MACHINE LEARNING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Handout

ML Project

Group 55

Adrian Radoi – 20250353

Koen van der Heijden – 20250527

Pedro Xavier – 20250476

Tiago Anastácio – 20250497

Fall Semester 2025

TABLE OF CONTENTS

1. Introduction

We built a supervised predictive model to estimate used-car prices. The dataset contained roughly 75,000 labeled listings with technical and market attributes (like engine size, number of previous owners, mileage, brand, model) and a separate test set of about 32,000 cars without prices. Our goal was to design a robust pipeline – data cleaning, preprocessing, feature engineering, feature selection, modeling and evaluation – that would yield accurate predictions on unseen cars. To the test set we applied the pipeline created from training.

2. Explore and Pre-process the data

We began with exploratory analysis to understand distributions, relationships with *price* and overall data quality. Most cars fell between 2015 and 2020, indicating a relatively modern fleet. *Mileage* was strongly right skewed, while the variable *hasDamage* was constant at 0 and therefore carried no signal. Bivariate checks aligned with expectations: *mileage* displayed a negative association with *price* and *engineSize* a positive one, whereas *paintQuality%* showed no stable pattern.

These observations informed our preprocessing plan. For outliers, we used the IQR method: values below Q1−1.5·IQR were capped at that limit, and values above Q3+1.5·IQR were reduced to it. This stabilizes linear models by preventing a handful of extreme points from dominating the loss, at the cost of some loss of fidelity when predicting outliers for the test set. We also corrected impossible or clearly invalid entries using simple rules – for example, *engineSize* equal to 0 was marked as null, while for example when *paintQuality%* above 100% were set as 100% – and standardized categorical text (case, spacing) before mapping brand/model variants to canonical forms so that categories reflected genuine groups rather than spelling variations. Missing data were imputed in a structured way: for categorical variables we used the mode within similar groups of cars (typically defined by Brand/Model/Year), and for numerical variables we used the median within comparable groups. This strategy retains as much information as possible. Before imputing the null values we split our training data into training and validation sets so we could train the model on the training set and only after evaluating the corresponding results on the validation set.

We engineered two features to improve linearity and interpretability. First, we derived *age* from *year* to make interpretation easier. Second, we applied a logarithmic transform to *mileage* to reduce its strong right skew and make its relationship with *price* closer to linear, which is advantageous for regularized linear models. We then one-hot encoded categorical variables and we applied the standard scaler to numerical features to transform their distribution in standard normal distributions of mean 0 and variance 1. Scaling ensures fair coefficient regularization and prevents large-scale features (like raw *mileage* or *tax*) from dominating simply because of their units.

3. Feature Selection

Feature selection combined statistical screening and model-based methods. For categorical variables, a chi-square (χ^2) test indicated that *hasDamage* had no detectable association with price, consistent with its lack of variation. For numerical variables, correlation diagnostics flagged collinearity among *year*/*age*, between *mileage*/*year*, and between *mileage*/*log_mileage*, implying these pairs should not be included simultaneously. We then applied Recursive Feature Elimination (RFE) with a linear estimator to iteratively remove weak predictors, which suggested that nine numerical features carried signal overall. Finally, we used Lasso (L1-regularized regression) both as a selector and as our modeling approach under multicollinearity. Lasso retained year over age and favored *log_mileage* over raw *mileage*, shrinking *previousOwners* and *age* toward zero, behavior consistent with the correlation diagnostics and with L1's tendency to pick one representative from a correlated group.

4. Results and Future Improvements

The final model was a Lasso regression trained on the cleaned, encoded, and scaled data with the feature set informed by the steps above. We evaluated performance on a held-out validation split using R^2 and Mean Absolute Error (MAE). The model achieved an R^2 of about 0.84, meaning it explained roughly 84% of the variance in *price*, and an MAE of approximately 2,404, which indicates that on average predictions differed from actual prices by about €2,400 per car. After selecting this model, we applied the identical preprocessing to the test set and generated predictions. As is typical when moving from validation to unseen data, MAE increased, reflecting distributional differences and the presence of atypical vehicles in the new dataset.

To further improve predictive performance, we plan to broaden both the modeling and preprocessing search space. On the modeling side, we will compare Ridge, Elastic Net, Decission Tree and Random Forest to test alternative regression structures under our specific collinearity patterns. We will automate hyperparameter search with GridSearchCV and RandomizedSearchCV to systematically explore regularization strength and, where relevant, model-specific settings. On the preprocessing side, we will experiment with alternative scalers (such as Min-Max Scaler) and different imputation strategies for missing values for variables with structured missingness. Additional feature engineering, interaction terms such as *mileage* × *age*, alternative monotonic transforms and selected polynomial terms, may further improve linear fit without overfitting the model. Finally, we will modularize the pipeline to make experimentation easier and fully reproducible, ensuring that all transformers are fit on training data only. This disciplined approach should reduce variance between folds, improve generalization and facilitate efficient iteration toward a higher-performing model.

5. Appendix – Pipeline Representation

```
EXPLORATORY DATA ANALYSIS
│
├── Univariate distributions
│     • Example: mileage shows right-skew
│
├── Bivariate relationships with price
│     • Mileage ↓ → Price tends to decrease
│     • Engine size ↑ → Price tends to increase
│
└── Data quality assessment
      • Missing values (NaNs)
      • Outlier detection
      • Constant features (such as 'hasDamage = 0')

PREPROCESSING
│
├── Outlier treatment
│     • Clean ouliers with IQR caps: [Q1 − 1.5·IQR, Q3 + 1.5·IQR]
│
├── Impossible or unrealistic values
│     • Example: engineSize = 0 → NaN
│     • Example: paintQuality% > 100 → capped at 100
│
├── Categorical normalization
│     • String cleaning and standardization
│     • Mapping brand/model variants to consistent names
│
├── Missing value imputation
│     • Categorical → mode within (Brand / Model / Year) groups
│     • Numerical   → median within similar groups
│
└── Feature engineering
      • age = reference_year − year
      • log_mileage = log(mileage) to reduce right skew

ENCODING & SCALING
│
├── One-hot encoding for categorical variables
└── Standardization (mean = 0, std = 1) for numerical variables

FEATURE SELECTION
│
├── Correlation diagnostics
│     • Collinearity checks (e.g., year vs. age, mileage vs. log_mileage)
│
├── RFE (Recursive Feature Elimination)
│     • Model-based selection on numerical variables
│
└── Lasso regression
      • Regularization + sparsity
      • Handles multicollinearity

MODELING & EVALUATION
│
├── Final model: Lasso Regression
├── Evaluation metrics on validation:
│     • R² (explained variance)
│     • MAE (mean absolute error)
│
└── Refit model on training data and apply to prepared test set
```