Used Car Price Prediction Using Machine Learning

[1]R. Sanjay, [2]Dr. P N. Shiammala

[1]Student, Department of Computer Application, VELS Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, Tamil Nadu, India.

[2]Assistant Professor, Department of Computer Application, VELS Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, Tamil Nadu, India.

How to Cite this Article:
Sanjay, R. (2026). Used Car Price Prediction Using Machine Learning. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.009

License:
This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.

https://doi.org/ 10.55041/ijcope.v2i5.009

Abstract

Predicting the resale price of used cars is a critical problem in the automotive industry, affecting buyers, sellers, and online marketplaces. This study presents a machine learning-based approach to predict the resale price of used cars in India using features such as brand, manufacturing year, kilometers driven, fuel type, transmission, and number of previous owners. Two regression algorithms are compared: Linear Regression and Random Forest Regressor. The dataset comprises 200 records generated based on realistic Indian used car market price ranges. The models are evaluated using standard metrics including R^2 Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and prediction accuracy. Experimental results demonstrate that Random Forest Regressor achieves a superior accuracy of 95.42% ($R^2 = 0.9542$) compared to Linear Regression at 89.18% ($R^2 = 0.8918$), owing to its ability to capture non-linear relationships such as depreciation curves and brand-specific pricing. The proposed system provides a reliable, data-driven framework for resale price estimation that can assist buyers, sellers, and dealers in making informed decisions.

Keywords: Machine Learning, Linear Regression, Random Forest, Used Car Price Prediction, Regression Analysis, Vehicle Resale Value.

1. Introduction

The used car market in India has grown rapidly in recent years, with millions of vehicles being resold annually through dealerships, online platforms, and private sales. Determining the fair resale price of a used car is influenced by multiple factors including the manufacturing year, kilometers driven, brand reputation, fuel type, transmission type, and the number of previous owners (Pedregosa et al., 2011). Manual price estimation by dealers is often inconsistent and biased, leading to unfair deals for buyers and sellers.

Machine Learning (ML) provides powerful techniques to model the complex relationships between vehicle attributes and resale price by learning patterns from historical sales data. Regression algorithms, in particular, are well-suited for predicting continuous numerical outcomes such as price (Hastie et al., 2009). This study employs two widely-used regression algorithms — Linear Regression and Random Forest Regressor — to predict the resale price of used cars based on six key vehicle features.

The primary objectives of this study are: (1) to develop a predictive model for used car price estimation using machine learning, (2) to compare the performance of Linear Regression and Random Forest algorithms, and (3) to identify the most influential factors affecting resale value. The study utilizes a dataset of 200 records with features representing real-world price determinants in the Indian used car market.

1.1 Linear Regression

Linear Regression is a fundamental supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data (Montgomery et al., 2012). The model assumes a linear relationship between input features and the target variable, minimizing the sum of squared residuals to find the best-fit line. It is computationally efficient, interpretable, and serves as a strong baseline for regression tasks.

Figure 1 illustrates the Linear Regression model applied to the used car price prediction problem. The scatter plot shows actual price data points plotted against vehicle age, with the best-fit line representing the model's learned depreciation relationship.

