Title: Pre-Owned Car Price Prediction

Author: Mayur Ramesh Bhurle (Roll no. 23M1027), ET IIT Bombay.

1. Introduction

The pre-owned car market is a significant segment of the automotive industry, providing consumers with affordable alternatives to new vehicles. Accurate pricing is crucial for both buyers and sellers, ensuring fairness and competitiveness. However, predicting the price of a pre-owned car involves many variables, including the car's age, mileage, brand, and condition. This study aims to optimize regression models for predicting the price of pre-owned cars by carefully managing missing data and selecting relevant features.

2. Problem Statement

The primary objective of this project is to develop a robust predictive model for estimating the price of pre-owned cars using regression techniques. Specifically, the study seeks to:

- Optimize regression models by handling missing data effectively.

- Evaluate the performance of Linear Regression and Random Forest models.

- Compare the accuracy of models trained on datasets with missing data omitted versus datasets with missing data imputed.

3. Methodology

The methodology section outlines the steps taken to prepare the data, select features, and build the regression models. The approach was divided into several stages, including data cleaning, feature engineering, and model training.

3.1 Data Cleaning and Preparation

The dataset used in this study was sourced from a public repository and consisted of various features such as the car's registration year, price, power, and more. Initially, the dataset contained some anomalies, including unrealistic values for the year of registration and power, as well as missing data.

- Data Cleaning: The first step involved removing outliers and unrealistic values. For instance, cars registered before 1950 or after 2018 were removed from the dataset. Similarly, cars with power values below 10 PS or above 500 PS were excluded.

- Handling Missing Data: Missing data was managed in two ways. In the first approach, rows with any missing values were omitted from the dataset. In the second approach, missing values were imputed using median values for numerical features and the most frequent values for categorical features.

- Feature Engineering: New features were created to enhance the model's predictive power. For example, the "Age" of the car was calculated by subtracting the registration year from 2018 and adding a fraction of the month of registration.

3.2 Feature Selection

Several features were considered for the regression models, including:

- Age: Derived from the year and month of registration.
- PowerPS: The power of the car's engine in PS.
- VehicleType: The type of the vehicle, such as sedan, SUV, etc.
- Gearbox: Type of gearbox, either automatic or manual.
- FuelType: The type of fuel used by the car, such as petrol, diesel, etc.
- Brand: The manufacturer of the car.
- NotRepairedDamage: Indicates whether the car has unrepaired damage.

Features that showed little to no variation or had no significant impact on the price were dropped from the model, such as the seller and offer type.

3.3 Model Building

Two regression models were built on two different datasets: one where missing data was omitted and one where missing data was imputed.

- Linear Regression: A linear approach to modeling the relationship between the dependent variable (price) and independent variables (features).

- Random Forest Regression: An ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction.

The models were trained on the prepared datasets and evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

4. Results

The results section presents the performance of the Linear Regression and Random Forest models on both the omitted and imputed datasets.

4.1 Performance on Omitted Data

- Linear Regression:
  - RMSE: 0.545
  - R-squared (Train): 0.7
  - R-squared (Test): 0.7

- Random Forest:
  - RMSE: 0.436
  - R-squared (Train): 0.85
  - R-squared (Test): 0.83

The Random Forest model outperformed the Linear Regression model in terms of both RMSE and R-squared values, indicating better accuracy and model fit.

4.2 Performance on Imputed Data

- Linear Regression:
  - RMSE: 0.63
  - R-squared (Train): 0.708

- R-squared (Test): 0.706

- Random Forest:
  - RMSE: 0.511
  - R-squared (Train): 0.82
  - R-squared (Test): 0.806

Again, the Random Forest model demonstrated superior performance compared to the Linear Regression model. Moreover, the results were not better than those obtained from the dataset with omitted data, suggesting that imputing missing values is not improving model accuracy.

5. Discussion

The analysis highlights several key findings. First, the Random Forest model consistently outperformed the Linear Regression model in predicting the price of pre-owned cars. This is likely due to the Random Forest's ability to capture complex interactions between variables that a linear model might miss.

Second, handling missing data through omission rather than imputation proved to be beneficial. Feature selection also played a crucial role in the model's success. By focusing on significant features such as the car's age, power, vehicle type, and brand, the models were able to make more accurate predictions. Features like the gearbox type and fuel type also contributed to the model's accuracy, although to a lesser extent.

6. Conclusion

This study successfully developed and optimized regression models for predicting the price of pre-owned cars. The Random Forest model emerged as the superior model, outperforming Linear Regression in both cases—whether missing data was omitted or imputed.

The study underscores the importance of data cleaning, feature engineering, and appropriate handling of missing data in building accurate predictive models. For future work, the model could be further improved by incorporating additional features, such as the car's condition, previous ownership history, and market trends.

The results have practical implications for stakeholders in the pre-owned car market, including dealers, buyers, and financial institutions, enabling them to make more informed decisions based on accurate price predictions.