증가한 것으로 나타났다(Lim, 2023).

대부분의 선행연구에서 차종과 상관없이 전체 차량 또는 브랜드별 차량을 대상으로 중고차의 가격을 예측하였다. 대표적으로 Yim et al.(2023)은 브랜드별 중고차 가격 예측 모델을 제안하였고, 이를 위해 브랜드별 Lasso 회귀 모델을 실행하여 특성의 영향력과 방향성을 확인한 후 브랜드별 트리 기반(Decision Tree) 머신러닝 모델을 실행하였다. 그 결과 쌍용을 제외한 브랜드에서 차종이 중고차 가격에 영향을 미친다는 것을 확인하였다. 그러나 이 연구는 두 유형의 머신러닝 모델의 순차적 실행에서 동일 샘플링이 아니라는 한계점이 존재하였다.

본 연구는 국내 시장에서 SUV와 하이브리드 차량의 점유율이 지속적으로 상승함에 따라 국내 브랜드의 차량 정보를 차종으로 구분하여 차종별 특성의 기여 수준을 확인하였다. 또한 머신러닝 모델의 실행을 위한 특성 선별 과정인 특성 공학(Feature Engineering) 과정에서 트리 기반 머신러닝 모델을 활용한 특성 선택(Feature Selection)은 특성 간 상관관계를 내포한 특성의 중요도가 실제보다 더 낮거나 높게 나타난 단점이 있었다(Géron, 2022). 이에 따라 차종별 특성 선택을 위한 과정으로 Lasso 회귀 모델을 활용하였으며, 차종별 동일 샘플링으로 앙상블(Ensemble) 모델을 실행하여 최우수 모델을 선정하였다. 마지막으로 차종별 최우수 모델을 대상으로 SHAP Value의 시각화를 실행하였다.

본 연구의 목표는 우선 국내 브랜드를 대상으로 차종별 중고차 가격 예측 모델을 제안하는 것이다. 다음으로 차종별 중고차 가격 예측을 위한 특성의 기여 수준과 방향성을 확인하는 것이다. 본 연구 결과를 통해 중고차 온라인 플랫폼 서비스를 이용하는 매매관계자에게 차종별 제원과 옵션의 수준을 기반으로 더욱 정확한 중고차 가격 예측에 도움이 되고자 하며, 이들 간 정보의 비대칭성으로 인해 발생한 문제 해결에 지원이 되고자 한다.

2. 선행연구

2.1 국내 중고차 시장

2018년에서 2020년까지 3년 동안 중고차의 거래량은 증가하는 추세에 있었다. 그러나 2021년 중고차 거래량은 387만대로 2020년의 380만대와 큰 차이가 없는 것으로 나타났다. 또한 2021년 중고차 거래량은 신차 판매량(167만 대)의 2.32배로 나타나 2020년의 2.06배보다 상승한 것으로 나타났다. 이에 따라 국내의 신차 판매량 대비 중고차 판매량의 비율은 자동차 선진국인 미국의 비율인 2.4배와 비슷한 수치로 나타나 국내 중고차 시장은 성숙기에 돌입했다고 볼 수 있다(Lee, 2022).

국내와 해외를 막론하고 중고차 온라인 플랫폼 서비스를 통한 중고차 거래량은 매년 증가하고 있다. 미국 시장의 중고차 온라인 플랫폼 서비스는 대표적으로 Carvana, Vroom, 그리고 Carmax 등이 있으며, 국내 시장의 중고차 온라인 플랫폼 서비스는 대표적으로 엔카, KB차차차, 그리고 K-Car 등이 있다. 국내 중고차 시장의 경우 온라인 플랫폼 서비스를 활용한 중고차 거래량은 2021년 4분기 기준으로 전체 중고차 거래량 대비 38.3%로 나타나 오프라인 중고차 거래량 41.0%에 근접한 것으로 확인되었다(이창희, 2022). 국내의 중고차 온라인 플랫폼 서비스는 차량 제원인 차량의 연식, 사용기간, 주행거리, 배기량, 그리고 마력 등의 정보와 차량 옵션인 내·외장, 편의·멀티미디어, 안전, 그리고 시트 구성의 정보를 중고차 온라인 플랫폼 서비스 이용자에게 제공하고 있다(Yim et al., 2023).

2.2 머신러닝 모델을 활용한 중고차 가격 예측

Chaudhary et al.(2022)은 인도 중고차 시장에서 차량의 제원 특성(Feature)을 활용하여 중고차 가격 예측 모델을 제안하였다. 이를 위해 선형 회귀 모델, CART(Classification and Regression Tree) 모델, Lasso 회귀 모델, Ridge 회귀 모델, 그리고 RF(Random Forest) 모델을 실행하였으며, 그 결과 RF 모델이 최우수 모델로 선정되었다. Gegic et al.(2019)은 보스니아 중고차 시장에서 차량의 제원, 시트와 크루즈 컨트롤 옵션 특성을 활용하여 중고차 가격 예측 모델을 제안하였다. 이를 위해 SVM(Support Vector Machine), ANN(Artificial Neural Network), 그리고 RF 모델을 실행하였으며, 그 결과 RF 모델이 가장 우수한 모델로 선정되었다.

Shanti et al.(2021)은 아랍 중고차 시장에서 차량의 제원, 편의 장비, 썬루프, 에어백, 시트 그리고 마그네슘휠 옵션의 특성을 활용하여 중고차 가격 예측 모델을 제안하였다. 이를 위해 ANN 모델, RF 모델, GB(Gradient Boosting) 모델, 그리고 SVM 모델을 실행하였으며, 그 결과 GB 모델이 가장 우수한 모델로 선정되었다. Huang et al.(2022)은 중국 중고차 시장에서 차량의 제원, 신차 가격, 시트, 그리고 계기판 옵션의 특성을 활용하여 중고차 가격 예측 모델을 제안하였다. 이를 위해 CB(Categorical Boosting) 모델, XGB(eXtream Gradient Boosting) 모델, 그리고 LGB(Light Gradient Boosting) 모델을 실행하였으며, 그 결과 XGB 모델이 가장 우수한 모델로 선정되었다.

Nasiboglu and Akdogan(2020)은 Kaggle의 자료를 대상으로 차량의 제원 특성을 활용하여 브랜드별 중고차 가격 예측 모델을 제안하였다. 이를 위해 선형 회귀 모델, Lasso 회귀 모델, Elastic.Net 회귀 모델, Ridge 회귀 모델, RF 모델, XGB 모델, 그리고 KNN(K Nearest Neighbor) 모델을 실행하였으며, 브랜드별 가장 우수한 모델을 선정하였다. Yim et al.(2023)은 한국 중고차 시장에서 차량 제원과 옵션 특성 72개를 활용하여 브랜드별 중고차 가격 예측 모델을 제안하였다. Lasso 회귀 모델을

통해 특성을 선택한 후 CART 모델, RF 모델, XGBR 모델, 그리고 LGBR 모델을 실행하여 브랜드별 최우수 모델을 선정하였다. 그러나 더미 변수의 문제로 Lasso 회귀 모델과 트리 기반 머신러닝 모델 간 서로 연계되지 못한 한계점이 있었다.

