Germany 중고차의 판매 가격 예측을 위한 머신러닝 기반 모델 개발

김지수 [01] 김태윤 [2] 양진주 [1] 김성현 [1] 이준범 [3] 조석헌*
[1] 숙명여자대학교 [2] 국립금오공과대학교 [3] 광주과학기술원
*Qualcomm Institute, University of California, San Diego (UCSD)
jisugim168@gmail.com, kty08062@gmail.com, jinjuyang80@gmail.com,
smwuai2004@gmail.com, kevin80681@gmail.com, justinshcho@gmail.com

Development of a Machine Learning-Based Prediction Model For Used Car Listing Prices in Germany

Jisoo Kim[01] TaeYoon Kim[2] JinJoo Yang[1] SungHyun Kim[1] JunBeom Lee[3] Seokheon Cho*
[1]Sookmyung Women's University, [2]Kumoh National Institute of Technology, [3]Gwangju Institute of Science and Technology (GIST), *Qualcomm Institute, University of California, San Diego (UCSD)

요 약

In recent years, the demand for used cars has grown significantly due to the rising costs of new vehicles. Accurately predicting fair listing prices for used cars can help alleviate information asymmetry between buyers and sellers, reducing the risk of financial losses during transactions. This study proposes a machine learning-based approach to predict the final listing price of used cars on online platforms. Two regression models—Multiple Linear Regression (MLR) and Random Forest Regression (RFR)—were trained on a dataset collected from German online car marketplaces. To evaluate the impact of encoding strategies on model performance, we applied both one-hot encoding and target encoding to the categorical variables in the dataset. The MLR-based model performed better with one-hot encoding, while the RFR-based model achieved higher accuracy when trained with target-encoded data. These results suggest that, target encoding better captures the relationship between categorical variables and price, and RFR is well-suited for modeling the nonlinear characteristics of used car data.

1. 서 론

차량 구매자는 증가하는 신차의 가격에 부담을 느끼고 이에 비교적 가격이 낮은 중고차에 대한 수요가 증가하고 있다. 이러한 이유로 중고차 시장은 점점 규모가 커지고 있다 [1]. 하지만 중고차 매매 시 구매자와 판매자 간의 정보 불균형 때문에 구매자뿐만 아니라 판매자도 피해를 볼 수 있는 상황이 존재한다 [2, 3]. 과거에는 오프라인으로 중고차 거래가 이루어 졌다면 최근에는 온라인 중고차 거래도 활발히 이루어지고 있다. 이는 오프라인 거래 시에 선호하는 차량에 대한 정보와 가격 등을 수집하기 위해 필요한 비용을 온라인 거래를 통해 절감할 수 있기 때문이다 [3]. 만약 판매자와 소비자 모두 합리적인 가격을 미리 예측할 수 있다면 차량 판매에 대한 정보 불균형으로 인한 피해를 방지하고 보다 적극적인 온라인 차량 매매가 이루어질 것이다. 따라서, 차량 및 시점에 따른 차량 판매자가 선호하는 차량 판매 가격을 예측하는 모델을 개발하고 이를 차량 구매자가 효과적으로 이용할 수 있도록 하는 관련 시스템 개발이 필요하다.

* This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation) in 2025(2022-0-01087) & (2024-0-00062).
* Following are results of a study on the "Leaders in Industry-university Cooperation 3.0 "Project, supported by the Ministry of Education and National Research Foundation of Korea.

