Final Report: Second-Hand Car Price Prediction

Project Overview

This project aimed to predict second-hand car prices using machine learning models trained on vehicle attributes such as brand, model, mileage, year, and derived features like mileage penalties and encoded brand/model information. The primary objective was to assess how well these factors can explain price variations and determine the most influential predictors.

Dataset Summary

- **Total Records**: 531 used car listings
- **Features**: 15 columns (including engineered features like `brand_classification_encoded`, `mileage_per_year`, `mileage_penalty`, etc.)
- **Target**: Car price (€)

Model Evaluation

🔹 Training Set Performance

- **R^2 (Coefficient of Determination)**: **0.8855**
- **Mean Absolute Error (MAE)**: **€1,505.93**
- **Average Price**: **€15,940.52**
- **MAE % of Average**: **9.4%**

➡️ **Interpretation**:
The model fits the training data very well, explaining nearly 89% of the variance in car prices. The error is relatively small (under 10% of the average price), indicating good internal consistency.

🔹 Test Set Performance

- **R^2**: **0.3976**
- **MAE**: **€3,222.20**
- **Average Price**: **€15,788.88**
- **MAE % of Average**: **20.4%**

➡️ **Interpretation**:
Performance on unseen data is significantly lower. The R^2 value of ~0.40 suggests moderate predictive power, and the error has more than doubled compared to the training set. This gap points to **overfitting**, where the model may have learned training data too well and lacks generalizability.

Top Predictive Features

Rank	Feature	Importance
1	`brand_classification_encoded`	**0.435**
2	`car_age`	0.117
3	`mileage_per_year`	0.081
4	`brand_encoded`	0.066
5	`brand_model_encoded`	0.066
6	`model_encoded`	0.061
7	`mileage_penalty`	0.058

➡️ **Interpretation**:

- **Brand classification** is the most significant predictor, highlighting that perceived brand prestige or category (e.g., economy vs. luxury) heavily influences pricing.
- **Car age** and **mileage-related features** are also strong predictors, aligning with expected depreciation patterns.

Prediction Case Studies

🔹 **BMW X5 (2011, 137,566 km)**

- **Predicted**: €15,704.47
- **Actual**: €13,490.00
- **Δ**: +€2,214.47 (model slightly overestimates)

Mileage Impact:

- 10,000 km → €37,970.42
- 90,000 km → €17,656.32
- **Depreciation**: ~€20,314

🔹 **SEAT Leon (2018, 40,250 km)**

- **Predicted**: €20,680.07
- **Actual**: €21,990.00
- **Δ**: -€1,309.93 (model slightly underestimates)

Mileage Impact:

- 10,000 km → €22,898.07
- 90,000 km → €17,107.35
- **Depreciation**: ~€5,791

➡️ **Insight**:

- Luxury cars (BMW) show **steeper depreciation** with mileage.
- Economy cars (SEAT) have **milder depreciation**, confirming the hypothesis that **brand classification affects mileage sensitivity**.

Key Takeaways

1. **High training accuracy** shows the model captures the data patterns well.
2. **Lower test performance** suggests overfitting; future iterations should include cross-validation, more data, or regularization techniques.
3. **Brand and mileage are dominant factors**, with brand classification being the strongest single feature.
4. **Mileage sensitivity varies by brand class**, supporting your hypothesis that luxury/performance cars depreciate faster per kilometer.

Potential Next Steps

- Explore **additional features**: accident history, ownership count, location-based price modifiers.