MACHINE LEARNING PROJECT (FIRST DELIVERABLE)-REPORT

Master's in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

CARS 4 YOU DELIVERABLE REPORT

Group 48

Project Members	Student Number
Khadija Ennaifer	20250439
Zeineb Hajji	20250535
Batoul Abdullah	20250536
Christian Chukwuebuka Ozougwu	20250449

1.0 Pipeline Structure

Our machine learning pipeline was built in a clear and logical order to ensure the data is clean, consistent, and ready for prediction. Each part of the process builds on the previous one, starting from cleaning the raw data and ending with the evaluation of the final model. The goal was to create a reliable and interpretable model that can estimate car prices with good accuracy.

2.0 Schematic Representation and Processes

Below is the schematic representation of the different steps we applied

Below, we have captured the respective activities and processes conducted in each step

Stage 1: Data Cleaning and Preprocessing

1. **Handling Missing Values:** For numerical columns like mileage, tax, mpg, and engineSize, missing values were replaced with the **mean** because it keeps the data balanced. Categorical columns like Brand, model, fuelType, and transmission, with missing values, were filled with the **most frequent value** to keep the categories consistent.
2. **Fixing Errors and Inconsistencies:** Many columns had typos or different spellings for the same category (like "anual" instead of "Manual" or "etrol" instead of "Petrol"). These were corrected using mapping dictionaries to make all values uniform. The same cleaning was applied to Brand, model, fuelType, and transmission.
3. **Correcting Invalid Values:** Negative numbers in mileage, tax, and engineSize were replaced with their **absolute values**. Values in paintQuality% above 100% were capped at **100**. The previousOwners column was rounded and converted to integers because it represents a count.
4. **Outlier Treatment:** Outliers in numerical columns (year, mileage, engineSize, tax, mpg, and paintQuality%) were capped using the **Z-score method** to avoid the influence of extreme values.
5. **Duplicate Removal:** Duplicate rows were removed to make sure each car appears only once in the dataset.
6. **Scaling and Encoding:** Numerical features were **scaled** using StandardScaler to bring them to a similar range. Categorical features were **encoded** in two different ways:
   a. **One-Hot Encoding** – turned categories into binary columns.
   b. **Target Encoding** – replaced each category with its average car price.
   Both methods were tested later to compare their impact on performance.

The above steps helped to make the data complete, clean, and consistent before moving to feature selection and modelling.

Stage 2: Feature Selection: The following techniques were tested and compared:

1. **Filter Method (SelectKBest):** Selected the top 10 numerical and encoded features most correlated with price (like year, mileage, and engineSize).

2. **Wrapper Method (RFE):** Used Recursive Feature Elimination to remove less important variables step by step until only the best ones remained.

3. **Embedded Method (Lasso Regularisation):** Applied Lasso regression, which automatically reduces the weight of weak predictors to zero, keeping only strong variables.

4. **Correlation and Chi-Square Tests: We used correlation to identify** numerical features strongly related to price (engineSize, year, mileage, mpg, and tax). Chi-square tests were used for categorical features and showed that Brand, transmission, and fuelType are important predictors.

5. **Tree-Based Feature Selection (Random Forest):** Random Forest automatically calculates the importance of each variable. The top 50% of features were kept based on their contribution to the model.

Stage 4: Model Training and Validation: After feature selection, two models were trained and compared:

1. **Ridge Regression:** used as a baseline model after applying feature selection. It gave decent performance (R^2 around 0.73).
2. **Random Forest Regressor:** the final model, which gave the best results. It performed better on both training and validation data and handled mixed data types well.

The data was then divided into **75% for training** and **25% for validation** to evaluate how well the model could generalise to unseen data. Among all methods, the **Random Forest** approach gave the best results, achieving **R^2 = 0.8966**, **MSE = 10,023,069**, and **RMSE = 3,165.9**.
This method worked best because it captured both linear and non-linear relationships between features and the target variable.

Stage 5: Evaluation and Interpretation: The model's performance was evaluated using:

1. **R^2 (Coefficient of Determination):** to show how much of the variation in car prices is explained by the model.
2. **MSE (Mean Squared Error)** and **RMSE (Root Mean Squared Error):** to measure how far predictions are from real prices.

Visual comparisons between actual and predicted prices were also made using scatter plots to check prediction accuracy.

3.0 Summary

The project followed a structured process from cleaning raw data to building the final model. Each step, cleaning, encoding, selecting features, and training models, was essential to reach accurate and reliable results. After testing several methods, the **Random Forest model** proved to be the best choice, offering strong prediction accuracy and showing how machine learning can be used to estimate fair car prices effectively.