S. Kumar *et al.*은 중고차 판매 가격을 예측하기 위해 단순한 Linear Regression (LR)모델만 사용하는 것에서 더 나아가 다양한 형태의 회귀 모델을 제안하였다 [4]. 또한, 가장 우수한 성능을 가진 모델을 선정하기 위해 Akaike Information Criterion (AIC) 및 Bayesian Information Criterion (BIC)를 사용하였다. 해당 연구에서는 불필요한 변수를 제거하고 AIC와 BIC 기법을 적용하여 R-squared (R^2)값을 0.903으로 향상시켰다. J. HE는 중고차 가격 예측 모델을 제시하는데 있어서, 고려하는 데이터세트에 포함된 범주형 변수를 여러 가지 방식으로 인코딩하였다 [5]. 이 중 원핫 인코딩 (One-hot Encoding)과 라벨 인코딩 (Label Encoding)을 혼합한 방식이 성능 향상을 가져왔다. 또한, LR 알고리즘 기반 모델이 저가 및 중가에 해당하는 차량에 대해서만 준수한 성능을 보였고 고가 중고차 판매 가격 예측 모델의 성능 개선을 위해서 비선형 모델이 필요함을 제시하였다. LR 알고리즘이 아닌 다른 머신 러닝 (Machine Learning) 알고리즘을 사용하여 중고차 판매 가격 예측 모델을 개발한 연구도 진행되었다. 특히, F. Wang *et al.*은 Decision Tree Regression, Extra Tree Regression (ETR), Random Forest Regression 그리고 Ridge Regression 알고리즘을 고려하였다 [6]. 4개의 알고리즘에서 ETR 기반 모델을 사용하여 특성 중요도를 구하고 이에 따라 선택한 주요 특성을 가지고 학습을 수행한 ETR 알고리즘 기반 모델이 0.9807의 R^2를 가짐으로써 최고의 성능을 보였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 Used Cars 데이터세트에 대한 설명과 데이터 전처리 과정을 언급한다. 제 3장에서는 본 연구에서 사용하는 머신러닝 알고리즘과 성능 평가 지표에 대해서 언급한다. 제 4장에서는 중고차 선호 판매 가격 예측 모델의 성능에 대해 비교 및 분석한다. 마지막으로, 제 5장에서는 본 연구의 결과와 향후 연구 과제에 대해 말한다.

2. 데이터세트 구성 및 전처리 과정

2.1 원본 데이터세트 설명

본 연구에서는 Kaggle에서 제공하는 독일에서 거래되는 Used Cars 데이터세트를 사용하였다 [7]. 이 데이터세트에는 중고차 매물이 웹사이트에 등록된 날짜, 각 중고차 매물 정보를 수집한 날짜와 마지막으로 수집한 날짜, 차량의 출고 년도와 달, 연료 유형, 주행 거리, 마력, 손상 여부 그리고 차량 판매 가격 등 21개의 특징으로 구성되어 있다. 본 데이터세트는 여러 중고차 웹사이트에서 2016년 03월부터 2016년 04월까지 수집된 데이터이기에 Car_Name (차량 이름), Car_Brand (차량 브랜드), Car_Model (차량 모델), Car_Type (차량 유형), Horsepower (마력), Fuel_Type (주유 유형)과의 값이 모두 동일할 경우 Date_Created (차량 정보 수집 날짜)의 값이 가장 최신인 것에 해당하는 데이터만 남기고 이전 일시에 수집한 동일한 차량에 대한 데이터는 삭제하였다. 이 후 Car_Name, Date_Created 그리고 Lastseen (차량 정보 최종 수집 날짜) 변수를 삭제하였다. 또한, 중고차 가격에 많은 영향을 끼치는 차량의 나이 (Car_Age)를 구하기 위해 차량이 출고된 시점과 차량 데이터를 수집한 시점의 차이를 이용해 계산하고 이를 새로운 독립 변수로 고려하였다.

2.2 Used Cars 데이터세트 설명과 전처리 과정

표 1은 본 연구에서 개발할 중고차 판매 가격 예측 모델을 위해 사용할 데이터세트에 포함된 변수의 특징을 나타내고 있다. 고려하는 Used Cars 데이터세트에는 9개의 독립 변수와 1개의 종속 변수를 포함하고 있다. 중고차 가격 예측 모델이 안정적인 성능을 가지기 위해서 중고차 시장에 나오는 다수의 차량 모델에 대해 예측을 한정하고자 한다. 이에 Car_Brand중 빈도수 기준 상위 5개의 독일차량 브랜드인 Volkswagen, BMW, Audi, Mercedes_Benz 그리고 Opel을 선정하였다. 또한, 이 5개의 Car_Brand에서 출시하는 수많은 차량 모델 중에서 Used Cars 데이터세트에서 500개 이상의 빈도수를 가지는 차량 모델만 사용하였다. Car_Model, Car_Type 및 Fuel_Type 변수에는 기타라는 값을 의미하는 Andere를 포함하는 데이터 샘플은 제거하였다. 차량의 최초 출고 연월이 데이터 수집 시점인 2016년 3월 또는 4월 이전인 경우에 한해 데이터를 추출하였다. Car_Age는 데이터 수집 시점에서 차량 출고 연월을 차감하여 계산하였다. Horsepower에 대한 이상치 (outlier)를 제거한 후 30 ~ 800 [PS]로 제한하였다. 또한, Horsepower 변수 값 중 Series를 의미하는 Reihe를 포함하는 데이터 샘플이 있는데 이 값을 특정할 수 없어 해당 데이터 샘플을 제거하였다. 종속 변수인 Listing_Price에 있어서 이상치를 제거한 후 500 [EUR: €]에서 24,500 [€]까지 값을 가지는 데이터 샘플만 고려하였다.

