MACHINE LEARNING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Car Pricing Model for Cars 4 You

Group 37

Filipa Pereira, 20240509

Gonçalo Silva, 20250354

Marta La Feria, 20211051

Tomás Coroa, 20250394

Fall Semester 2025-2026

ABSTRACT

This project was developed for **Cars 4 You** to accelerate their car evaluation process, which is currently slowed by manual inspections and leads potential sellers to competitors. The primary goal was to build a robust regression model to accurately predict a car's resale price based on seller-provided features. Due to the presence of outliers, the modeling focus was set on predicting the **median price**, using the **Mean Absolute Error (MAE)** as the primary loss function, which aligns with the Kaggle competition's metric.

The methodology began with extensive data preprocessing, including cleaning categorical inconsistencies using fuzzy matching and a validated dictionary built with the **wheel-size.com API**. Logical errors like negative numerical values were corrected, and missing data was imputed using the **KNNImputer** for MAR variables. Predictive power was enhanced by engineering features like *age* and *miles per year*. Categorical features were prepared using **Frequency Encoding** and **One-Hot Encoding**, and data was scaled using **RobustScaler** to mitigate the influence of legitimate outliers. Features were then selected using an aggregate consensus approach combining Filter, Wrapper, and Embedded methods to retain only the most stable and predictive subset.

After benchmarking several regression models, the **Gradient Boosting Regressor** achieved the best overall performance. The final model demonstrated satisfactory generalization performance, achieving a validation MAE of **1,414.71** and a Pinball Loss ($\alpha = 0.5$) of **707.36**, indicating good accuracy without significant overfitting. This solution directly addresses business needs. Future work will explore leveraging Pinball Loss with $\alpha < 0.5$ to deliver conservative estimates, minimizing the risk of overpaying and enhancing profit protection.

TABLE OF CONTENTS

Abstract ii
Introduction 1
Data Exploration and Preprocessing 1
Inconsistencies 1
Missing Values 1
Outliers 1
Feature Engineering 1
Encoding Categorical Variables 1
Data Scaling 2
Feature Selection 2
Modelling & Evaluation Metrics 2
Benchmarking and Final Model 2

Introduction

The primary business need for Cars 4 You is to accelerate the car evaluation process, as current manual inspections cause delays and lead potential sellers to competitors. This project addresses this by developing a robust regression model to accurately predict a car's resale price based on seller-provided features. The **holdout** method was used due to its simplicity and allowing for a quick evaluation of the models' performance. This choice is supported by the fact that we have a large data set with similar distributions between the training and test sets.

Data Exploration and Preprocessing

The project began with a comprehensive **Exploratory Data Analysis**, meticulously checking data types, distributions, misspellings, impossible values, and exploring multivariate relationships. Key insights from this full assessment were documented in the Notebook. To ensure a clean foundation, duplicate rows (entries with identical features, regardless of the price) were subsequently removed. Crucially, to prevent data leakage, the data was then **split** into an **80/20** holdout for training and validation before any data preparation. The split used shuffling to remove any possible ordering bias and a fixed random state for reproducibility.

Inconsistencies

Categorical inconsistencies were standardized using **fuzzy string matching** (similarity-based correction) against their respective valid metadata values. For brand and model, we leveraged the **wheel-size.com API** to validate and correct all combinations against real-world data. For numerical features, non-physical percentages were clipped to the nearest valid limit (0 or 100). Furthermore, impossible negative values were converted to absolute values after we confirmed their prices aligned closely with their positive counterparts, suggesting sign errors.

Missing Values

Categorical missings were replaced with "**unknown**" since the absence of information may itself carry predictive value. For numerical variables, missingness was statistically classified as MCAR or MAR using t-tests. **MCAR** values were imputed with the **median/mean**, while **MAR** values were handled using the **KNNImputer** with $k = 5$ to preserve multivariate relationships.

Outliers

Outliers were detected during EDA using **IQR** and **MAD** (a robust Z-score equivalent). Neither removal nor winsorization was deemed appropriate because these extreme values reflect real market variability, and the distribution consistency between train and test sets supported retaining them. Instead, to mitigate their influence, we employed outlier-resistant techniques, specifically **Robust Scaling** and naturally robust models like **Random Forest**. For model outliers, rare models (and any potential new models appearing in the validation and/or test set) were grouped into a single "**other**" category, creating a more stable and generalized feature.

