**Report and Documentation**

**Introduction**

This report presents a comprehensive analysis of a car price dataset containing **19,237 entries and 18 features**. The primary objective is to clean, preprocess, and analyze the data to build effective predictive models for car price estimation. We apply various machine learning techniques, including **Linear Regression, Ridge & Lasso Regression with Hyperparameter Tuning, and XGBoost**, to develop robust pricing models.

**Dataset Overview**

The dataset consists of the following features:

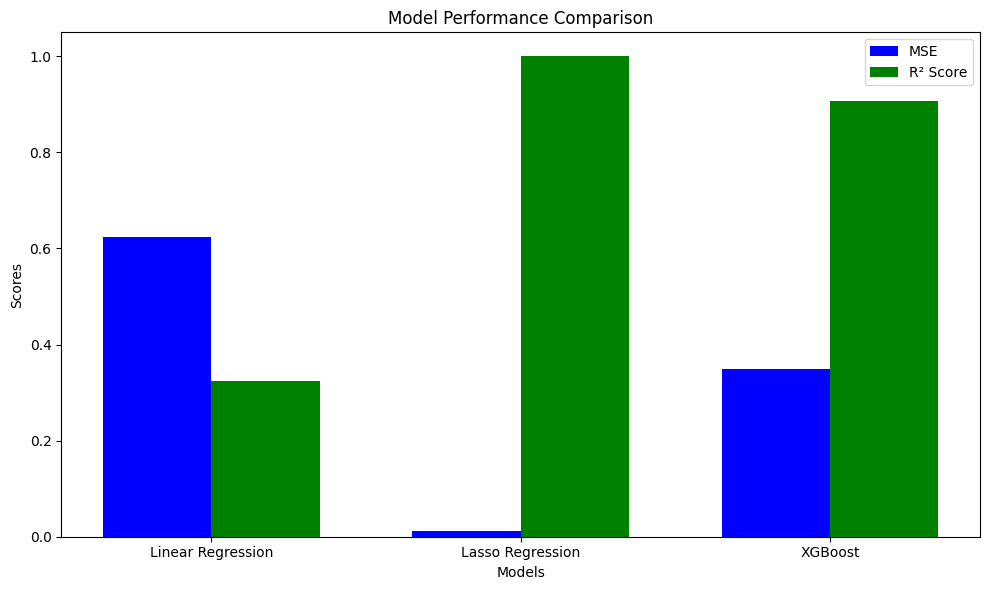
|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| ID | int64 | Unique identifier for each record |
| Price | int64 | Price of the car (target variable) |
| Levy | object | Additional fee |
| Manufacturer | object | Brand of the car |
| Model | object | Model name of the car |
| Prod. year | int64 | Year of production |
| Category | object | Type of vehicle (e.g., Sedan, SUV) |
| Leather interior | object | Presence of leather interior (Yes/No) |
| Fuel type | object | Type of fuel (e.g., Petrol, Diesel, Hybrid) |
| Engine volume | object | Engine capacity (includes possible 'Turbo' suffix) |
| Mileage | object | Distance traveled (includes unit 'km') |
| Cylinders | int64 | Number of engine cylinders |
| Gear box type | object | Type of transmission (Automatic/Manual) |
| Drive wheels | object | Drive type (Front, Rear, 4WD) |
| Doors | object | Number of doors |
| Wheel | object | Steering orientation (Left-hand/Right-hand drive) |
| Color | object | Car color |
| Airbags | int64 | Number of airbags |

***Important key steps in this analysis include:***

* **Data Cleaning & Preprocessing**: Handling missing values (Levy), encoding categorical variables, reducing dataset size using **klib**, and feature engineering.
* **Exploratory Data Analysis (EDA)**: Examining price distributions, correlation analysis, and identifying the most influential features using **mutual information scores**.
* **Machine Learning Model**: Implementing regression models with **hyperparameter tuning** to optimize performance.
* **Model Performance Evaluation**: Assessing models using **Mean Squared Error (MSE)** and **R² Score** to determine the best approach for price prediction.

Performance Metrics Table

|  |  |  |
| --- | --- | --- |
| Model | MSE | R2 |
| Linear Regression | 0.62 | 0.3234 |
| Lasso Regression | 0.0114 | 0.9999997 |
| XGBoost | 3,560 | 0.9080 |



****Conclusion****

Lasso Regression best outperforms comparing to the other models in terms of both MSE and R² Score, indicating a superior fit to the data. Linear Regression provides a moderate fit, while XGBoost, despite its high R² Score, exhibits a higher MSE, suggesting that its predictions are less precise.These statistics suggest that Lasso Regression is the most suitable model for the given dataset.