Das Adhikary et al.(2022)은 인도 중고차 시장에서 차량 제원과 시트 옵션 특성을 활용하여 중고차 가격 예측 모델을 제안하였다. 이를 위해 KNN 모델, RF 모델, 그리고 LGB 모델을 실행하였으며, 그 결과 가장 우수한 모델로 LGB 모델이 선정되었다. Wang et al.(2022)은 중국 중고차 시장에서 차량의 제원, 차량의 전 주인 수, 그리고 색상 옵션 특성을 활용하여 중고차 가격 예측 모델을 제안하였다. 이을 위해 CART 모델, AB(Adaptive Boosting) 모델, CB 모델, GB 모델, LGB 모델, 그리고 XGB 모델을 실행하였으며, 그 결과 가장 우수한 모델로 LGB 모델이 선정되었다.

중고차 가격 예측이 주제인 대부분의 선행연구는 첫째, 차량의 제원 특성과 일부 옵션 특성만을 활용한 한계점이 존재하였다. 둘째, 차량 전체 또는 브랜드를 대상으로 머신러닝 모델을 실행하였고, 차종(세단, SUV, 하이브리드)을 고려하지 못하였다. 이에 따라 차종별 중고차 가격 예측에 있어 특성의 중요도를 알 수 없었다. 셋째, 결과값인 중고차 가격을 서열척도로 설정하여 척도 내 오차가 존재하였다.

본 연구는 최근 일부 차종의 인기를 반영하여 차종을 세단, SUV, 그리고 하이브리드로 구분하였다. 이후 차종별로 차량의 제원과 옵션 특성 72개를 활용하여 Lasso 회귀 모델을 통해 특성 선택을 실행한 후 영향력이 0으로 나온 특성을 소거하였다(범주형 특성 제외). 다음으로 동일 샘플링으로 앙상블 모델을 실행하여 차종별 최우수 모델을 선정하였다. 마지막으로 최우수 모델을 대상으로 SHAP Value의 시각화를 통해 특성의 기여도와 방향성을 확인하였다. 또한 본 연구는 차량 가격을 연속형으로 구성하여 선행연구의 척도 내 오차를 해결하고자 하였다.

3. 자료 수집 및 특성 설정

3.1 자료 수집

본 연구는 중고차 온라인 플랫폼 서비스 중 국내 시장 점유율이 가장 높은 엔카(encar.com)의 중고차 정보를 활용하였다. 또한 본 연구의 분석 대상은 국내 시장에서 자국 브랜드의 중고차이며, 현세대와 직전 세대의 차량이다. 이에 따라 본 연구는 엔카의 차량 정보(제원, 옵션, 가격)를 대상으로 크롤링(Crawling)을 실행하여 이들 정보를 획득하였다.

3.2 특성의 정의와 측정

특성(Feature)은 머신러닝이 패턴을 학습한 후 예측 또는 분류를 수행하는 데 필요한 입력 자료로 정의된다(Sikora, 2015). 즉 특성은 독립변수를, 결과값은 종속변수를 의미한다.

본 연구에서 차량의 제원 특성은 중고차 가격 예측을 위한 선행연구를 참조하여 브랜드(1: 제네시스, 2: 현대, 3: 기아, 4: 쉐보레, 5: 르노, 6: 쌍용), 마력, 토크, 연비, 연료(1: 디젤, 2: 가솔린), 구동방식(1: 전륜, 2: 후륜, 3: 4륜), 공차중량, 사용기간, 주행거리, 배기량, 그리고 색상으로 구성하였다.

본 연구에서 차량의 옵션 특성은 Gegic et al.(2019), Shanti et al.(2021), Huang et al.(2022), Das Adhikary et al.(2022), 그리고 Yim et al.(2023)의 선행연구를 참고하였으며, 네 개의 유형으로 구분하였다. 첫째, 내·외관 관련 옵션은 썬루프, 헤드램프, 전동 사이드미러, 전동 트렁크, 고스트도어 클로징, 알루미늄휠, 루프랙, 열선 스티어링, 전동 스티어링, 파워 스티어링, 패들시프트, ECM 룸미러, 하이패스, 스티어링 리모콘, 파워 도어락, 그리고 파워 윈도우의 유무(무: 0, 유: 1)로 구성하였다. 둘째, 안전 관련 옵션은 에어백(운전

석), 에어백(동승석), 에어백(커튼형), 에어백(측면), 브레이크 잠금장치, 미끄럼방지 장치, 타이어 공기압 센서, 차체 자세 제어장치, 전자제어 서스펜션, 차선 이탈 경보, 주차 감지 센서(전방), 주차 감지 센서(후방), 후방카메라, 후측방 경보, 그리고 어라운드 뷰의 유무(무: 0, 유: 1)로 구성하였다. 셋째, 편의 및 멀티미디어 관련 옵션은 크루즈 컨트롤, 전자식 주차브레이크, 헤드업 디스플레이, 자동 에어컨, 비 센서, 오토 라이트, 블라인드(뒷좌석), 블라인드(후방), 스마트키, 무선 도어잠금, 내비게이션, 전면 AV 모니터, 후면 AV 모니터, USB 단자, 블루투스, CD 플레이어, 그리고 AUX 단자의 유무(무: 0, 유: 1)로 구성하였다. 넷째, 시트 관련 옵션은 전동시트(운전석), 전동시트(동승석), 전동시트(뒷좌석), 가죽시트, 열선시트(앞좌석), 열선시트(뒷좌석), 메모리 시트(운전석), 메모리 시트(동승석), 통풍 시트(운전석), 통풍 시트(동승석), 그리고 통풍 시트(뒷좌석)의 유무(무: 0, 유: 1)로 구성하였다.

차량의 옵션 특성 중 가격에 따른 차등을 위해 헤드램프의 경우 옵션 없음은 0, HID 램프는 1, LED 램프는 2로 코딩하였다. 또한 크루즈 컨트롤의 경우 옵션 없음은 0, 일반 크루즈 컨트롤은 1, 아답티브 크루즈 컨트롤은 2로 코딩하였다. 마지막으로 LPG 차량은 장애인 여부 확인의 문제, 렌탈 및 리스 승계 차량은 중고차 가격 측정의 난해함으로 인해 이들 차량은 분석 대상에서 제외하였다.

3.3 자료의 분류

본 연구는 차량을 차종에 따라 분류하였다. 차종은 세단과 SUV가 대표적이며, 하이브리드 차종은 제조사가 기존 차종과 별개의 차종으로 판매하고 있으며, 소비자 역시 별개의 차종으로 인식하고 있다. 이에 따라 하이브리드를 차종의 범주로 포함하였다. 본 연구의 자료에서 경차를 제외한 세단, SUV, 그리고 하이브리드 차량은 총 24,943대로 확인되었다.