Feature Engineering

New features were engineered to enhance predictive power, such as **age** (providing a direct relationship with price); **miles per year** (capturing a car's usage intensity) and **brand-model** (capturing the interaction between a brand and its model). Given the creation of these new features, the variables model and year were dropped, since keeping them would introduce the issue of multicollinearity.

Encoding Categorical Variables

Two encoding techniques were applied depending on feature cardinality. For **high-cardinality** variables (brand-model and brand), we used **Frequency Encoding**. This method effectively handles many unique categories without increasing dimensionality, thus avoiding the curse of dimensionality. It also encodes category popularity, which can be a useful predictive signal for the model. For remaining categorical variables (transmission and fuel type), **One-Hot Encoding** was adopted. It offers advantages such as simplicity, straightforward interpretability, and the explicit representation of each category, which can be vital for models to capture their distinct effects accurately.

Data Scaling

For data scaling, **RobustScaler** method was used. This technique centers data by removing the median and scaling it based on the IQR, making it highly resistant to outliers and well-suited for the dataset's skewed features. This choice aligns with our overall strategy of retaining valid outliers while mitigating their influence on the model's training process. Although it may slightly compress informative variance, we believe its robustness is important for modeling price prediction accurately.

Feature Selection

Our comprehensive feature selection strategy employed a multi-faceted approach, combining Filter Methods (**Spearman correlation** and **Variance Threshold**), Wrapper Methods (**Recursive Feature Elimination** with both **Logistic Regression** and **Random Forest**), and Embedded Methods (**Lasso** and **Ridge** regression). To maximize model performance and interpretability, an aggregate consensus approach was applied: only features recommended by a majority (at least 4 out of 5 methods, or 4 out of 4 for binary ones) were retained. The final, reduced feature set, comprising the most stable and predictive attributes across models, included: *mileage*, *tax*, *mpg*, *engineSize*, *age*, *Brand_freq_enc*, *Brand_model_freq_enc* and *transmission_manual*.

Modelling & Evaluation Metrics

The project addresses a **supervised regression problem**, where the objective is to model the relationship between vehicle characteristics and the continuous numerical target variable, price. Given the business context, we believe predicting the **median price** is preferred, as car prices contain outliers and extreme values that can distort mean-based estimates, making the median a more robust and representative prediction for typical vehicles.

Consequently, median-based loss functions were adopted. The primary metric for optimization and evaluation was the **Mean Absolute Error** (MAE), which is strictly consistent for the median and aligns perfectly with the Kaggle competition's evaluation metric. Additionally, the **Root Mean Squared Error** (RMSE) was monitored to track the influence of larger deviations. We also utilized the **Pinball Loss** (Quantile Loss), which, while equivalent to MAE at $\alpha = 0.5$, but holds significant future potential. Given that Cars 4 You prioritizes **minimizing the risk of overpaying** (i.e., preferring to under-predict the value rather than over-predict it to protect profit margins), future work will explicitly explore Pinball Loss with $\alpha < 0.5$.

Benchmarking and Final Model

The following models were tested: **Gradient Boosting**, **Random Forest**, **K-Neighbors**, **Decision Tree**, **Support Vector Regression (SVR)**, **ElasticNet**, and **Linear Regression**. For each model, six different hyperparameter combinations were evaluated (except for the deterministic Linear Regression) to ensure fair comparison. Among these, **SVR demonstrated the lowest overfitting** and proved to be the most robust under this holdout split. However, the **Gradient Boosting Regressor** achieved the best overall performance across all validation metrics, leading to its selection as the final predictive model for the Kaggle competition.

The Gradient Boosting model demonstrated satisfactory generalization performance, achieving a validation MAE of **1,414.71** and a training MAE of 1,228.13, along with a validation RMSE of **2,332.31** and a training RMSE of 1,840.51. For the median-focused metric, the Pinball Loss (α = 0.5) was **707.36** on the validation set and 614.07 on the training set, indicating good generalization power without significant overfitting. Finally, the best configuration of the Gradient Boosting model was used to train on the full training set (train + validation) and generate the final price predictions for the test set.