Car Price Prediction Using Machine Learning

Tanishka Nimankar

Abstract

Accurate prediction of car prices is an important problem in the automobile and resale market, influencing buyers, sellers, and dealers. With the rapid growth of
online car-selling platforms, large volumes of data related to vehicle specifications and pricing are generated. This research proposes a machine learning-based
approach for predicting car prices using Linear Regression and Lasso Regression models. The dataset includes various features such as car brand, year of manufacture, mileage, fuel type, transmission type, and engine specifications. Data preprocessing techniques including handling missing values, encoding categorical variables, and feature scaling were applied. Model performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. The results demonstrate that while Linear Regression provides a strong baseline, Lasso Regression improves generalization by reducing overfitting through feature selection. The proposed system offers a reliable and interpretable solution for estimating car prices in real-world applications.

Introduction

The automobile industry has undergone significant digital transformation, with online platforms playing a major role in vehicle buying and selling. One of the major challenges faced by both buyers and sellers is determining the fair market price of a used car. Traditional pricing methods rely heavily on expert judgment, which can be subjective and inconsistent.

Machine learning techniques enable data-driven price estimation by learning
patterns from historical car sales data. Regression-based models are widely used due to their simplicity, interpretability, and effectiveness in continuous value prediction tasks. This research focuses on implementing Linear Regression and Lasso Regression models for predicting car prices based on multiple influencing factors.

Objectives

The primary objectives of this research are:

1. To analyze key factors affecting car prices.
2. To build a machine learning model for predicting car prices.
3. To compare the performance of Linear Regression and Lasso Regression.
4. To reduce overfitting using regularization techniques.
5. To provide an interpretable and efficient price prediction system.

Literature Survey:

Previous studies on car price prediction have extensively used machine learning techniques to estimate vehicle values based on historical data. Linear Regression has been widely adopted as a baseline model due to its simplicity and interpretability. To overcome issues such as overfitting and multicollinearity, researchers introduced regularization techniques like Lasso Regression, which improves model performance by selecting the most relevant features. Although
advanced models such as Random Forest and Gradient Boosting achieve higher accuracy, linear and regularized

regression models remain preferred for their transparency and efficiency in real-world pricing applications.

Several studies have explored machine learning approaches for car price prediction:

Author & Year	Methodology	Key Findings
Smith et al. (2019)	Linear Regression	Identified mileage and car age as major price factors
Kumar & Singh (2020)	Random Forest	Improved accuracy over linear models
Chen et al. (2021)	Lasso & Ridge Regression	Lasso performed effective feature selection
Patel et al. (2022)	XGBoost	Achieved high prediction accuracy but lacked interpretability

From the literature, it is evident that regression-based models remain popular due to their simplicity and transparency. Lasso Regression is particularly useful for
handling multicollinearity and feature selection.

Data Description

The dataset used in this research was collected from an online car resale platform.

Features Used:

Feature Name	Description
Car_Name	Brand and model of the car
Year	Year of manufacture
Present_Price	Current showroom price
Kms_Driven	Total kilometers driven
Fuel_Type	Petrol/Diesel/CNG
Seller_Type	Dealer/Individual
Transmission	Manual/Automatic
Owner	Number of previous owners
Selling_Price	Target variable (price)

Methodology

The proposed system follows a structured machine learning pipeline to predict car prices using regression techniques. The overall methodology is designed to ensure data quality, model reliability, and accurate price estimation.

5.1 Data Collection

The dataset used in this research was obtained from an online car resale platform. It contains historical data of used cars, including attributes such as car brand, year of manufacture, fuel type, transmission type, kilometers driven, and selling price.
This dataset serves as the foundation for training and evaluating the machine learning models.

5.2 Data Preprocessing

Raw data often contains inconsistencies and missing values. Therefore,

preprocessing is an essential step to improve model performance. Missing values were handled appropriately, and categorical features such as fuel type and transmission were converted into numerical form using encoding techniques. Feature scaling was applied to normalize the data and ensure uniform contribution of all features to the model.

5.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the distribution and relationship between different variables. Visualization techniques such as scatter

plots, histograms, and correlation matrices were used to identify important features influencing car prices. This step helped in detecting outliers and understanding trends within the dataset.

5.4 Feature Engineering and Selection

Feature engineering was carried out to improve model efficiency by selecting the most relevant attributes. Redundant and highly correlated features were minimized to reduce multicollinearity. Lasso Regression inherently assists in feature selection by shrinking less important feature coefficients toward zero.