표 1. Used Cars 데이터세트 변수 및 특징

Variables	Type	Values	Unit
Independent Variables			
Car_Brand	String	{Volkswagen, BMW, Audi, Mercedes Benz, Opel}	–
Car_Model	String	–	–
Odometer	Integer	–	Km
Gearbox	String	{Automatic, Manual}	–
Car_Type	String	{Limousine, Station wagon, Compact cars, Bus, Coupe, Convertible, SUV}	–
Fuel_Type	String	{EV, CNG, Diesel, Gasoline, Hybrid, LPG}	–
Car_Age	Double	–	Year
Horsepower	Integer	–	PS
Unrepaired_Damage	String	{Yes, No}	–
Dependent Variable			
Listing_Price	Integer	–	EUR[€]

3. 머신러닝 알고리즘 및 성능 평가 지표

3.1 머신러닝 모델 및 하이퍼파라미터 설정

본 연구에서는 Multiple Linear Regression (MLR)과 Random Forest Regression (RFR) 알고리즘을 사용하여 중고차별 판매자가 최종 판매 가격을 예측하였다. 과적합 방지 및 모델 성능의 평균값을 구하기 위해 K-Fold Cross Validation을 사용하였다. 이 때 $k = 5$로 설정하여 학습 데이터세트와 테스트 데이터세트를 각각 80%과 20%의 비율로 나누었다.

3.2 성능 평가 지표

중고차 판매 가격 예측 모델에 대한 성능 평가 지표로 R^2 (R-squared: 결정계수)와 RMSE (Root Mean Squared Error)를 사용하였다. R^2는 모델이 종속 변수인 Listing_Price의 변동성을 얼마나 잘 설명하는지를 나타내는 지표이다 [8]. 해당 값이 1에 가까울수록 모델의 설명력이 높음을 의미하며, 회귀 모델의 예측 성능을 종합적으로 판단하는 데 효과적인 지표이다.

RMSE는 오차의 제곱에 대한 평균을 구한 후 제곱근을 취한 값으로 다음과 같이 정의될 수 있다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

여기에서, y_i와 $\hat{y}_i$는 각각 i번째 데이터 샘플의 실제 가격과 예측 가격이다. 또한 n은 데이터 샘플의 총 개수를 나타낸다.

4. 중고차 선호 판매 가격 예측 모델 분석

본 연구에서 고려하는 Used Cars 데이터세트에서 포함하고 있는 범주형 변수를 변환 방식에 따라 중고차 선호 판매 가격 예측 모델의 성능을 분석하고자 한다. 범주형 변수에 대한 고려하는 변환 방식은 원핫 인코딩 (One-hot Encoding)과 타겟 인코딩 (Target Encoding)이다. 특히 타겟 인코딩은 범주형 변수의 각 범주를 Listing_Price에 대한 평균값으로 변환하는 기법이다. k번째 범주에 해당하는 샘플들의 타겟 인코딩 값은 다음과 같다.

$$TE(k) = \frac{1}{n_k}\sum_{j=1}^{n} y_j * \mathbb{1}_{\{x_j \in k\}} \quad (2)$$

여기에서, $\mathbb{1}_{\{x_j \in k\}}$는 범주형 변수의 j번째 샘플 x_j가 범주 k에 속할 경우 1의 값을 가지거나, 그렇지 않으면 0인 indicator function 이다. y_j는 j번째 샘플의 실제 가격을 나타내고 n_k는 범주 k에 해당하는 샘플들의 수이다.