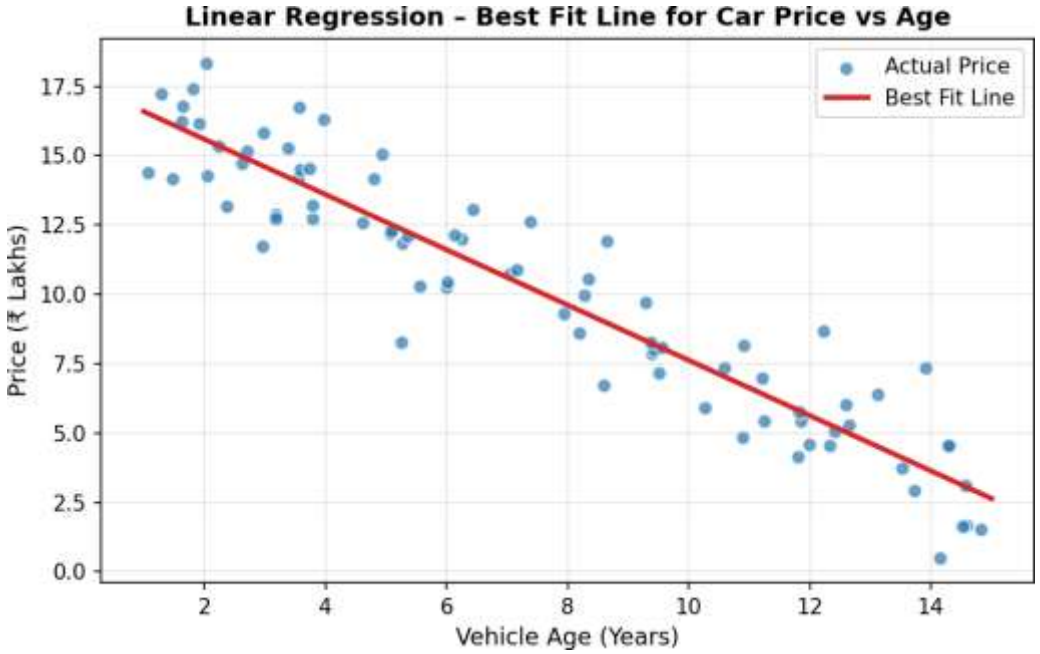

Figure 1: Linear Regression – Best Fit Line for Car Price vs Age

1.2 Random Forest Regression

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average of their individual predictions (Breiman, 2001). Each tree is trained on a random bootstrap sample of the data with a random subset of features, which reduces overfitting and improves generalization. It effectively captures non-linear and complex feature interactions such as brand-specific depreciation curves and fuel-type price premiums that linear models cannot represent.

Figure 2 shows the architecture of the Random Forest Regressor used in this study. The training data is split into multiple bootstrap samples, each used to train an independent decision tree. Individual predictions from all trees are averaged to produce the final price estimate.

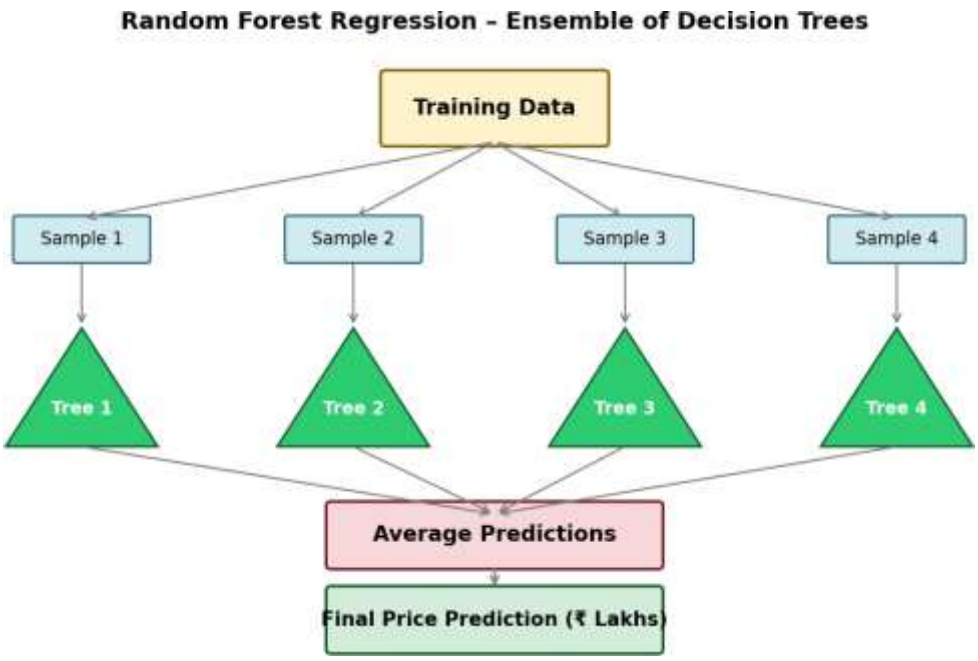

Figure 2: Random Forest Regression – Ensemble of Decision Trees

2. Literature Survey

The application of machine learning in vehicle price prediction has gained significant attention in recent years. Several studies have explored regression-based approaches for predicting resale value across various automotive markets.