4. 머신러닝 모델의 실행 과정

예측을 위한 머신러닝 모델의 실행 과정은 우선, 자료의 전처리 작업 후 머신러닝에 투입되는 특성을 선별하는 특성 공학으로 시작한다. 이후 자료의 샘플링을 실행한 후 K-Fold 교차검증을 실행하여 샘플링 타당성을 확인한다. 다음으로 머신러닝 모델별 하이퍼 파라미터 튜닝 과정을 통해 최적화된 초매개변수를 도출한 후 이를 활용하여 머신러닝 모델을 실행한다. 마지막으로 머신러닝 모델 간 비용함수를 비교하여 최우수 모델을 선정하는 과정을 거친다.

4.1 특성 공학

특성 공학(Feature Engineering)이란 특성 공간(Feature Space)의 변환을 통한 자료를 활용하여 예측 모델의 성능을 향상시키는 과정이다(Nargesian et al., 2017). 즉 머신러닝 모델에 활용될 특성을 선별하는 과정이며, 크게 두 가지 방법이 존재한다. 첫째, 특성 추출(Feature Extraction)로 주성분 분석(PCA), 선형 판별 분석(LDA), 그리고 2차 판별 분석(QDA) 등이 대표적이다(Seo et al., 2023). 둘째, 특성 선택(Feature Selection)으로 회귀 기반 머신러닝 모델 또는 트리(Decision Tree) 기반 머신러닝 모델을 실행하여 특성을 선택한다. 특히 트리 기반 머신러닝 모델을 활용한 특성 선택은 Cardinality(서로 연관된 실체 유형 간 관계 유형)가 높은 특성을 선호하는 편향성(bias)이 존재할 가능성이 높은 단점이 존재한다(Géron, 2022).

Lasso 회귀 모델은 회귀 기반 머신러닝 모델 중 하나로 선형 회귀의 목적함수인 MSE의 최소화에 페널티인 알파(α)를 부여한다. 이를 통해 회귀계수(W_i)의 크기를 제어하며, 이를 L1 규제(Regularization)라고 한다(Chaudhary et al., 2022). 아래 식은 L1 규제의 과정이다.

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

L1 규제는 회귀계수의 절댓값에 패널티를 부여함으로써 영향력이 미미한 회귀계수를 0으로 도출한다. 본 연구는 앞서 언급한 트리 기반 머신러닝 모델을 통한 특성 선택의 단점을 해결하고자 Lasso 회귀 모델을 통한 특성 선택을 실행하였다.

4.2 과대적합 및 K-Fold 교차검증

머신러닝 모델의 실행에 앞서 자료를 훈련 세트와 시험 세트로 분리하는 샘플링(Sampling) 작업을 실행한다. 머신러닝 모델의 실행 시 훈련 세트 점수가 시험 세트 점수에 비해 과도하게 높을 경우 과대적합(Overfitting), 반대는 과소적합(Underfitting)이라 한다(Ramampiandra et al., 2023). 과대적합과 과소적합 사이의 균형점을 찾는 것은 매우 중요하며, K-Fold 교차검증의 실행을 통해 샘플링의 타당성을 확인할 수 있다. K-Fold 교차검증은 훈련 세트를 검증 세트로 분리한 후 k번 반복하는 과정을 거친다(Staartjes et al., 2022).

4.3 앙상블 모델

앙상블(Ensemble) 모델은 두 개 이상의 CART 모델을 생성한 후 다수결(결과값: 범주형) 또는 평균(결과값: 연속형)으로 최종 결과값을 산출하는 방식이다(Shanti et al., 2021). 본 연구의 결과값인 중고차 가격은 연속형이므로 본 연구에서 실행할 앙상블 모델은 모두 Regressor가 추가된 모델을 실행한다. <Tab. 4-1>은 결과값의 유형에 따른 앙상블 모델의 종류이다.

Tab. 4-1 Ensemble Model

결과값: 범주형	결과값: 연속형
Random Forest	RF Regressor(RFR)
Light Gradient Boosting	LGB Regressor(LGBR)
eXtream Gradient Boosting	XGB Regressor(XGBR)
Categorical Boosting	CB Regressor(CBR)

RF(Random Forest) 모델은 CART(Classification and Regression Tree) 모델의 과대적합을 해결하기 위해 제안되었다. 결과값이 연속형일 때 활용하는 RFR(Random Forest Regressor) 모델은 n개의 CART(Regression Tree) 모델의 결과값에 대한 평균을 도출한다. RF와 RFR 모델은 n의 크기가 어느 정도 증가했을 때 이후 성능의 향상이 작아지는 단점이 존재한다(Breiman, 2001).

LGB(Light Gradient Boosting) 모델은 리프 노드를 지속적으로 분할하여 예측 오차를 감소시키는데, 이를 리프 중심 트리 분할이라 한다. 또한 LGB 모델은 앙상블 모델 중 실행시간이 상대적으로 짧은 편이나, 과대적합의 위험성이 높은 단점이 존재한다(Huang et al., 2022; Wang et al., 2022).

GB(Gradient Boosting) 모델은 LGB 모델과는 다르게 균형 트리 분할이 특징이며(Huang et al., 2022), 처음에는 여러 개의 약한 학습기를 순차적으로 학습하고 예측한 후 잘못 예측한 결과값에 가중치를 부여해 오류를 개선한다. GB 모델은 실행시간이 타 모델에 비해 길다는 단점이 존재하였으나, 이러한 단점을 해결하기 위해 XGB(eXtream Gradient Boosting) 모델이 제안되었다(Chen and Guestrin, 2016). 본 연구는 GB 모델의 비용함수와 차이가 거의 없는 XGB 모델을 실행하였다.

CB(Categorical Boosting) 모델은 모든 훈련 세트가 대상이 아니라 일부 훈련 세트를 대상으로 오차의 계산을 한 결과를 토대로 모델을 다시 생성한다. 특히 CB 모델은 특성들이 범주형일 때 우수한 결과를 기대할 수 있다(Chelgani et al., 2023).

4.4 하이퍼 파라미터 튜닝

하이퍼 파라미터 튜닝(Hyper Parameter Tuning) 과정은 최적화된 머신러닝 모델을 실행하기 위한 초매개변수(Hyper Parameter)의 값을 도출하는 과정이다. <Tab. 4-2>는 일반적으로 사용되는 초매개변수의 목록이다(Yim et al., 2023).

Tab. 4-2 Hyper Parameter

초매개변수	정의
cv(k)	교차검증 횟수
min_samples_split	노드 분할을 위한 최소 샘플 수
min_samples_leaf	리프 노드를 위한 최소 샘플 수
max_features	최적 분할의 feature 수
max_depth	트리의 최대 깊이 수준
max_leaf_nodes	리프 노드의 최대 개수
n_estimate	트리(CART) 개수
learning_rate	경사하강의 지점 이동 수준
subsample	학습 시 사용할 데이터 비율

4.5 비용함수

머신러닝의 목적함수(Objective Function)는 비용함수(Cost Function)의 최소화이다.

