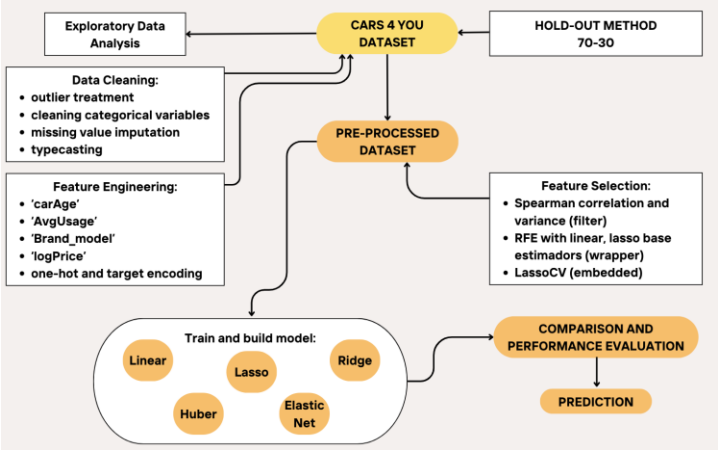
Machine Learning 2025/2026 – Deliverable 1

Andreea Roica (20250361), Barbara Franco (20250388),
Beatriz Varela (20250367), Marisa Esteves (20250348)

Our Pipeline



Pre-processing

Treatment of Numerical Variables: The first step in cleaning the data was replacing the inconsistent numerical values such as negative values in *mileage*, *previousOwners* and *tax*, years above 2020, negative or zero *mpg* and *engineSize* and *paintQuality*% above 100, as NA, since they are impossible and for this project it is not recommended to drop observation due to the Kaggle submission. For *mpg* and *engineSize* lower limits of 40 miles and 1 L, respectively, were imposed. *year* and *previousOwners* were rounded to the nearest lower integer. The remaining variables were rounded to 2 decimal points, except for *hasDamage* which was typecast into boolean. In this case, it was considered that zero values reflect no damage while NA values do the opposite. Categorical variables were first typecast into strings, stripped of leading and trailing spaces and converted to uppercase.

As to prevent data leakage, the following steps were done after applying the hold-out method to the train set. A 70-30 train-test split was used for this, originating new train and validation sets.

Treatment of Categorical Variables: The *fuzzywuzzy* librabry was used to group the categorical strings into clusters based on a similarity threshold, assigning them the most frequent value in their cluster. The clusters were created based only on the train dataset. *models* had to be corrected using *get_close_matches* from the *difflib* library, but the logic applied was the same. Categorical variables in the validation and test sets, were corrected according to those. In the case that a value was not found in any cluster, its similarity with the representative values (i.e., the most frequent value of each cluster) was computed. The value was then replaced with the most similar match if the similarity score exceeded the defined threshold; otherwise, it remained unchanged.

Outliers: Most variables have under 5% outliers while *tax* has 26%. Winsorization was applied to all the sets at either the 1st percentile (for lower-bound IQR outliers, for *year*), the 99th percentile (for upper-bound IQR outliers, for *mileage* and *mpg*), or both (for *tax*), affecting 1-2% of each variable. Percentiles were calculated based only in the train dataset. This method was chosen as to preserve most outliers, which represent genuine car characteristics, while still addressing extreme values.

Missing values: In first place, a mapping of the models to brands was created. For the brands that were missing but had a corresponding model that was present in the mapping, then the corresponding brand was used to replace the missing value. If the model is missing, the global mode of the column *Brand* of the train set is returned.

The treatment of missing values of the column *model* involved creating 5 different mappings: mapping_6 : Uses *Brand*, *fuelType*, *engineSize*, and *transmission*; mapping_5: Uses *Brand*, *fuelType*, *engineSize*; mapping_4 : Uses *Brand* and *fuelType*; mapping_3 : Uses *Brand* and *transmission*; mapping_2: Uses *Brand* and *engineSize*; mapping_1: Uses *Brand*. Each mapping uses the model's mode for the given feature combination. mapping_6, the most detailed, is the preferred choice, however, in the occurrence of missing values in those features, another mapping may be needed (descending order). Since the column *Brand* is free of missing values, this method is adjusted for all possibilities. These features were chosen since *Brand* and *model* are highly correlated according to Cramers V and the other ones are characteristics of the car model that are unchangeable.

year and *mileage* are highly correlated features, so they were be used to input each other's missing values. Bins of 'low', 'average', 'high' and 'very high' milage were created, followed by the imputation of the median of the year of each bin. To account for observations where both year and mileage are missing, the imputation based on the most correlated features (first *tax*, then *mpg*). If they are also missing, the median of year is returned. For *mileage*, the train dataset was grouped by year and for each group the median of mileage was computed.

For *tax*, the cars were grouped by *model* and *year* followed by the imputation of the median of *tax*, as cars of the same model tend to have similar road tax values and age of the car might also be a factor. For cars that are the only model in a certain year, the median of the tax of the specific model is returned. When there are no observations of the same model, the median of *tax* of the corresponding year is returned. The same logic was applied to the columns *previousOwners* and *paintQuality*% but with the mode instead of the median.