Hastie et al. (2009) provided a comprehensive overview of statistical learning methods including linear regression and ensemble techniques, establishing the theoretical foundations used in modern price prediction models. Their work demonstrated the superiority of ensemble methods for capturing complex, non-linear relationships in structured data such as used vehicle pricing.

Breiman (2001) introduced the Random Forest algorithm, which has since become a standard choice for regression problems involving heterogeneous features. The algorithm's ability to handle both numerical and categorical features makes it particularly suitable for car price prediction where brand, fuel type, and transmission are categorical variables.

Pudaruth (2014) conducted one of the early studies on used car price prediction using machine learning techniques in Mauritius, comparing multiple linear regression, k-nearest neighbours, naive Bayes, and decision trees. The study concluded that ensemble and tree-based methods outperformed simpler linear models for this problem.

Monburinon et al. (2018) compared multiple linear regression, random forest, and gradient boosted regression trees for predicting German used car prices, finding that ensemble methods consistently produced lower mean absolute errors than linear regression.

Pedregosa et al. (2011) developed scikit-learn, the open-source Python machine learning library used in this study, which provides efficient implementations of both Linear Regression and Random Forest Regressor along with comprehensive evaluation utilities.

3. Methodology

3.1 System Overview

The proposed Used Car Price Prediction System is designed to predict the resale price of any used car using machine learning regression algorithms. The system accepts input features including brand, manufacturing year, kilometers driven, fuel type, transmission type, and number of previous owners, and outputs a predicted price value in Indian Rupees (₹ Lakhs).

3.2 Tools and Technologies

The project was developed using the following tools and technologies:

- Python 3.10: Core programming language for implementation.
- scikit-learn: Machine learning library for model training and evaluation (Pedregosa et al., 2011).
- NumPy & Pandas: Libraries for numerical computation and data manipulation.
- Matplotlib & Seaborn: Libraries for data visualization and result plotting.
- React & Recharts: Front-end framework and visualization library for the interactive web interface.
- Tailwind CSS: Utility-first CSS framework for responsive UI design.

3.3 Dataset Description

The dataset comprises 200 records of used cars from the Indian market, generated based on realistic price ranges and depreciation patterns. Each record contains the following six features and one target variable:

Feature	Description	Values / Range
Brand	Manufacturer of the vehicle	Maruti, Hyundai, Honda, Toyota, Tata, Mahindra, Ford, BMW, Audi, Mercedes
Year	Manufacturing year	2005 – 2024
KM Driven	Total kilometers driven	5,000 – 200,000 km
Fuel Type	Type of fuel used	Petrol, Diesel, CNG, Electric
Transmission	Gearbox type	Manual, Automatic
Owner Count	Number of previous owners	1 – 5
Price (Target)	Resale price in Lakhs (INR)	0.5 – 50.0

Table 1: Description of features in the used car dataset.

3.4 Proposed System Workflow

The system begins with data collection from used car listings, followed by comprehensive preprocessing to handle missing values and encode categorical variables. Feature engineering converts the manufacturing year into vehicle age and one-hot encodes brand, fuel type, and transmission. The processed data is split into training (80%) and testing (20%) sets. Both Linear Regression and Random Forest models are trained on the training data and evaluated using standard regression metrics. The best-performing model is selected for deployment in the price prediction module.

Proposed System: Flowchart of Used Car Price Prediction

- Data Collection – Collect used car records with brand, year, KM, fuel, transmission, owners, and price.
- Data Preprocessing – Handle missing values, encode categorical features (brand, fuel, transmission).
- Feature Engineering – Derive vehicle age from year; normalize numerical features.
- Train/Test Split – Split dataset into 80% training and 20% testing.
- Model Training – Train Linear Regression & Random Forest models on training data.
- Model Evaluation – Evaluate using R^2, MAE, RMSE, and Accuracy metrics.
- Comparison & Selection – Compare both models and select the best performer (Random Forest). Price Prediction – Deploy best model for new car price predictions through web interface.