Tab. 4-3 Cost Function & R^2

비용함수와 R^2	정의
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)	$\frac{1}{n}\sum_{i=1}^{n}\lvert\frac{(Y_i-\hat{Y}_i)}{Y_i}\rvert\times 100$
R^2(Coefficient of Determination)	$Var(\hat{Y})/Var(Y)$

본 연구의 결과값은 연속형이므로 본 연구에서 활용될 비용함수는 MSE, RMSE, MAE, 그리고 MAPE이며, 상대적으로 수치가 작을수록 우수한 모델이다. 또한 R^2은 비용함수가 아니지만 상대적으로 수치가 클수록 우수한 모델이다. <Tab. 4-3>에서 $\hat{Y}_i$는 머신러닝 모델을 실행한 자료의 예측값, Y_i는 자료의 실제값, 그리고 n은 자료의 총수를 의미한다(Yim et al., 2023).

4.6 SHAP Value

SHAP Value란 개별 특성들의 결합을 통해 도출한 예측값에 대한 특성의 기여도(Feature Attribution)를 평균으로 계산한 값을 의미한다(Chelgani et al., 2023). Lundberg et al.(2018)은 앙상블 모델에서 특성의 기여도가 개별 트리(CART)마다 달라지는 것을 반영하기 위해 Tree SHAP Value를 제안하였다.

$$Tree\ SHAP\ Value = E_i[f(x)|x_s]$$

위 식은 개별 트리(CART; Regression Tree)에서 x_s는 소속된 최종 노드(Leaf Node)에 저장된 결과값의 평균값(예측값)이며, 이 예측값과 전체 훈련 세트에서 x_s가 소속된 최종 노드 속 훈련 세트의 비중을 곱한 조건부 기댓값(Conditional Expectation)을 의미한다. 또한 이들 개별 트리의 조건부 기댓값의 합산을 통해 특성의 기여도를 도출한다.

5. 머신러닝 모델의 실행 결과

본 연구의 목적은 차종별 최우수 앙상블 모델을 선정하는 것이며, 선정된 차종별 최우수 모델을 대상으로 주요 특성의 절대적 기여도와 특성의 방향성을 확인하는 것이다. 이를 위해 본 연구는 첫째, 자료를 차종별 샘플링을 실행하였다. 둘째, 차종별로 특성 선택인 Lasso 회귀 모델을 실행하여 중고차 가격에 미치는 영향력이 0으로 나타난 특성을 도출

하여 소거하였다(범주형 특성 제외). 셋째, 동일 샘플링으로 앙상블 모델의 비용함수를 도출하여 차종별 최우수 모델을 선정하였다. 마지막으로 차종별 최우수 모델을 대상으로 Tree SHAP Value의 시각화를 통해 특성의 절대적 기여도, 방향성, 그리고 특성과 결과값 간 관계를 확인하였다.

5.1 세단 중고차

세단 중고차는 총 11,823대(제네시스: 3,343대, 현대: 3,957대, 기아: 3,377대, 쉐보레: 72대, 르노: 1,074대)로 나타났으며, 차량의 제원 특성은 다음 <Tab. 5-1>과 같다.

Tab. 5-1 Sedan's Specifications Feature

특성	마력(hp)	토크 (kgf · m)	연비 (km/l)	중량(kg)
mean	225.102	30.560	11.344	1628.773
std	75.428	10.631	2.377	257.558
특성	가격(만원)	사용기간 (month)	주행거리 (km)	배기량 (cc)
mean	2739.419	43.811	61076	2516.474
std	1357.038	16.906	38001	744.791

우선 세단 중고차를 대상으로 샘플링(30%, 20%, 10%)을 실행하였다. 그 결과 훈련 세트 점수(예측 정확도)와 시험 세트 점수(예측 정확도) 간 차이가 가장 작게 나타난 비율은 10%로 나타났고, CART 모델로 K-Fold 교차검증을 실행한 결과 k가 6일 때 가장 우수하였다. 이에 따라 동일 샘플링(10%)과 k값(6)으로 Lasso 회귀 모델과 앙상블 모델을 순차적으로 실행하였다.

세단 중고차의 앙상블 모델에 활용하기 위한 특성 선택으로 Lasso 회귀 모델(α=1)을 실행하였다. 그 결과 범주형 특성을 제외하고 영향력이 0으로 나타난 특성은 파워도어록, 블루투스, smartkey, 파워스티어링, 오토라이트, 블라인드(후방), usb단자, 루프랙, 가죽시트, 전동시트(운전석), 통풍시트(동승석), 통풍시트(뒷좌석), 전동사이드미러, 무선도어잠금, ESC, 에어백동승석, 파워윈도우, 후방카메라, 스티어링 리모콘, 주차감지센서(후방), ECM, 에어백(커튼), ABS, 에어백(운전석), 에어백(측면), 그리고 TCS의 유무로 확인되었다. 이에 따라 이 특성들을 제거한 다음 동일 샘플링으로 하이퍼 파라미터 튜닝 과정을 거쳐 앙상블 모델을 실행하였다. <Tab. 5-2>는 세단 중고차의 앙상블 모델별 비용함수 수치이다.

Tab. 5-2 Sedan's K-Fold, Cost Function, R^2

M.L	RFR	LGBR	XGBR	CBR
K-Fold	0.971	0.971	0.972	0.975
RMSE	199.058	195.766	195.313	193.241
MSE	39624.098	38324.16	38147.3	37342.2
MAE	132.424	127.096	128.132	122.017
MAPE(%)	5.552	5.303	5.351	5.002
R^2	0.978	0.979	0.979	0.979

세단 중고차의 앙상블 모델의 실행 결과 비용함수(MAPE 기준) 수치가 가장 우수한 모델로 CBR(Categorical Boosting Regressor) 모델(n_estimators=600)이 선정되었으며, CBR 모델의 비용함수는 RMSE가 193.241, MSE가 37342.251, MAE가 122.017, 그리고 MAPE가 5.002(%)로 나타났다.

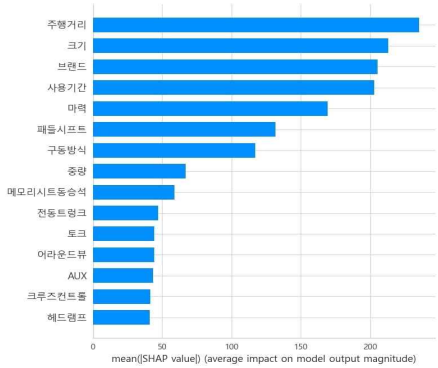

Fig. 5-1 Sedan's CBR SHAP Summary Plot

<Fig. 5-1>은 세단 중고차의 CBR 모델 실행에 따른 시험 세트의 예측값에 대한 Tree SHAP Value의 시각화로써 특성의 절대적 기여 수준을 보여준다. 대표적으로 주행거리, 차량 크기(소형 ~ 대형), 브랜드, 사용기간, 마력, 패들시프트의 유무(옵션 선택률: 50.7%), 그리고 구동방식(전륜: 64.2%, 후륜: 11.5%, 사륜: 24.2%) 순으로 세단 중고차 가격 예측에 기여하는 것으로 확인되었다.

