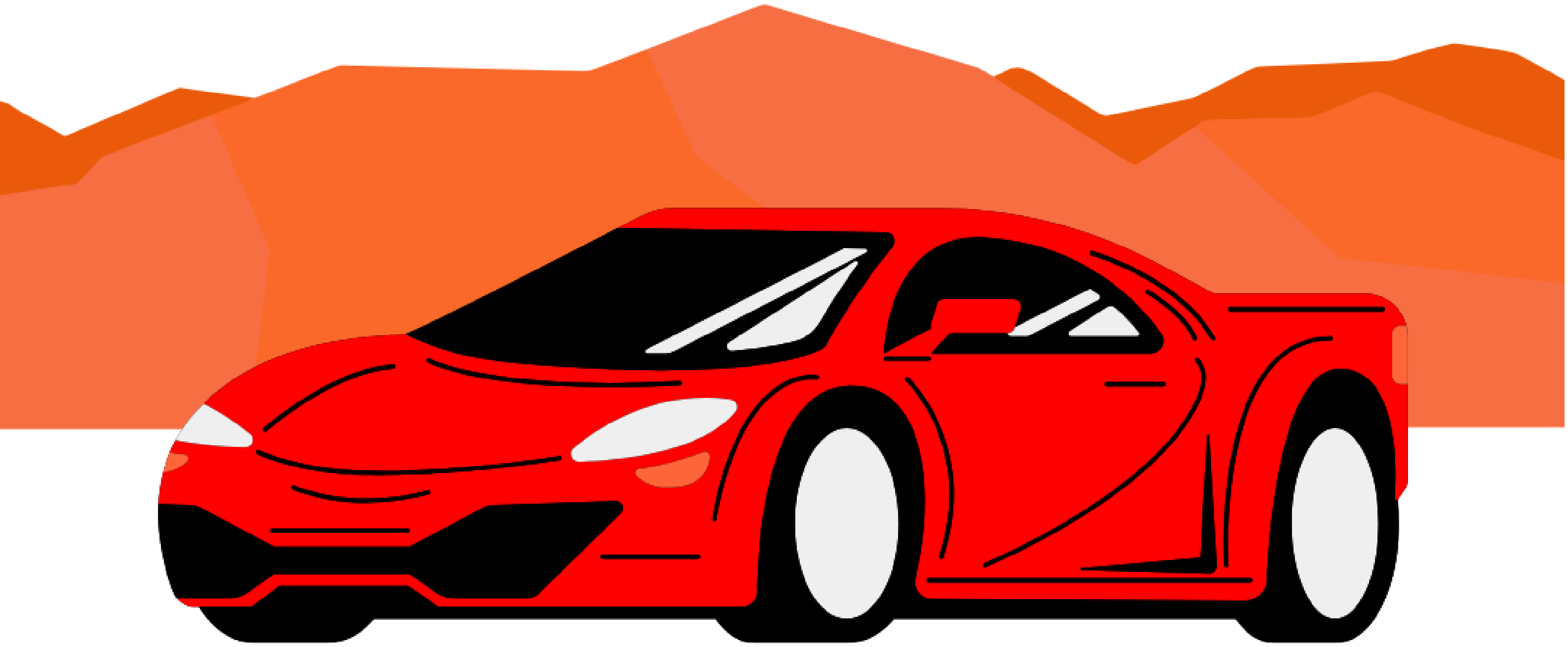
Master in **Data Science and Advanced Analytics**
2025/2026
**GROUP 27**

NOVA IMS
INFORMATION MANAGEMENT SCHOOL

# CARS 4 YOU: WE BUY YOUR CAR!

# MACHINE LEARNING

**Instructors: Roberto Henriques, Leon Debatin, Ricardo Santos**
HANDOUT PDF



**GONÇALO RILHAS, 20250490**
**GUILHERME SANTOS, 20250510**
**LUANA PINTO, 20250481**
**MARTA PEDRO, 20250500**

# 1. PROJECT OBJECTIVES

This Machine Learning (ML) project[1] was developed for “Cars 4 You”, an online car resale platform dependent on manual car evaluation workflow.

To address the company's challenge, the project aims to design and implement a supervised machine learning regression model capable of predicting car prices (being ‘*price*’ our target variable) directly from user-provided attributes, effectively automating the valuation process and reducing the reliance on manual assessments.

The solution follows a data-driven pipeline extensive data preprocessing, feature engineering, model training and evaluation. Multiple regression algorithms, including Linear Regression, SVR and Random Forest Regressor, were benchmarked to identify the most accurate and robust approach.

# 2. PIPELINE OVERVIEW

The project followed a structured ML pipeline designed to resolve data challenges. The process culminated in the selection of a robust feature set for the final model training.