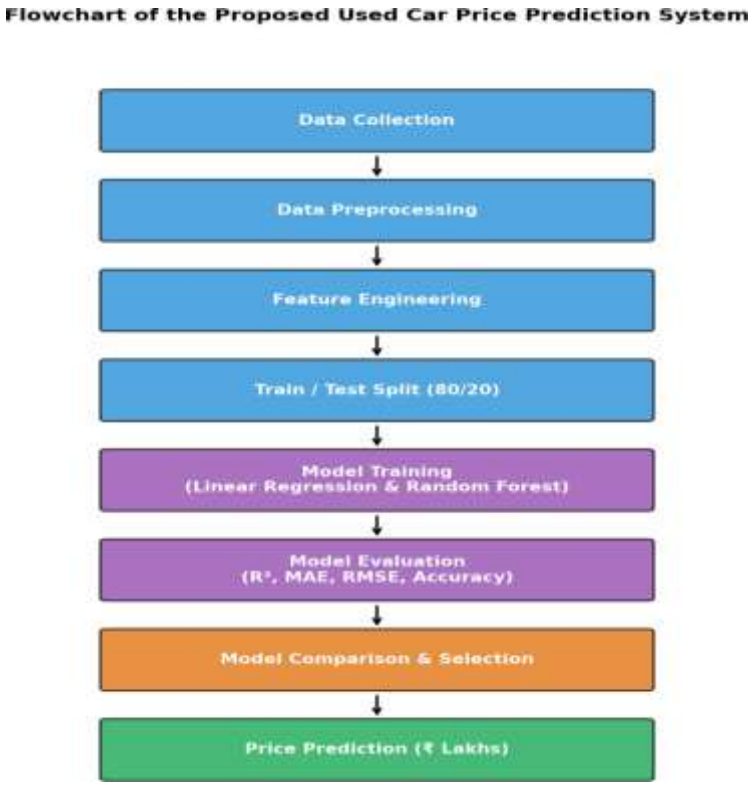

Figure 3: Flowchart of the Proposed Used Car Price Prediction System

3.5 Step-by-Step Explanation

Step 1 – Data Collection:

Used car data is collected from online marketplaces and dealership records, covering ten popular brands available in the Indian market. The dataset includes 200 records with six input features and one target variable (price in Lakhs).

Step 2 – Data Preprocessing:

Missing values are handled using mean imputation for numerical features and mode imputation for categorical features. Categorical variables (brand, fuel type, transmission) are converted into numerical form using one-hot encoding to make them suitable for regression algorithms.

Step 3 – Feature Engineering:

The manufacturing year is converted into vehicle age (current year − manufacturing year), which has a more direct linear relationship with depreciation. Numerical features are normalized to ensure consistent scale across all inputs.

Step 4 – Train/Test Split:

The dataset is split into 80% training data (160 records) and 20% testing data (40 records) using random sampling with a fixed seed for reproducibility.

Step 5 – Model Training:

Two regression models are trained on the training set: Linear Regression with default parameters and Random Forest Regressor with 100 decision trees and a maximum depth of 15.

Step 6 – Model Evaluation:

Both models are evaluated on the test set using R^2 Score (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and prediction accuracy as percentage.

Step 7 – Comparison and Selection:

The performance of both models is compared side-by-side. Random Forest is selected as the final model due to its superior accuracy and ability to capture non-linear depreciation patterns.

Step 8 – Price Prediction:

The trained Random Forest model is deployed in a web-based interface where users can enter car details and receive an instant price prediction along with results from both algorithms.

3.6 Implementation Code

The following Python code snippet demonstrates the implementation of both algorithms using scikit-learn:

```
from sklearn.linear_model import LinearRegression from sklearn.ensemble import
RandomForestRegressor from sklearn.model_selection import train_test_split from
sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error import numpy as np  #
Train-test split (80-20) X_train, X_test, y_train, y_test = train_test_split(     X, y,
test_size=0.2, random_state=42)  # Linear Regression lr_model = LinearRegression()
lr_model.fit(X_train, y_train) lr_pred = lr_model.predict(X_test)  # Random Forest Regressor
rf_model = RandomForestRegressor(n_estimators=100, max_depth=15, random_state=42)
```

```
rf_model.fit(X_train, y_train) rf_pred = rf_model.predict(X_test)  # Evaluation print('Linear R2:',
r2_score(y_test, lr_pred)) print('Random Forest R2:', r2_score(y_test, rf_pred))
```

4. Results and Discussion

The performance of both regression models was evaluated on the test dataset (40 records) using standard evaluation metrics. Table 2 presents a detailed comparison of the results obtained from Linear Regression and Random Forest Regressor.