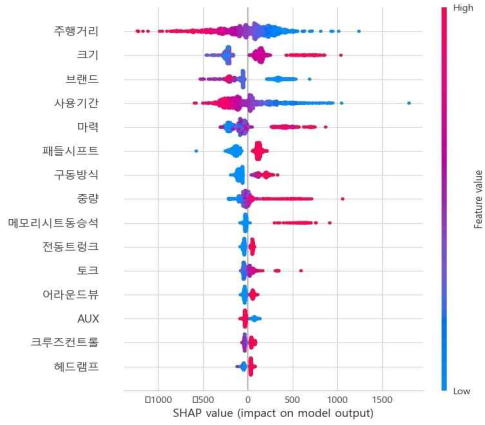

Fig. 5-2 Sedan's CBR SHAP Dot Plot

<Fig. 5-2>는 세단 중고차의 CBR 모델 실행에서 시험 세트의 예측값에 대한 Tree SHAP Value의 시각화로써 특성의 방향성 및 특성과 예측값 간 관계를 보여준다. Y축은 기여도가 높은 특성 순으로 나타나며, X축은 0을 기준으로 왼쪽의 경우 예측값 감소의 기여도, 오른쪽의 경우 예측값 증가의 기여 수준을 의미한다. 대표적으로 주행거리와 사용기간은 길수록(빨간점) 예측값에 음의 방향으로, 반대로 차량 크기와 마력은 클수록(빨간점) 예측값에 양의 방향으로 기여한 것을 보여준다. 또한 브랜드(제네시스: 파란점, 현대/기아: 찐한 파란점, 쉐보레/르노: 빨간점)에 따라 예측값에 음과 양의 방향으로 기여한 것을 보여주며, 차량의 옵션 특성은 빨간점이 옵션 선택, 파란점이 옵션 미선택을 나타낸다.

SHAP Value의 시각화를 통해 세단 중고차의 경우 중고차 가격 예측에 있어 차량의 크기(소형 ~ 대형)와 브랜드의 종류가 사용기간에 비해 기여 수준이 높은 것으로 나타났다. 또한 세단 중고차는 다운사이징(Turbo 엔진) 엔진의 유행으로 인해 배기량이 아닌 중량이 중고차 가격 예측에 기여하는 것으로 확인되었다.

5.2 SUV 중고차

SUV 중고차는 총 11,277대(제네시스: 387대, 현대: 3,737대, 기아: 3,604대, 쉐보레: 702대, 르노: 988대, 쌍용: 1,859대)로 나타났으며, 차량의 제원 특성은 다음 <Tab. 5-3>과 같다.

Tab. 5-3 SUV's Specifications Feature

특성	마력(hp)	토크 (kgf · m)	연비 (km/l)	중량(kg)
mean	187.600	38.068	12.614	1715.709
std	47.119	10.711	1.960	336.754
특성	가격(만원)	사용기간 (month)	주행거리 (km)	배기량 (cc)
mean	2471.042	44.675	60892	2036.653
std	1176.942	17.982	40430	526.727

우선 SUV 중고차를 대상으로 샘플링(30%, 20%, 10%)을 실행하였다. 그 결과 훈련 세트 점수와 시험 세트 점수간 차이가 가장 작게 나타난 비율은 10%로 나타났고, CART 모델로 실행한 결과 k가 6일 때 가장 우수하였다. 이에 따라 동일 샘플링(10%)과 k값(6)으로 Lasso 회귀 모델과 앙상블 모델을 순차적으로 실행하였다.

다음으로 앙상블 모델에서의 특성 선택을 위해 Lasso 회귀 모델(α=1)을 실행하였다. 그 결과 범주형 특성을 제외하고 영향력이 0으로 나타난 특성은 smartkey, 메모리시트(동승석), 무선도어잠금, 블라인드(후방), 통풍시트 운전석, 후면AV모니터, usb단자, 가죽시트, 전동시트(운전석), 블루투스, 후

방카메라, 스티어링 리모콘, 파워도어록, 알루미늄휠, 파워스티어링, 전동사이드미러, 파워윈도우, 에어백(운전석), 에어백(동승석), 에어백(측면), 에어백(커튼), ABS, TCS, ESC, TPMS, 주차감지센서(전방), 그리고 주차감지센서(후방)의 유무로 확인되었다. 이에 따라 이 특성들을 제거한 다음 동일 샘플링으로 하이퍼 파라미터 튜닝 과정을 거쳐 앙상블 모델을 실행하였다. <Tab. 5-4>는 SUV 중고차의 앙상블 모델별 비용함수 수치이다.

Tab. 5-4 SUV's K-Fold, Cost Function, R^2

M.L	RFR	LGBR	XGBR	CBR
K-Fold	0.978	0.979	0.978	0.981
RMSE	159.798	152.576	160.515	148.630
MSE	25535.441	23279.56	26764.8	22090.8
MAE	110.684	106.306	107.993	101.770
MAPE(%)	4.963	4.780	4.858	4.581
R^2	0.980	0.982	0.980	0.982

SUV 중고차의 앙상블 모델의 실행 결과 비용함수(MAPE 기준) 수치가 가장 우수한 모델로 CBR 모델(n_estimators=500)이 선정되었다. CBR 모델의 비용함수는 RMSE가 148.630, MSE가 22090.807, MAE가 101.770, 그리고 MAPE가 4.581(%)로 나타났다.

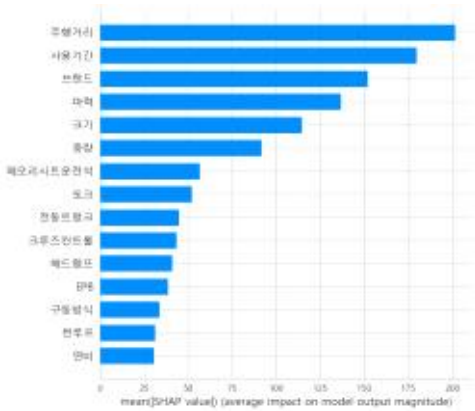

Fig. 5-3 SUV's CBR SHAP Summary Plot

<Fig. 5-3>은 SUV 중고차의 CBR 모델 실행에 따른 시험 세트의 예측값에 대한 Tree SHAP Value의 시각화로 특성의 절대적 기여 수준을 보여준다. 대표적으로 주행거리, 사용기간, 브랜드, 마력, 차량 크기, 중량, 그리고 메모리시트 운전석(옵션 선택률: 28.9%)의 유무 순으로 SUV 중고차 가격 예측에 기여한 것으로 확인되었다.

