Machine Learning Project Report

Used Car Price Prediction

ESILV – Data and Artificial Intelligence Major
2025

Contents

Chapter 1

Business Scope

The goal of this project is to develop a machine learning model capable of accurately predicting the market price of a used car based on a variety of characteristics such as brand, mileage, engine power, model, fuel type, and year of manufacture. Used car pricing is a complex and highly dynamic problem influenced by numerous dependent and independent factors.

As automotive marketplaces continue to grow, precise and automated valuation models are essential for sellers, buyers, insurance companies, and dealerships. Machine learning creates the opportunity to replace subjective pricing with objective, data-driven predictions.

The choice of this subject is also deeply personal: all three members of the group have been passionate automobile enthusiasts from a very young age. We grew up comparing cars, following new releases, studying performance metrics, and debating pricing differences among brands and models. Naturally, when given the opportunity to choose a project topic, we felt drawn to a field that truly motivates us. This personal connection has allowed us to approach the project with greater curiosity and dedication.

Finally, this project aligns perfectly with our DIA major, as it requires real data preprocessing, exploratory data analysis, supervised learning, evaluation of machine learning models, and critical interpretation of results.

Chapter 2

Problem Formalisation and Methods

2.1 Problem Definition

The objective of this project is to design a predictive model capable of estimating the selling price of a used car based on its technical, mechanical, and descriptive characteristics. This is framed as a supervised regression problem, where the target variable is the vehicle's price in euros.

The problem presents several challenges:

- Non-linear relationships between features (e.g., mileage vs. price).
- Imbalanced target distribution, particularly for high-end vehicles.
- High-cardinality categorical variables such as brand and model.
- Presence of outliers which may distort model training.

2.2 Algorithm Description

Three families of models were considered:

Baseline Model — Dummy Regressor

Predicts the median price of the training set. Serves as a minimum performance benchmark.

Linear Regression

A simple parametric model used to evaluate linear relationships. Highly interpretable but unable to capture non-linearities.

Tree-Based Ensemble Methods

- **Random Forest**: Handles non-linear relationships and is robust to noise.
- **XGBoost**: A gradient boosting algorithm optimized for tabular data, excellent at modeling complex interactions.
- **CatBoost**: Efficient with categorical variables and known for strong generalization performance.

2.3 Limitations

Despite strong performance, the models have limitations:

– Difficulty interpreting ensemble models.

– Sensitivity to hyperparameters.

– Target imbalance affecting predictions of high-end cars.

– Strong dependency on data quality and preprocessing.

Chapter 3

Methodology

3.1 Data Description and Exploration

The dataset contains numerical features (mileage, engine power, age, price) and categorical features (brand, model, fuel type, transmission). Price is highly variable, from low-budget cars to premium vehicles, increasing modeling complexity.

3.2 Missing Values

Several features contained missing values:

- color (166 missing)
- power_kw and power_ps (128 missing each)
- fuel_consumption_l_100km (26,873 missing)
- mileage_in_km (62 missing)

Since these features were important and missing values were limited relative to the dataset size (except for fuel consumption), rows with missing values were removed. The column `offer_description` was dropped entirely due to its lack of predictive usefulness.

3.3 Imbalanced Data and Outliers

The price distribution is heavily right-skewed, with most cars in the low-to-mid price ranges and relatively few luxury models. This leads to:

– highly accurate predictions for common vehicles,

– increasing prediction errors for rare high-end cars.

Outliers were removed to stabilise model behavior.

3.4 Feature Encoding

Two encoding strategies were applied:

- **Target Encoding** for high-cardinality variables (brand, model),
- **One-Hot Encoding** for low-cardinality variables (fuel type, transmission, color).

These transformations were integrated in a preprocessing pipeline for consistent training and evaluation.

3.5 Data Splitting

The dataset was split into:

- **80% training data**
- **20% test data**

A fixed random seed ensured reproducibility.

3.6 Algorithm Implementation and Hyperparameters

All models were integrated into a unified pipeline combining preprocessing and regression. Hyperparameter tuning was conducted for Random Forest and XGBoost using `RandomizedSearchCV`, optimizing parameters such as:

  – number of estimators,

  – learning rate (boosting),

  – max depth,

  – subsampling ratios.

CatBoost was also tested; despite installation constraints, it performed strongly with default hyperparameters.

Chapter 4

Results

4.1 Metrics

Three metrics were used to evaluate regression performance:

- **MAE** – average absolute error in euros,
- **RMSE** – penalizes large errors,
- **R^2** – proportion of explained variance.

4.2 Overfitting, Underfitting, and Imbalance

The optimized XGBoost model achieved:

- Train MAE: 278 €
- Test MAE: 457 €
- Train R^2: 0.9978
- Test R^2: 0.9925

Close train–test performance indicates no overfitting. Error-vs-price plots confirmed that errors increase for high-end vehicles due to target imbalance.

4.3 Evaluation Compared to Baseline

Dummy Regressor

MAE: 9704 €, RMSE: 12596 €, R^2: –0.02

Linear Regression

MAE: 3877 €, RMSE: 5586 €, R^2: 0.80

Random Forest

MAE: 431 €, RMSE: 1400 €, R^2: 0.99

XGBoost (Best Model)

MAE: 439 €, RMSE: 1031 €, R^2: 0.99

CatBoost

MAE: 566 €, RMSE: 1086 €, R^2: 0.99

Chapter 5

Discussion and Conclusion

Several challenges were identified during this project:

- Non-linear relationships between key features,
- High-cardinality categorical features,
- Outliers distorting error metrics,
- Imbalanced price distribution,
- Overfitting potential in ensemble models.

Tree-based ensemble methods significantly outperformed simpler baselines. **XGBoost emerged as the best-performing model**, delivering the strongest generalization and lowest prediction error.

Future improvements could include:

- expanding the dataset,
- using SHAP for interpretability,
- applying deep learning models,
- refining outlier management for luxury vehicles.

Overall, the project successfully met its objective. Beyond the technical achievement, this work allowed us to combine machine learning knowledge with a genuine passion for the automotive domain, resulting in a meaningful and motivating analytical experience.