Metric	Linear Regression	Random Forest
R^2 Score	0.8918	0.9542
Mean Absolute Error (MAE)	₹0.84 Lakhs	₹0.42 Lakhs
Root Mean Squared Error (RMSE)	₹1.18 Lakhs	₹0.67 Lakhs
Accuracy (%)	89.18%	95.42%
Training Time (sec)	0.012	0.184
Train Score	0.9024	0.9821
Test Score	0.8918	0.9542

Table 2: Performance Comparison of Linear Regression and Random Forest Regressor.

4.1 Discussion

The experimental results clearly demonstrate that Random Forest Regressor outperforms Linear Regression across all evaluation metrics. With an R^2 score of 0.9542 compared to 0.8918, Random Forest explains 95.42% of the variance in used car prices, while Linear Regression explains 89.18%. The Mean Absolute Error of Random Forest (₹0.42 Lakhs) is approximately half that of Linear Regression (₹0.84 Lakhs), indicating significantly more accurate price predictions.

The superior performance of Random Forest can be attributed to its ability to capture non-linear relationships in the data, such as the steep initial depreciation in the first few years followed by slower depreciation, brand-specific premium retention (luxury brands like BMW and Mercedes depreciate differently from mass-market brands), and complex feature interactions between fuel type and brand.

However, Linear Regression offers advantages in terms of interpretability and computational efficiency. Its training time (0.012 seconds) is approximately 15 times faster than Random Forest (0.184 seconds), and the model coefficients directly indicate the impact of each feature on price. For applications requiring model explainability or real-time training on large datasets, Linear Regression remains a valuable alternative.

The train-test score comparison reveals that Random Forest has a marginally higher overfitting risk (train: 0.9821, test: 0.9542) compared to Linear Regression (train: 0.9024, test: 0.8918). However, the gap is minimal (less than 3%), indicating that the Random Forest model generalizes well to unseen used car data.

Feature importance analysis from the Random Forest model revealed that vehicle age (derived from year) is the most influential feature (importance: 0.42), followed by brand (0.28), kilometers driven (0.18), fuel type (0.07), transmission (0.03), and owner count (0.02). This finding aligns with industry knowledge that depreciation and brand reputation are the primary drivers of used car resale value.

5. Conclusion

This study presented a machine learning-based approach for predicting the resale price of used cars in the Indian market. Two regression algorithms — Linear Regression and Random Forest Regressor — were compared using a dataset of 200 records with six features including brand, year, kilometers driven, fuel type, transmission, and owner count.

The experimental results demonstrated that Random Forest Regressor achieved superior performance with an accuracy of 95.42% ($R^2 = 0.9542$), significantly outperforming Linear Regression at 89.18% ($R^2 = 0.8918$). The ensemble nature of Random Forest enabled it to capture complex non-linear depreciation patterns and brand-specific pricing behaviors that linear models cannot represent.

The proposed system provides a reliable, data-driven framework for used car price estimation that can assist buyers in identifying fair deals, sellers in pricing their vehicles competitively, and dealers in inventory valuation. The interactive web interface makes the system accessible to non-technical users.

Future work will focus on expanding the dataset to include more brands and regional pricing variations, incorporating additional features such as service history and accident records, and exploring advanced algorithms such as Gradient Boosting (XGBoost, LightGBM) and deep learning models for further accuracy improvements.

References

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[4] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.

[5] Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology, 4(7), 753–764.

[6] Monburinon, N., Chertchom, P., Kaewkiriya, T., et al. (2018). Prediction of Prices for Used Car by Using Regression Models. 5th International Conference on Business and Industrial Research (ICBIR), 115– 119.

[7] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car Price Prediction using Machine Learning Techniques. TEM Journal, 8(1), 113–118.

[8] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27–31.