For *fuelType*, *mpg*, *engineSize* and transmission the missing values were replaced by the column's mode of the corresponding *model*. If the *model* is unique then the corresponding *Brand* is used. This method was applied because cars of the same model tend to have the same characteristics. Finally, since all the missing values from *hasDamaged* were interpreted as the car having damage, the column is free of null entries.

Feature Selection

Four new features were engineered: '*Brand_model'*, combines *brand* and *model* into a single string; '*carAge'*, age of the car; '*AvgUsage'*, *mileage* and *age* ratio; '*logPrice'*, a logarithmic transformation of the target. More variables came from the encoding of categorical ones. Target encoding can improve model performance by incorporating target information directly into the feature without increasing the input space, so it was used in '*Brand_model'* and '*Brand'* due to the large number of categories. For the other variables one-hot encoding was used. Previous to feature selection, variables were scaled in all datasets using *StandardScaler*.

Various feature selection methods were employed in the train dataset and compared (Table 2 in the Annex). For filter methods, variance and Spearman correlation were considered. *year* and *model* were dropped as they're perfectly correlated with *carAge* and *Brand_model*, which were kept for interpretability and completeness. *AvgUsage* is marked "keep?" because, if retained, it preserves *mileage* information without redundancy, though it is weakly correlated with the target. *EngineSize* is also marked 'keep?' since it is neither strongly correlated with the target nor with other features. For wrapper methods, RFE was used along with linear and Lasso regression base estimators. To maintain a balance between precision and complexity, we chose the number of features at which the score stabilizes. For embedded methods, Lasso with cross-validation was used. Regarding the one-hot encoded features, all categories were kept, for now. The selected features were: *mileage*, *engineSize*, *carAge*, *AvgUsage*, *fuelTyper_cleaned_encoded*, *transmission_cleaned_encode* and *Brand_model_encoded*.

Linear, Ridge, Lasso, Elastic Net and Huber models were employed using the selected features. Huber Regressor was elected as the best model, since MAE, MedAE, and Adjusted R^2 were prioritized. The distribution of error (true value – predicted value) is approximately cantered around 0 and bell-shaped, but it is right-skewed (Fig. 3 in Annex), indicating outliers that correspond to very high prices that were substantially underestimated by the model.

Refer to the accompanying Jupyter notebook for detailed data exploration and additional insights.

Annex

Table 1 - Missing value imputation techniques used for each variable (refer to notebook for exceptions).

'year'	Creation of 'mileage' bins and imputation of the median of the corresponding bin.
'mileage'	Median 'mileage' of car's 'year'.
'mpg'	Median from group of cars sharing the same 'model'.
'engineSize'	Median from group of cars sharing the same 'model'.
'paintQuality%'	Median from group of cars sharing the same 'model' and 'year'.
'previousOwners'	Median from group of cars sharing the same 'model' and 'year'.
'tax'	Median tax of cars grouped by 'model' and 'year'.
'Brand'	The car's main brand (e.g. Ford, Toyota).
'model'	Imputation according to mode in the biggest group possible of categorical variables in a fallback system.
'transmission'	Mode of the corresponding 'model'.
'fuelType'	Imputation of mode from group of cars sharing the same 'model'.

Table 2 - Feature selection results using different methods.

Predictor	Variance	Spearman	RFE Linear	RFE Lasso	Lasso CV	What to do?
'year'	keep	discard	keep	keep	keep	discard
'mileage'	keep	discard	keep	keep	keep	keep
'mpg'	keep	discard	discard	discard	keep	discard
'engineSize'	keep	keep?	keep	keep	keep	keep
'paintQuality%'	keep	discard	discard	discard	keep	discard
'previousOwners'	keep	discard	discard	discard	keep	discard
'tax'	keep	discard	discard	discard	keep	discard
'hasDamage'	keep	discard	discard	discard	discard	discard
'carAge'	keep	keep	keep	discard	keep	keep
'AvgUsage'	keep	keep?	discard	discard	keep	keep
'Brand_cleaned_encoded'	keep	discard	discard	discard	keep	discard
'model_cleaned_encoded'	keep	discard	keep	keep	keep	discard
'fuelType_cleaned_encoded'	---	---	---	---	---	keep

‘transmission_cleaned_encode’	---	---	---	---	---	keep
‘Brand_model_encoded’	keep	keep	keep	discard	keep	keep