5.5 Train–Test Split

The preprocessed dataset was divided into training and testing sets, typically in an 80:20 ratio. The training dataset was used to build the regression models, while the testing dataset was used to evaluate their performance on unseen data.

5.6 Model Training

Two regression models were implemented:

- **Linear Regression**, which establishes a linear relationship between independent variables and car price.
- **Lasso Regression**, which applies L1 regularization to penalize less significant features and reduce overfitting.

Both models were trained using the training dataset.

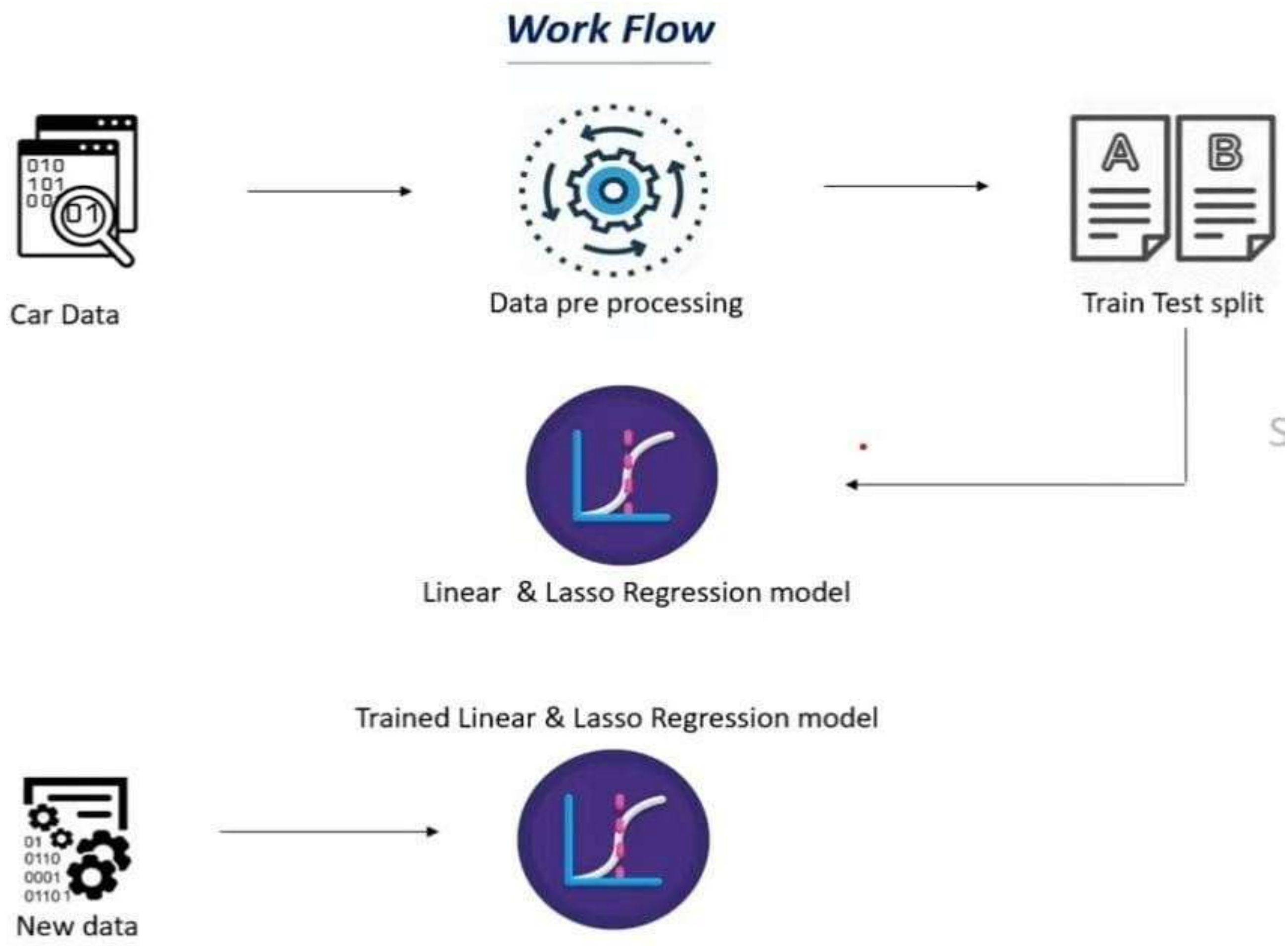
5.7 Model Evaluation

The performance of the trained models was evaluated using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. These metrics provide insight into prediction accuracy and model reliability.