4.1 Multiple Linear Regression (MLR)

MLR 알고리즘 기반 중고차 판매 가격 예측 모델에 있어서 참조할 범주형 변수를 인코딩하는 방식으로는 one-hot encoding

과 target encoding을 선정하였다.

표 2는 인코딩 방식에 따른 두 개의 데이터세트를 학습한 MLR 알고리즘 기반 모델의 성능 결과를 비교하고 있다. MLR 알고리즘 기반 중고차 판매 가격 예측 모델을 학습하는데 사용하는 범주형 변수를 target encoding 방식보다 one-hot encoding 방식으로 변환하는 것이 성능 향상을 가져옴을 확인할 수 있다. 이는 one-hot encoding이 R^2 향상 및 RMSE 감소에 효과적이기 때문이다. 반면, target encoding은 모델이 범주 간 차이를 충분히 반영하기 어렵다. 또한, MLR 알고리즘은 비선형 관계를 학습하기 어렵기 때문에 target encoding에서 도출한 평균값이 독립 변수의 범주 간에 미묘한 차이를 설명하지 못하는 경우 예측 정확도가 저하될 수 있다. 따라서, MLR 기반 모델과 같이 독립 변수와 종속 변수와의 관계가 단순한 모델에서는 one-hot encoding 방식이 target encoding 보다 더 효과적임을 알 수 있다.

표 2. 인코딩 방식에 따른 MLR 알고리즘 기반 예측 모델 성능

	One-hot Encoding	Target Encoding
R^2	0.741	0.699
RMSE [€]	4,214.52	4,546.95

4.2 Random Forest Regression (RFR)

RFR 알고리즘 기반 모델을 학습하기 위해 사용하는 범주형 변수를 대상으로 인코딩을 하지 않은 경우와 target encoding을 적용한 경우에 대해서 얻은 성능을 분석하였다. 범주형 변수를 그대로 사용하였을 경우와 범주형 변수 모두에 타겟 인코딩 방식을 적용하였을 경우, 트리 깊이와 잎 노드 크기 등 하이퍼파라미터 (hyperparameter)는 각각 최적값(21, 10) 및 (13, 6)으로 설정되었다. 표 3은 인코딩 여부에 따른 두 개의 데이터세트를 학습한 RFR 알고리즘 기반 모델의 성능 결과를 비교하고 있다. 범주형 변수를 그대로 사용한 경우보다 target encoding 방식을 적용한 경우 R^2 값이 약 0.025만큼 높게 나타났고 RMSE는 약 307.76 [€]만큼 감소하여 전반적인 예측 정확도가 향상되었음을 보여준다.

표 3. 인코딩 여부에 따른 RFR 알고리즘 기반 예측 모델 성능

	Non-Target Encoding	Target Encoding
R^2	0.867	0.892
RMSE [€]	3,042.17	2,734.41

4.3 머신러닝 알고리즘 기반 모델 별 성능 결과 분석

그림 3은 target encoding을 사용한 동일한 데이터세트에 대해 학습한 MLR 알고리즘 및 RFR 알고리즘 기반 중고차 판매 가격 예측 모델의 성능을 비교한 것이다. RFR 알고리즘 기반 모델이 MLR 알고리즘 기반 모델보다 0.193의 R^2값 상승과 1,812.54 [€]의 RMSE 하락함을 보였다. 이는 본 연구에서 고려하는 데이터세트의 독립 변수가 종속 변수에 대해 단순한 선형 관계를 따르지 않기 때문으로 해석할 수 있다. 결과적으로 RFR 모델이 비선형 관계를 효과적으로 반영해 더 적합함을 확인하였다.