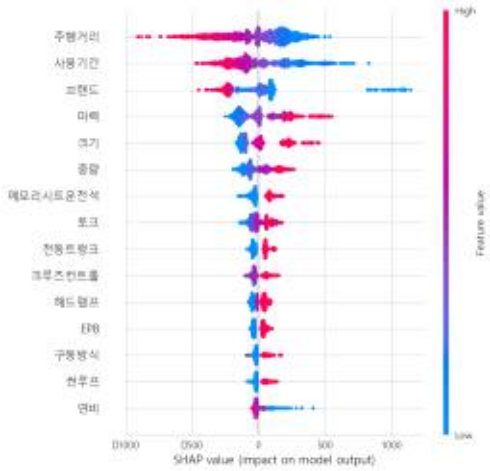

Fig. 5-4 SUV's CBR SHAP Dot Plot

<Fig. 5-4>는 SUV 중고차의 CBR 모델 실행에서 시험 세트의 예측값에 대한 Tree SHAP Value의 시각화로 특성의 방향성, 특성과 예측값 간 관계를 보여준다. 대표적으로 주행거리와 사용기간은 길수록(빨간점) 예측값에 음의 방향으로, 반대로 마력, 크기, 그리고 중량은 클수록(빨간점) 예측값에 양의 방향으로 기여한 것을 보여준다. 특히 브랜드에서 제네시스(파란점)는 예측값에 타 브랜드에 비해 크게 기여한 것을 보여주며, 차량의 옵션 특성은 빨간점이 옵션 선택, 파란점이 옵션 미선택을 나타낸다.

SHAP Value의 시각화를 통해 SUV 중고차의 가격 예측에 있어 주행거리와 사용기간의 기여 수준이 가장 높으며, 메모리시트(운전석) 유무가 기여 수준의 상위권(7위)에 있는 것으로 나타났다. 또한 SUV 차종은 세단 차종과 같이 다운사이징 엔진(터보 엔

진)의 보급으로 인해 배기량이 아닌 중량이 중고차 가격 예측에 기여하는 것으로 확인되었다.

5.3 하이브리드 중고차

하이브리드 중고차는 총 1,843대(현대: 897대, 기아: 946대)로 나타났으며, 차량의 제원 특성은 다음 <Tab. 5-5>과 같다.

Tab. 5-5 Hybrid's Specifications Feature

특성	마력(hp)	토크 (kgf · m)	연비 (km/l)	중량(kg)
mean	192.099	20.534	17.341	1611.942
std	29.933	3.878	2.017	130.366
특성	가격(만원)	사용기간 (month)	주행거리 (km)	배기량 (cc)
mean	3163.476	34.471	51192	1999.36
std	907.364	18.241	41390	361.059

우선 하이브리드 중고차를 대상으로 샘플링(30%, 20%, 10%)을 실행하였다. 그 결과 훈련 세트 점수와 시험 세트 점수 간 차이가 가장 작게 나타난 비율은 10%로 나타났고, CART 모델로 K-Fold 교차검증을 실행한 결과 k가 6일 때 가장 우수하였다. 이에 따라 동일 샘플링(10%)과 k값(6)으로 Lasso 회귀 모델과 앙상블 모델을 순차적으로 실행하였다.

다음으로 앙상블 모델에서의 특성 선택을 위해 Lasso 회귀 모델(α=1)을 실행하였다. 그 결과 범주형 특성을 제외하고 영향력이 0으로 나타난 특성은 블라인드(뒷좌석), 통풍시트(동승석), autoair, smartkey, 무선도어잠금, 오토라이트, 블루투스, 네비게이션, 전면AV모니터, 후면AV모니터, usb단자, 가죽시트, 후방카메라, 통풍시트(운전석), 메모리시트(동승석), 열선시트(앞좌석), 통풍시트(뒷좌석), 후측방경보, 에어백(동승석), 연비, 고스트도어클로징, 전동사이드미러, 알루미늄휠, 스티어링리모콘, 하이패스, 파워도어록, 주차감지센서(후방), 파워윈도우, 에어백(운전석), 파워스티어링, 에어백(커튼), ABS, 주차감지센서전방, TCS, ESC, TPMS, 그리고 에어백(측면)의 유무로 확인되었다. 이에 따라 이 특성들을 제거한 다음 동일 샘플링으로 하이퍼파라미터 튜닝 과정을 거쳐 앙상블 모델을 실행하였다. <Tab. 5-6>은 하이브리드 중고차의 앙상블 모델별 비용함수 수치이다.

Tab. 5-6 Hybrid's K-Fold, Cost Function, R^2

M.L	RFR	LGBR	XGBR	CBR
K-Fold	0.950	0.950	0.945	0.957
RMSE	187.427	173.776	196.905	172.247
MSE	35128.79	30198.23	38771.6	29669.1
MAE	140.758	130.010	144.621	125.544
MAPE(%)	4.620	4.190	4.580	4.000
R^2	0.957	0.963	0.953	0.964

하이브리드 중고차에 대한 앙상블 모델의 실행 결과 비용함수(MAPE 기준) 수치가 가장 우수한 모델로 CBR 모델(n_estimators=200)이 선정되었다. CBR 모델의 비용함수는 RMSE가 172.247, MSE가 29669.108, MAE가 125.544, 그리고 MAPE가 4.000(%)으로 나타났다.

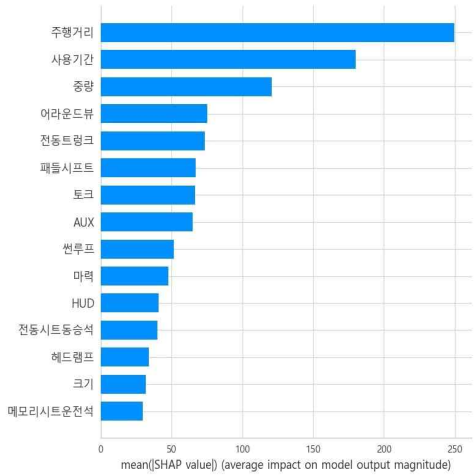

Fig. 5-5 Hybrid's CBR SHAP Summary Plot

<Fig. 5-5>는 하이브리드 중고차의 CBR 모델 실행에 따른 시험 세트의 예측값에 대한 Tree

SHAP Value의 시각화로 특성의 절대적 기여 수준을 보여준다. 대표적으로 주행거리, 사용기간, 중량, 어라운드뷰(옵션 선택률: 49.8%), 전동트렁크(옵션 선택률: 57.2.%), 패들시프트(옵션 선택률: 51.2%)의 유무, 그리고 토크 순으로 하이브리드 중고차의 가격 예측에 기여한 것으로 확인되었다.

<Fig. 5-6>은 하이브리드의 CBR 모델 실행에서 시험 세트의 예측값에 대한 Tree SHAP Value의 시각화로 특성의 방향성, 특성과 예측값 간 관계를 보여준다. 대표적으로 주행거리와 사용기간은 길수록(빨간점) 예측값에 음의 방향으로, 반대로 차량 중량과 토크는 클수록(빨간점) 예측값에 양의 방향으로 기여한 것을 보여준다. 또한 차량의 옵션 특성은 빨간점이 옵션 선택, 파란점이 옵션 미선택을 나타낸다.

