# Germany 중고차의 판매 가격 예측을 위한 머신러닝 기반 모델 개발

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

김지수[1], 김태윤[2], 양진주[1], 김성현[1], 이준범[3], 조석헌*
[1]숙명여자대학교, [2]금오공과대학교, [3]광주과학기술원, *University of California, San Diego (UCSD)

## Abstract

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.

## 서 론

### 1-1. 연구 필요성

- 중고차 매매 시 구매자와 판매자 간의 정보 불균형으로 인하여 구매자 뿐만 아니라 판매자도 피해를 볼 수 있는 상황이 존재
- 특히 최근에는 인터넷의 발달로 오프라인 외에 온라인 중고차 거래도 활발히 이루어지고 있음

### 1-2. 연구 목적

- 온라인 중고차량 판매자가 선호하는 차량 판매 등록 가격 예측 모델 개발
- 중고차 가격 변동에 영향을 미치는 다양한 요인 분석
- 머신러닝 알고리즘 및 범주형 변수에 적용한 인코딩 방식을 비교하여 최적의 중고차 가격 예측 모델 도출

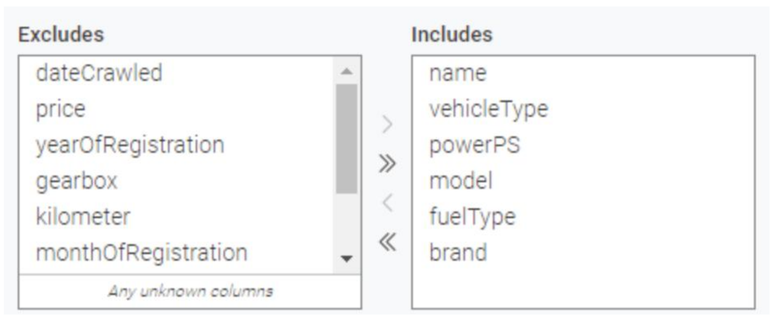
## 본 론

### 2-1. 원본 데이터셋

- 2016.03 – 2016.04 (1달): 독일 중고차 거래사이트 크롤링 데이터 사용

### 2-2. 데이터 전처리

① 분석 복잡성 감소 : 동일한 차량 매물 제거




② 데이터 오류 및 이상치 제거

- 판매 가격인 Listing_price는 500-24,500 EUR로 제한하여 극단값 제외
- 차량 출고 연월이 데이터 수집 시점(2016.03-04)보다 이전인 경우만 유지
- 마력인 Horsepower는 비정상적인 이상치를 제거하고 30-800 PS 범위로 제한하였고, 'Reihe'(기타)처럼 값을 특정할 수 없는 샘플도 제거

③ 파생 변수 생성

- 차량 출고 시점과 수집 시점 간 차이로부터 나이를 계산하여 'Car_age'를 예측 변수로 새롭게 추가

④ 범주형 변수 처리

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

- One-Hot Encoding & **Target Encoding**
- $\mathbb{1}_{\{x_j \in k\}}$: 범주형 변수의 $j$번째 샘플 $x_j$가 범주 $k$에 속할 경우 1, 아니면 0
- $y_j$: $j$번째 샘플의 실제 가격 / $n_k$: 범주 $k$에 해당하는 샘플들의 수
- Car_brand는 전체 중에서 상위 5개 브랜드(Volkswagen, BMW, Audi, Mercedes-Benz, Opel)만 유지
- Car_model도 500개 이상의 샘플이 존재하는 모델만 선택

### 2-3. 예측 데이터셋

- 총 80,946개의 데이터 샘플 수를 바탕으로 중고차 등록 가격 예측 진행
- 총 21개 변수 중 다음 주요 변수 10개를 선별하여 분석 진행

① 9개의 독립변수(Dependent Variable) : Car_brand, Car_model, Car_age, …

② 1개의 종속변수(Independent Variable) : Listing_Price

### 2-4. 알고리즘 및 성능지표

- 기계학습 알고리즘: Multiple Linear Regression(MLR), Random Forest Regression(RFR)
- 범주형 변수 인코딩 방식: One-hot encoding, Target encoding
- 성능 평가 지표: $R^2$ (결정 계수), RMSE(Root Mean Squared Error)

### 2-5. 중고차 가격 예측 모델 분석 결과

- Random Forest + Target encoding 조합이 가장 높은 예측 정확도를 기록

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

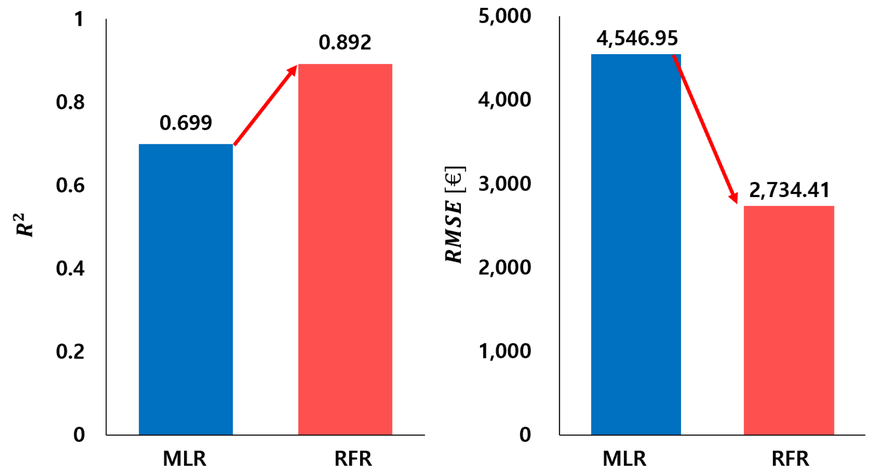
<인코딩 방식에 따른 MLR 알고리즘 기반 예측 모델 성능>

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

<인코딩 여부에 따른 RFR 알고리즘 기반 예측 모델 성능>

### 2-6. Target Encoding 방식 적용 데이터세트를 학습한 모델별 성능 비교

- RFR 기반 모델이 MLR 기반 모델보다 비선형 관계를 더 효과적으로 반영




## 결 론

- 중고차 가격 예측을 위한 머신러닝 모델에서 범주형 변수의 인코딩 방식은 핵심 요소 중 하나
- 비선형 모델(RFR)은 범주의 평균값을 반영하는 Target Encoding 방식에서 높은 성능을 보여준 반면, 선형 모델(MLR)은 다차원 표현이 가능한 One-hot encoding 방식에 적합
- 향후 연구에서는 가격대별로 구간화하여 분류 모델 기반의 추가 분석 계획