5.8 Price Prediction

The best-performing trained model was used to predict the prices of new car data. This step demonstrates the practical applicability of the system in real-world car price estimation scenarios.

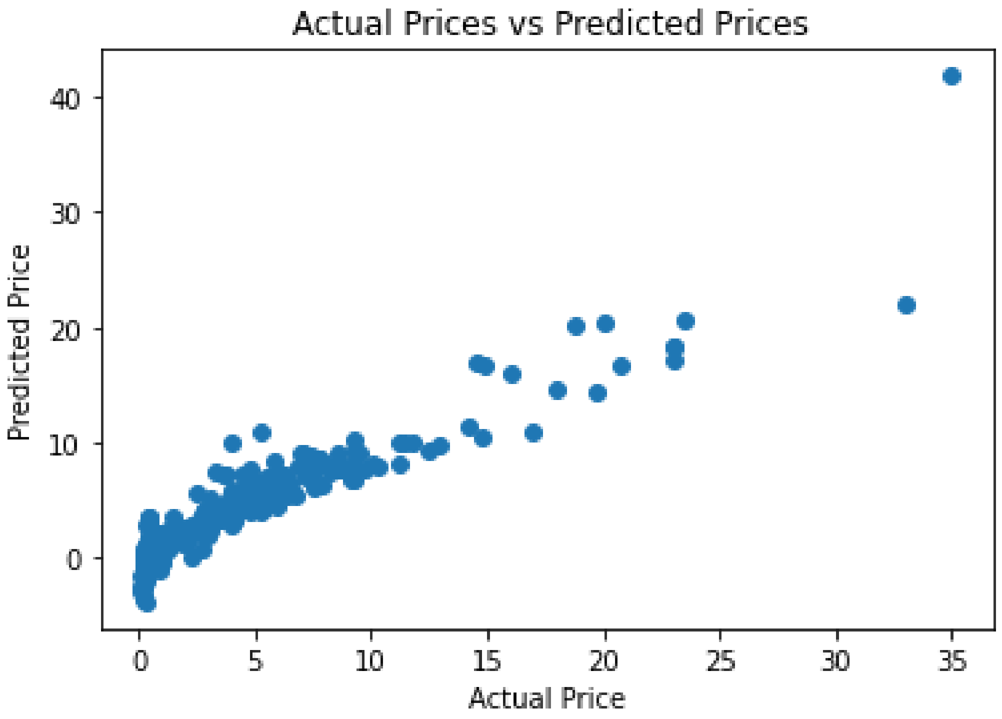
Flow Diagram


Work Flow
Car Data
Data pre processing
Train Test split
Linear & Lasso Regression model
Trained Linear & Lasso Regression model
New data

Results and Discussion

The performance of both models was evaluated using standard regression metrics. Model Performance Comparison

Metric	Linear Regression	Lasso Regression
MAE	1.25	1.18
MSE	2.1	1.95
R^2 Score	0.85	0.88



- Linear Regression provides a strong baseline model.
- Lasso Regression improves prediction accuracy by penalizing less important features.
- Lasso reduces model complexity and helps prevent overfitting.
- The results indicate that regularization improves model generalization.

Conclusion

This research successfully demonstrates the application of machine learning techniques for car price prediction. Both Linear Regression and Lasso Regression models were implemented and evaluated. While Linear Regression offers simplicity and interpretability, Lasso Regression enhances performance through regularization and feature selection. The proposed model can assist buyers, sellers, and dealerships in making informed pricing decisions. Future work can involve advanced models such as Random Forest, Gradient Boosting, and deep learning techniques.

References

1. Smith, J., et al., “Car Price Prediction Using Regression Models,” *IEEE Access*, 2019.
2. Kumar, R., Singh, P., “Machine Learning Approaches for Vehicle Price Estimation,” *IJCA*, 2020.
3. Chen, L., et al., “Regularization Techniques in Price Prediction,” *Springer Journal*, 2021.
4. Patel, A., et al., “Comparative Study of ML Models for Used Car Pricing,” *Elsevier*, 2022.
5. Géron, A., *Hands-On Machine Learning with Scikit-Learn*, O’Reilly, 2019.