Machine Learning Project: Deliverable 1

"Cars 4 You" Price Prediction Pipeline

Master in Data Science and Advanced Analytics
Group 34

Carolina Luz | 20250409
Margarida Quintino | 20250411
Pedro Castro | 20250467
Pedro Carrasqueira | 20250488

1 Introduction

This homework serves as a preparatory step for the final project "Cars 4 You: We Buy Your Car!", aiming to guide the full machine learning process – from data exploration to model evaluation. The project's goal is to derive actionable insights from structured data containing various vehicle and consumer attributes, ultimately enhancing the accuracy of car pricing models and recommendation strategies.

2 Pipeline Structure

Our pipeline is a sequential, end-to-end workflow designed to ensure data integrity and prevent leakage. It transforms raw data into a format suitable for modeling and evaluation.

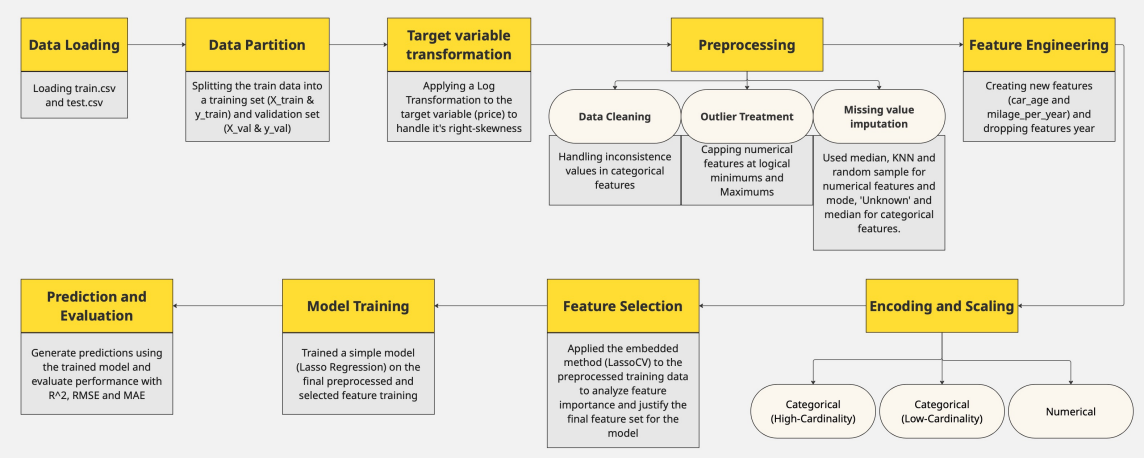

Figure 1: Schematic representation of the pipeline

3 Preprocessing and Feature Engineering

During data cleaning, we inspected unique values, focusing on categorical features like `brand`, `transmission`, `fuelType`, and `model`; corrected typos; and standardized naming for consistency. Impossible negative values were found in numerical columns such as `mileage`, `tax`, `mpg`, `engineSize`, and `previousOwners`, and were considered data-entry mistakes and converted to `NaN`. The `year` feature included float values that were rounded and converted to integers, while invalid or non-integer values in `previousOwners` were also treated as missing.

After cleaning, missing values were imputed using feature-specific strategies. Numerical variables with few missing values and skewed distributions (e.g. `year`, `mileage`, `engineSize`, `previousOwners`) were imputed using the median to be robust to outliers. Variables with more missing data and stronger local correlations (e.g. `tax`, `mpg`) were imputed with KNN to obtain context-aware estimates. Random sampling was applied to maintain the `paintQuality%` distribution. Categorical features such as `brand`, `model`, `transmission`, and `fuelType` were considered Missing Not At Random (MNAR), and a new "unknown" category was added rather than using the mode. Missing values in the binary variable `hasDamage` were replaced with 0, presuming that the majority of cars had no damage. To prevent data leakage, all imputations were fit on the training data and applied to both training and validation sets. The resulting dataset was examined to ensure that distributions remained consistent after imputation.

Next, we performed feature engineering to boost predictive power. We derived an `age` feature based on the `year` variable to reflect depreciation and simplify the pricing link, as well as a `mileage_per_year` feature to capture usage intensity and assess wear and tear. The target variable `price` is right-skewed, with a long tail of expensive cars that could bias linear models. To normalize the distribution, a logarithmic transformation was used, producing the new variable `log_price`, which is more symmetric and stable and therefore suitable for regression modeling.

Exploratory Data Analysis (EDA) was performed on the predictor features to understand their distributions and relationships with `log_price`. The analysis revealed sensible patterns: newer cars and those with larger engines tend to be more expensive, while age and usage are primary factors reducing price.

Before model training, a preprocessing pipeline (e.g. using scikit-learn's `ColumnTransformer`) was built to handle categorical encoding and numerical scaling. Because machine learning models require numerical input and features vary in scale, the transformer applies the proper transformations to each feature type. Before one-hot encoding, high-cardinality features like `brand` and `model` were grouped into an "Other" category, while low-cardinality variables were encoded directly. This systematic preprocessing guarantees consistent and leak-free transformations in both training and validation data.

4 Feature Selection

We used LASSO Regression (Least Absolute Shrinkage and Selection Operator) for model selection because it naturally performs embedded feature selection by shrinking less relevant coefficients to zero, which improves model simplicity and interpretability. The model was trained with `LassoCV`, which automatically adjusts the regularization parameter. The final `lasso_pipeline`, which included the preprocessor and LASSO model, was trained end-to-end, achieving strong results with a Validation RMSE of 0.2136 (log scale), equivalent to approximately 6,848.78 on the original price scale.

We assessed feature relevance by extracting model coefficients and mapping them to feature names (including those generated by one-hot encoding). This allowed us to identify which features the model retained (non-zero coefficients) and which it discarded. To study feature significance further, we compared models trained using only the original variables (`year`, `mileage`) with models using only engineered features (`age`, `mileage_per_year`). The findings indicated that `age` is a stronger predictor of depreciation than `year`, while `mileage` often captures wear better than `mileage_per_year`. Consequently, the final model retained `age` and `mileage` and dropped the other two.

The final LASSO model's feature importance analysis revealed that engine size is the most influential positive price driver, followed by premium brands (e.g. Mercedes, Audi, BMW, VW) and automatic transmission, all of which increase car value. In contrast, age remains the strongest negative predictor, indicating depreciation over time. The final model is accurate, interpretable, and consistent with market behavior.