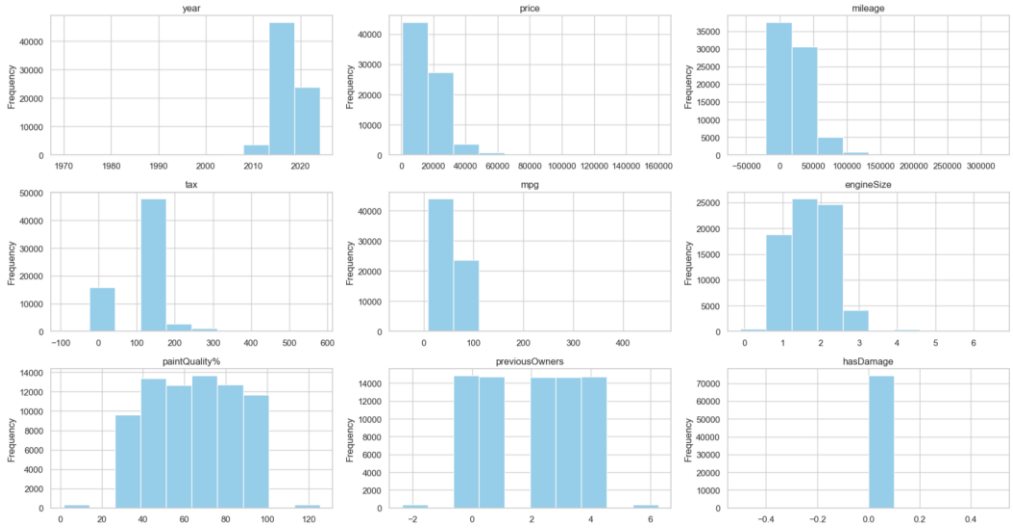


Figure 1 - Histograms of numerical variables.

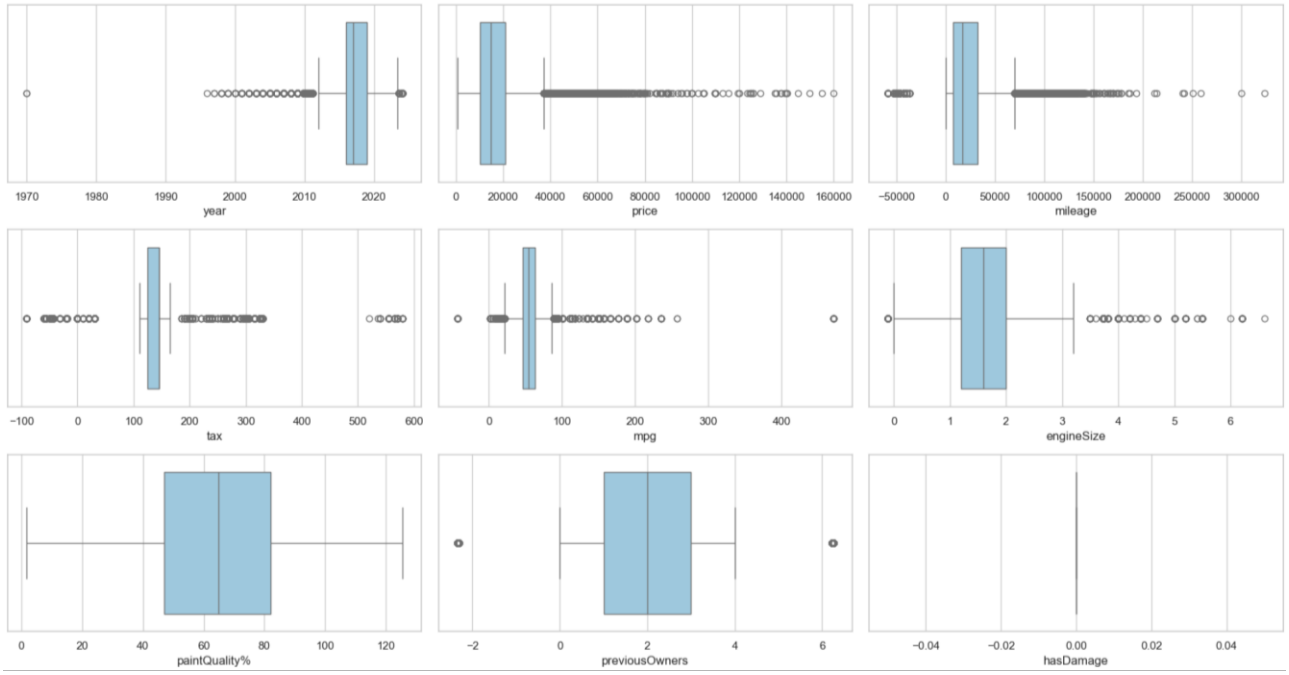


Figure 2 - Boxplots of numerical variables.

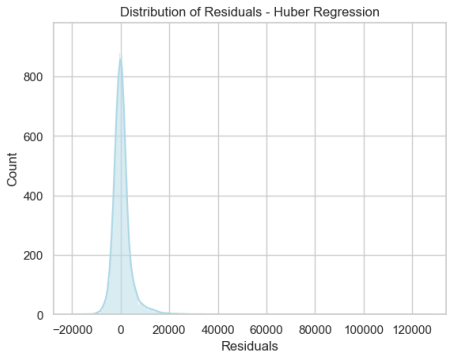


Figure 3 - Huber regressor residual distribution.