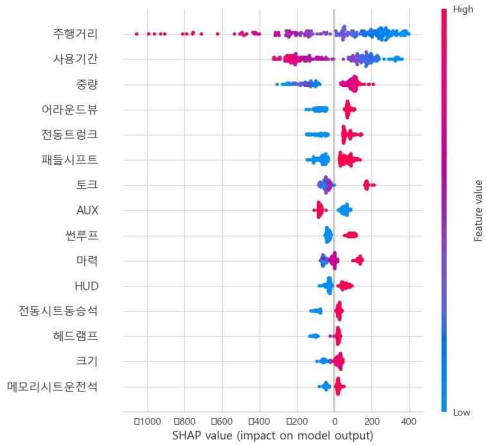

Fig. 5-6 Hybrid's CBR SHAP Dot Plot

Tree SHAP Value의 시각화를 통해 하이브리드 중고차의 가격 예측에 있어 주행거리와 사용기간의 기여 수준이 가장 높으며, 어라운드뷰, 전동트렁크, 그리고 패들시프트 유무가 기여 수준의 상위권에 있는 것으로 나타났다. 또한 하이브리드 차종의 배기량은 대부분 1600cc(일부 차량 제외)이며, 중형차부터 다운사이징 엔진(1600cc 터보 엔진)과 전기모터가 결합한 하이브리드 시스템으로 인해 배기량이 아닌 중량이 중고차 가격 예측에 기여하는 것으로 확인되었다.

6. 결론

본 연구는 국내 온라인 중고차 플랫폼 서비스를 대상으로 크롤링과 스크랩핑을 실행한 후 국내 브랜드의 차량 정보를 획득하였다. 이후 획득한 자료를 3개 차종으로 분류하여 차종별 중고차 가격 예측을 위한 최우수 모델을 제안하고, 차종별 특성의 기여 수준과 방향성을 확인하였다. 이를 위해 따라 본 연구는 우선 차종별 특성 공학(Feature Engineering) 과정으로 특성 선택(Feature Selection) 중 하나인 Lasso 회귀 모델을 실행하여 중고차 가격에 미치는 영향력이 0으로 나타난 특성을 선별하였다. 다음으로 차종별 동일 샘플링으로 영향력이 0인 특성을 소거한 후 앙상블(Ensemble) 모델을 실행하였다. 이와 같은 과정을 거쳐 차종별 앙상블 모델의 비용함수 수치를 비교하여 차종별 가장 우수한 모델을 선정하였다. 마지막으로 차종별 최우수 모델을 대상으로 Tree SHAP Value의 시각화인 Summary Plot을 통해 차종별 특성의 절대적 기여도를 확인하고, Dot Plot을 통해 차종별 특성의 방향성과 특성과 결과값 간 관계를 확인하였다.

본 연구의 결과는 다음과 같다. 우선 세단 중고차에서 CBR(Categorical Boosting Regressor) 모델이 최우수 모델로 선정되었으며, MSE가 37342.251, MAPE가 5.002(%), 그리고 R^2이 0.979로 나타났다. 또한 세단 중고차의 차량 제원 특성인 차량 크기(차급), 브랜드(제네시스 여부), 그리고 마력이 중고차 가격에 양의 방향으로 기여를, 차량 옵션 특성인 패들시프트 유무와 구동 방식은 중고차 가격에 양의 방향으로 기여하는 것으로 확인되었다. 다음으로 SUV 중고차에서 CBR 모델이 최

우수 모델로 선정되었으며, MSE가 22090.807, MAPE가 4.581(%), 그리고 R^2이 0.982로 나타났다. 또한 SUV 중고차의 차량 제원 특성인 브랜드(제네시스 여부), 마력, 차량 크기, 그리고 중량이 중고차 가격에 양의 방향으로 기여를, 차량 옵션 특성인 메모리시트(운전석)의 유무가 중고차 가격에 양의 방향으로 기여하는 것으로 확인되었다. 마지막으로 하이브리드 중고차에서 CBR 모델이 최우수 모델로 선정되었으며, MSE가 29669.108, MAPE가 4.000(%), 그리고 R^2이 0.964로 나타났다. 또한 하이브리드 중고차의 차량 제원 특성인 중량과 토크가 중고차 가격에 양의 방향으로 기여를, 차량 옵션 특성인 어라운드뷰, 전동트렁크, 그리고 패들시프트의 유무가 중고차 가격에 양의 방향으로 기여하는 것으로 확인되었다. 마지막으로 세단, SUV, 그리고 하이브리드 중고차 모두 공통적으로 차량 제원 특성인 주행거리와 사용기간이 중고차 가격에 음의 방향으로 기여하는 것으로 확인되었다.

본 연구의 시사점은 다음과 같다. 첫째, 모든 차종에서 CBR 모델이 최우수 모델로 선정되었다. CBR 모델은 특성이 범주형일 시 타 모델에 비해 비용함수 수치가 우수한 장점이 있다(Chelgani et al., 2023). 본 연구에 활용된 특성은 일부를 제외한 대부분이 서열척도인 동시에 범주형(0: 없음, 1: 있음)으로 구성되었다. 이와 같은 이유로 CBR 모델의 강점이 본 연구에서 발현된 것으로 판단되며, 범주형 특성이 많은 자료의 경우 CBR 모델의 활용을 고려할 수 있다. 둘째, 본 연구는 Lundberg et al.(2018)이 제안한 Tree SHAP Value의 시각화를 실행하여 차종별 특성의 기여도와 방향성, 특성과 결과값 간 관계를 확인하였다. 이를 통해 차종 간 특성의 기여도 차이를 확인하였다. 그러나 Summary Plot과 Dot Plot에서 특성의 기여도는 인과관계 검증을 위한 선형 회귀의 회귀계수와 동일한 의미가 아니며, 이에 따라 특성의 기여도 해석 시 주의가 필요하다. 즉 특성의 기여도는 조건부 기댓값(Conditional Expectation)으로서, 특성과 결과값 간 상관관계라 볼 수 있다. 셋째, 중고차 가격 예측에 관한 선행연구는 대부분 중고차 가격을 서열척도로 구성하였으나, 본 연구는 연속형으로 구성함으로써 척도 내 오차의 문제를 해결하고자 하였다.

본 연구의 결과는 중고차 온라인 플랫폼 서비스를 이용하는 매매관계자가 차종에 따른 차량 제원 및 차량 옵션 수준을 기반으로 대략적인 중고차의 가격 예측에 도움이 되고자 하였다. 또한 본 연구에서 선정된 중고차 가격 예측 모델이 중고차 구매자와 판매자 간 정보의 비대칭성으로 야기된 문제 해결에 지원이 되기를 희망한다.

본 연구의 한계점으로 우선 중고차 온라인 플랫폼 서비스에 공시된 중고차 가격은 판매자와 구매자 간 실제 거래가격과 다를 수 있다. 그러나 이들의 입장에서 온라인 플랫폼 서비스에 공시된 가격은 중고차의 거래를 위한 시작점인 동시에 거래의 기준점으로 볼 수 있다. 다음으로 2022년도 11월 기준 국내 자동차 시장에서 해외 브랜드의 신차 점유율이 19.55%로 나타났으며(Han, 2022), 동시에 해외 브랜드의 중고차 거래가 활발히 이루어지고 있다. 이에 따라 추후 해외 브랜드를 대상으로 중고차 가격 예측 모델을 제안할 필요가 있다.