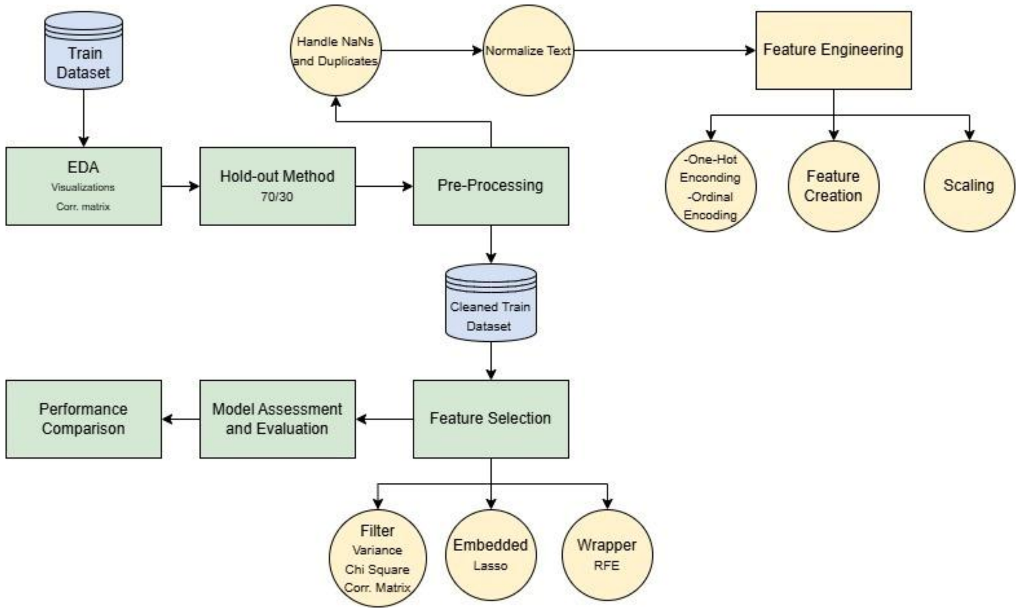


*Fig. 1: Flowchart of the regression problem pipeline.*

# 3. PREPROCESSING AND FEATURE SELECTION DETAILS

These vital stages were processed to ensure both statistical consistency and predictive robustness. Although the project initially aimed to satisfy the assumptions of Ordinary Least Squares (OLS) regression, the resulting transformations significantly improved the dataset’s structure and interpretability, supporting the performance of the final current chosen model.

[1] The project can be checked at our group GitHub repository: https://github.com/martap18/Machine-Learning-Project

### 3.1 PREPROCESSING AND FEATURE ENGINEERING

The preprocessing pipeline began with a set of transformations applied to numerical variables. Continuous variables such as '*mileage*', '*tax*', '*mpg*' and '*engineSize*' (and later in the notebook, our target variable) underwent logarithmic transformations to reduce strong right-skewness and to linearize their relationships with the target variable. With this step, it was ensured that these predictors contributed more evenly across the model.

Temporal and usage-based predictors were refined to reduce redundancy and improve model interpretability. The original '*year*' and '*mileage*' variables were replaced with '*car_age*' and '*log_avg_mileage*' (the logarithm of mileage per year of age), which better describe depreciation and usage intensity while minimizing multicollinearity.

Categorical variables with high cardinality, such as '*Brand*' and '*model*', were encoded using Ordinal Encoding to maintain efficiency and prevent dimensional inflation. This approach assigns each category a unique integer value, which is particularly suitable for non-linear models capable of capturing complex relationships. Variables with fewer discrete levels, such as '*previousOwners*' and '*paintQuality%*' and others, were binned into interpretable tiers (as example, "*owners_cat_SingleOwner*" and "*paint_condition_Excellent*" respectively) to reflect market perception and subsequently transformed using One-Hot Encoding to enhance model interpretability.

### 3.2 FEATURE SELECTION STRATEGY

The feature selection section was performed through a diverse methodology combining statistical testing, domain knowledge and model-driven regularization. Initially, correlation analysis and the Chi-squared test of independence were used to assess variable relevance, guiding the removal of predictors with slight predictive power, such as '*owners_cat_ManyOwners*' and others.

Next, RFE, a wrapper method, and LassoCV with L1 regularization, an embedded approach, were applied to identify and eliminate redundancy. LassoCV automatically shrank the coefficients of features that provided overlapping information to zero, confirming that engineered attributes such as '*car_age*' and '*log_avg_mileage*' effectively captured the information from original predictors. Minor divergences between RFE and Lasso, such as for '*owners_cat_SingleOwner*' and '*paint_condition_Excellent*', were resolved based on domain reasoning, retaining features with practical interpretability despite low coefficients.

The final selection resulted in a set of twelve robust features, balancing statistical relevance, model efficiency and reduced overfitting risk.

## 4. CURRENT BEST-PERFORMING MODEL

The Random Forest Regressor with 12 features and our target variable logarithmized outperformed alternative models ($R^2$ of 0.9302 and an RMSE of £2,640,30). Its robustness to heteroscedasticity and skewed targets, combined with the ability to capture complex non-linear relationships and interactions among categorical features, justified its selection. This model thus provides a reliable and efficient first foundation for Cars 4 You's automated car valuation system.