5. 결 론

본 논문에서는 중고차의 최종 선호 판매 가격을 예측하기 위해 Used Cars 데이터세트와 두 개의 머신러닝 알고리즘-Multiple Linear Regression (MLR)과 Random Forest Regression (RFR)-을 활용하였다. 해당 데이터세트에는 다양한 범주형 변수가 존재했으며, 이러한 범주형 변수에 대하여 적용할 수 있는 다양한 인코딩 방식에 따라 중고차 판매 가격 예측 모델의 성능에 미치는 영향을 살펴보았다. MLR 알고리즘 기반 모델에는 원핫 인코딩 (One-hot Encoding) 과 타겟 인코딩 (Target Encoding)을 각각 적용하였고, 그 결과 원핫 인코딩을 적용한 MLR 알고리즘 기반 모델이 R^2와 RMSE 값에서 더 우수한 성능을 보였다. 이와는 달리, RFR 알고리즘 기반 모델에서는 타겟 인코딩을 적용함으로써 R^2값이 증가하였고 RMSE가 감소함을 관찰할 수 있었다. 또한, 고려하는 데이터 세트에 타겟 인코딩을 적용한 RFR 알고리즘 기반 중고차 판매 가격 예측 모델이 원핫 인코딩 방식뿐만 아니라 타겟 인코딩 방식을 적용한 MLR 알고리즘 기반 알고리즘보다 성능이 우수함을 확인할 수 있었다. 이는 중고차 판매 가격에 영향을 미치는 다양한 입력 변수와의 관계가 복잡한 비선형 특징을 갖기 때문이며, RFR 알고리즘 기반 모델이 이러한 비선형 특성에 더 적합하여 중고차 판매 가격을 예측하는데 있어서 더욱 효과적인 것이다.

본 연구에서는 고려했던 데이터세트의 종속 변수인 Listing_Price의 전체 범위에 대해서 예측을 수행하였는데, 향후 연구에서는 Listing_Price의 범위를 저가, 중가 및 고가로 나눈 후 차량 판매 가격을 예측하는 모델 개발을 진행하여 예측 모델의 성능을 향상시키고자 한다.

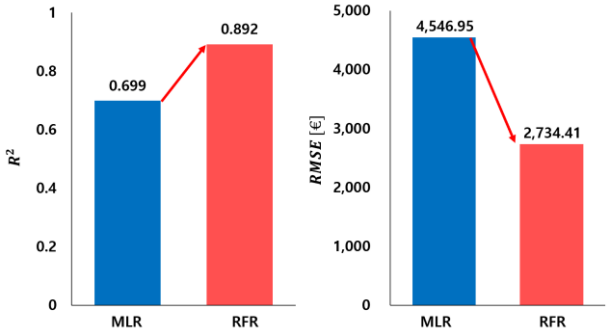

그림 3. Target Encoding 방식 적용 데이터세트를 학습한 MLR 및 RFR 알고리즘 기반 모델 별 성능

참 고 문 헌

[1] United Nations Environment Program (UNEP), “Global Trade in Used Vehicles,” UNEP, Available: https://www.unep.org/resources/report/global-trade-used-vehicles, [Accessed: Feb. 10, 2025].

[2] B. Ahn, “Construction for Safe Transaction System using Blockchain Technology (Case: Used Car),” Journal of Digital Convergence, vol. 18, no. 4, pp. 237-242, 2020.

[3] S. C. Lee, “Trust Formation of Buyers Who Perceive High Quality Risk in Online Used Car Transactions: Trust Between Buyers and Agents,” Journal of Distribution Science, vol. 7, no. 3, pp. 49-69, 2009.

[4] S. Kumar and A. Sinha, “Predicting Used Car Prices with Regression Techniques,” International Journal of Computer Trends and Technology, vol. 72, no. 6, pp. 132-141, 2024.

[5] J. He, “Predicting Vehicle Prices Using Machine Learning: A Case Study with Linear Regression,” Applied and Computational Engineering, vol. 99, no. 1, pp. 35-42, Nov. 2024.

[6] F. Wang, X. Zhang, and Q. Wang, “Prediction of Used Car Price Based on Supervised Learning Algorithm,” in Proceedings of the International Conference on Networking, Communications and Information Technology (NetCIT), pp. 143-147, 2021.

[7] The Devastator, “Used Cars,” Kaggle, Available: https://www.kaggle.com/datasets/thedevastator/uncovering-factors-that-affect-used-car-prices, [Accessed: Feb. 10, 2025].

[8] G. James, D. Witten, T. Hastie, “An Introduction to Statistical Learning,” Springer, pp. 78-79, Jul. 2013.