[References]

[1] Breiman, L.(2001), Random Forests, *Machine Learning*, 45, 5-32.

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

[3] Chelgani, S. C., Nasiri, H., Tohry, A., and Heidari, H. R.(2023), Modeling industrial hydrocyclone

operational variables by SHAP-CatBoost-A “conscious lab” approach, *Powder Technology*, 420, 118416.

[4] Chen, T. and Guestrin, C.(2016), XGBoost: A scalable tree boosting system, In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

[5] Das Adhikary, D. R., Sahu, R., and Pragyna Panda, S.(2022), Prediction of used car prices using machine learning, In *Biologically Inspired Techniques in Many Criteria Decision Making: Proceedings of BITMDM 2022*, 131-140.

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

[7] Géron, A.(2022), Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, O’Reilly Media, Inc.

[8] Han, J.H.(2022. 12. 7.) Imported cars have a nearly 20% market share this year, the largest “acceleration“, Dong-A ILBO from https://www.donga.com/news/Economy/article/all/20221206/116881585/1 (한재희(2022. 12. 7.) 수입차 올해 점유율 20% 육박 역대 최대 ‘급가속’, 동아일보)

[9] Huang, J., Yu, Z., Ning, Z., and Hu, D.(2022), Used Car Price Prediction Analysis Based on Machine Learning, In *2022 International Conference on Artificial Intelligence, Internet and Digital Economy, 2022 International Conference on Artificial Intelligence, Internet and Digital Economy*, 356-364.

[10] Kim, H.W.(2023) SUV Preferred Consumers…“The interior space is spacious, the body is high, and the ride feels good.“, UPI NEWS. from https://upinews.kr/newsView/upi202303200062 (김해욱(2023. 3. 20.) SUV 선호하는 소비자들…“실내 공간 넓고, 차체 높고, 승차감 좋아져“, UPI 뉴스)

[11] Lee, C. H.(2022), Opportunities in the used car industry, Samsung Securities, 1-31 (이창희 (2022), 중고차 산업에서의 기회, 삼성증권, 1-31)

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

[13] Lee, J. I.(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, 1-36)

[14] Lee, S. H.(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)

[15] Lim, K. C.(2023) The endless strides of hybrids…Number one increase in share by fuel, YONHAP NEWS. from

https://www.yna.co.kr/view/AKR20230527040100003 (임기창(2023. 5. 28.), 하이브리드의 끝없는 약진…연료별 점유율 상승폭 1위, 연합뉴스)

[16] Lundberg, S. M., Erion, G. G., and Lee, S. I.(2018), Consistent individualized feature attribution for tree ensembles, *arXiv preprint,* arXiv:1802.03888.

[17] 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.

[18] Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E.B., and Turaga, D.S.(2017), Learning Feature Engineering for Classification, *In Ijcai,* 17, 2529-2535.

[19] 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 Modelling*, 481, 110353.

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

[21] 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.

[22] Sikora, R.(2015), A modified stacking ensemble machine learning algorithm using genetic algorithms, In *Handbook of research on organizational transformations through big data analytics*, 43-53.

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

[24] Seo, H. J., Kim, D. H., and Byun, J. H.(2023), Data Pre-processing for Manufacturing Quality Improvement, *Journal of the Korean Institute of Industrial Engineers*, 49(3), 248-257 (서호진, 김도현, 변재현 (2023), 제조품질 향상을 위한 데이터 전처리 프로세스, *대한산업공학회지*, 49(3), 248-257.

[25] Wang, A., Yu, Q., Li, X., Lu, Z., Yu, X., and Wang, Z.(2022), Research on Used Car Valuation Problem Based on Machine Learning, In *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA),* 101-106.

[26] Yim, S. J., Lee, J. H., and Ryu, C. H.(2023), A Study on the Prediction Models of Used Car Prices for Domestic Brands Using Machine Learning, *Journal of Service Research and Studies*, 13(3), 106-127 (임승준, 이정호, 류춘호 (2023), 머신러닝을 활용한 브랜드별 국내 중고차 가격 예측 모델에 관한 연구, *서비스 연구*, 13(3), 106-127)

Yim, Seungjun (yimdgz@hotmail.com)

Seungjun Yim completed a bachelor's degree in Naval Architecture & Ocean Engineering at Hongik University and a master's degree in Service Operations Management from the Department of Business Administration at Hongik University. And He received his Ph.D. in Operations Research & Operations Management from the Department of Business Administration at Hongik University. His current research interests include Service Quality, Product Service-System Design(PSS), and Data Science.

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

Joungho Lee is an associate professor at the School of Business of Konkuk University. He completed a bachelor's degree in mechanical engineering at Hongik University and a master's degree in Operations Management from the Department of Business Administration at Hongik University. And he received his Ph.D. in Management Science 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, Seoul, Korea. He received a bachelor's degree in economics from Seoul National University and a master's degree in management science from KAIST. 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(especially, combinatorial optimization), productivity and teacher evaluation.

A Study on the Prediction Models of Used Car Prices Using Ensemble Model And SHAP Value: Focus on Feature of the Vehicle Type

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

ABSTRACT

The market share of online platform services in the used car market continues to expand. and The used car online platform service provides service users with specifications of vehicles, accident history, inspection details, detailed options, and prices of used cars.

SUV vehicle type's share in the domestic automobile market will be more than 50% in 2023, Sales of Hybrid vehicle type are doubled compared to last year. And these vehicle types are also gaining popularity in the used car market. Prior research has proposed a used car price prediction model by executing a Machine Learning model for all vehicles or vehicles by brand. On the other hand, the popularity of SUV and Hybrid vehicles in the domestic market continues to rise, but It was difficult to find a study that proposed a used car price prediction model for these vehicle type.

This study selects a used car price prediction model by vehicle type using vehicle specifications and options for Sedans, SUV, and Hybrid vehicles produced by domestic brands. Accordingly, after selecting feature through the Lasso regression model, which is a feature selection, the ensemble model was sequentially executed with the same sampling, and the best model by vehicle type was selected. As a result, the best model for all models was selected as the CBR model, and the contribution and direction of the features were confirmed by visualizing Tree SHAP Value for the best model for each model.

The implications of this study are expected to propose a used car price prediction model by vehicle type to sales officials using online platform services, confirm the attribution and direction of features, and help solve problems caused by asymmetry fo information between them.

Keywords: Used Car Online Platform Service, Used Car Price, Vehicle Type, Ensemble Model, SHAP Value

* First Author, Ph.D., Department of Business Administration, Hongik University
** Corresponding Author, Associate Professor, School of Business, Konkuk University
*** Coauthor, Professor, Department of Business Administration, Hongik University