Cars 4 You: Expediting Car Evaluations with ML

Group Project Machine Learning 2024/2025

GROUP 12 - HANDOUT

Project members:
- Ricardo Isidro - 20250374
- Rodrigo Santos - 20250387
- Rodrigo Texieira - 20250393

Overall structure of our project pipeline:

1. Data Import & Exploration

Training and ***test*** datasets are imported using pandas. Exploratory Data Analysis (EDA) includes **descriptive statistics**, **correlation heatmaps**, and **distribution plots** to identify **trends**, **missing values**, and **outliers**.

---

2. Preprocessing & Data Cleaning.

- **Dropped Columns:** ***hasDamage*** was removed due to a single value (**0**), providing no predictive information.
- **Text Standardization:** All categorical columns (***Brand***, ***model***, etc.) were converted to lower-case to consolidate unique values and fixed typos using **fix_typos()**.

2.1 Outlier Handling: **Negative**, **unrealistic**, and **outlier numeric values** were set to ***NaN*** using boxplot analysis. Outliers were removed only for **Linear Regression**, since **Random Forest** is **robust** to them.

- ***mileage*** < 0
- ***tax*** not in range [0, 400]
- ***mpg*** not in range [0, 150]
- ***engineSize*** not in range [1, 6]
- ***paintQuality%*** > 100
- ***year*** not in range [1990, 2020]
- ***previousOwners*** not in range [0,4]

The **Linear Regression** model is extremely sensitive to outliers. Therefore, removal and clipping were crucial:

- **Unrealistic Values (e.g., *mileage* < 0, *paintQuality%* > 100, *mpg>150*)**: These were removed to prevent **distortion of the regression coefficients**, ensuring the model learns valid and logical relationships.
- **Extreme Outliers (Interval Clipping)**: The removal of very rare/extreme values (e.g., tax, mpg, year, engineSize) was performed because these points would have **high leverage**, pulling the regression line and violating the assumption of **linearity** in the main sample.
- **Percentile Clipping (*tax*, *engineSize*)**: Limiting to the 1st/99th percentile **before** logarithmic transformation to **stabilize the model**, reducing the extreme variance introduced by milder outliers.
- **Log-Transformation (np.log1p)**: Applied to the target (price) and skewed variables (mileage, mpg and tax), this is essential for Linear Regression because it:
  - **Normalizes distributions**, bringing them closer to the symmetry required for the assumption of **normal residuals** (model error).
  - **Linearizes relationships** (e.g., log(price) vs. log(mileage).
  - **Reduces Heteroscedasticity** (disproportionate errors at high price values).

2.2.1 Categorical Missing Values

Theil's U (Asymmetric Uncertainty) was used to guide the selection of auxiliary variables, ensuring contextual imputations:

- **Contextual Imputation (*Brand*, *model*):** A custom helper function (***fill_NaN_with_categorical***) was used to fill NaNs with the group mode from the most correlated group. For example, ***Brand*** was inferred based on the combination of ***model***, ***transmission***, and ***fuelType***, preserving feature dependency.
- **Simple Imputation (*transmission*, *fuelType*):** NaNs were filled using the global mode of the respective column.

2.2.1 Numerical Missing Values

The Correlation Ratio R^2 scores were used to identify the best predictors for numerical variables:

- **Mixed and Contextual Imputation (*year*, *mileage*, *tax*, *engineSize*):** A mixed-method function (***fill_NaN_with_mixed***) was used, which fills the NaN with the mode within a group defined by one categorical variable (model) and one highly correlated, discretized (*binned*) numerical variable. This highly granular approach maintains the data dependency structure.

- **Robust Median Imputation (*paintQuality%*, *previousOwners*):** The median was used to fill NaNs. The median is a robust statistic that minimizes the impact of imputation on skewed distributions or those with residual outliers.

---

3. Feature Engineering

- **Categorical Encoding:**
  - ***One-Hot Encoding*** applied to ***Brand***, ***transmission***, and ***fuelType*** to convert nominal categories into binary indicator variables.
  - **K-Fold Target Encoding** was applied to the ***Brand*** and ***model*** variables due to the **high cardinality** of the *model* feature. This approach allows the model to effectively capture price-related information (average price patterns by Brand and model) while **reducing overfitting** through cross-validation.
  - **Min–Max normalization** applied to numerical features to standardize scales between 0 and 1.It offers stability to the model and is ideal when the data distribution is already well-behaved (without outliers).

---

4. Feature Selection Strategy

Feature relevance was evaluated using both statistical and model-driven methods:

- **Pearson Correlation**: Used to evaluate linear relationships between numerical variables and the target (price), as well as to identify and remove multicollinear predictors that could distort regression coefficients.
- **Recursive Feature Elimination (RFE):** Performed using **Linear Regression** model to iteratively rank and select the most informative subset of predictors based on their contribution to model performance.
- **LassoCV:** Applied to penalize less relevant predictors through L1 regularization, automatically setting weak feature coefficients to zero and improving model generalization.

Final retained features (used in both models): ***year***, ***engineSize***, ***tax***, ***mpg***, ***mileage***, ***Brand***, ***model***, and ***transmission***.

---

5. Model Training and Evaluation

- The task was formulated as a supervised regression problem.
- Linear Regression served as a baseline model for interpretability and benchmarking.
- Random Forest Regressor was employed as the final model to capture nonlinear dependencies and feature interactions.
- Evaluation Metrics:
  - **R² Score:** Measures the proportion of variance explained by the model.
  - **Mean Absolute Error (MAE):** Represents the average prediction error.
  - **Root Mean Squared Error (RMSE)**: Highlights larger deviations more heavily.

Pipeline overview table

Stage	Techniques Used
Exploration	Descriptive statistics, histograms, correlation matrix, missing value analysis
Preprocessing	Outlier filtering (IQR & quantile clipping), missing value imputation (median or inferred), typo correction, data type adjustments
Feature Engineering	*One-Hot Encoding and* K-Fold target encoding (for categorical variables), log transformation for skewed numeric variables, Min–Max normalization
Feature Selection	Pearson correlation,Recursive Feature Elimination (RFE), LassoCV
Modeling	Linear Regression (baseline) and Random Forest Regressor (final model) using 70/30 train–validation split
Evaluation	R², MAE, RMSE — Random Forest achieved better generalization and lower error

- Results: The Random Forest Regressor consistently outperformed Linear Regression across all metrics, demonstrating better predictive accuracy, robustness, and generalization to unseen data.