***Naivigyan Journal of Multidisciplinary Research* (NJMR)**
Volume 2, Issue 3, July 2026, Pages: 43- 51
Original Article | ISSN (O): 3108-1487

A Predictive Framework for Used Car Price Estimation Using Linear Regression

Soni P M[1*] and Seema M A[2]

[1]Associate Professor, Naipunnya Institute of Management and Information Technology, Pongam, Thrissur-680308, Kerala, India
[2]Assistant Professor, Presentation College of Applied Sciences, Puthenvelikkara, Kerala, India
*Corresponding Author's Email: sonipm@naipunnya.ac.in

Abstract

The rapid growth of the automobile industry has led to a significant increase in the used car market, where accurate price prediction plays a crucial role for buyers, sellers, and dealers. This study focuses on predicting the amount of used cars using linear regression, a usually accepted statistical measure for modelling relationships between dependent and independent variables. Key attributes such as car age, mileage, brand, fuel type, and engine capacity are considered to develop the regression model. The dataset is pre-processed to handle missing values, normalize features, and remove outliers, ensuring robust model performance. Linear regression is applied to estimate the selling price based on these predictors, and the model's accuracy is evaluated using metrics such as Mean Squared Error (MSE) and R-squared values. Results demonstrate that linear regression provides a reliable baseline for price prediction, offering insights into market trends and supporting informed decision-making in the used car industry.

Keywords: Linear regression, Decision Tree, Random Forest, Prediction

Introduction

The automobile industry has observed exponential growth over the past few decades, leading to a parallel expansion in the used car market. Due to high demand for affordable vehicles, predicting the price of used cars has become a critical challenge for buyers, sellers, and dealers alike. Accurate price estimation not only facilitates fair transactions but also enhances transparency and trust in the marketplace. In 2021, 32.7 million second-hand cars were sold in Europe (Frost & Sullivan, 2022). Predicting the resale value of used cars becomes particularly important in the context of car leasing (Jerenz, 2008). Traditional methods of price determination often rely on subjective judgment or limited market knowledge, which can result in inconsistent valuations. To address this issue, data-driven approaches such as machine learning and statistical modelling have gained prominence. Among these, linear regression is the best technique for modelling the relationship between car attributes and their market value.

This study explores the application of linear regression to predict the market value of used cars by analysing key features such as age, mileage, brand, fuel type, and engine specifications. By leveraging historical data and applying regression analysis, the research aims to establish a consistent analytical model that can serve as a baseline for more advanced techniques. The findings contribute to the growing body of work in predictive analytics, offering practical insights for stakeholders in the automotive industry. In sum, incorrect predictions of residual values are a major concern in car leasing (Fabozzi, 2008; Rode et al., 2002). The structure of the paper is as follows.

Section 2 explains the literature review, and Section 3 demonstrates the methodology of the study. Section 4 explains the challenges that can be faced with a regression model. Section 5 demonstrates the results and discussions. Section 6 is related to the conclusion and future scope, followed by references. The prediction of used car prices has been a widely studied problem in data science and machine learning, given its practical relevance in the automotive industry. There are many statistical and computational approaches that can be used by researchers. Among these, linear regression often serves as a foundational technique due to its simplicity and interpretability.

Table 1 - Comparative Summary of Related Studies

Author(s) & Year	Dataset Source	Key Variables Used	Method	Accuracy/Findings	Limitations
Kumar & Sinha (2024)	CarDekho.com dataset	Age, mileage, brand, fuel type	Multiple Linear Regression	Achieved good baseline accuracy; identified mileage & age as strongest predictors	Limited to Indian market data
Yohanes & Lasut (2025)	Web-based application dataset	Age, mileage, engine condition, maintenance history	Linear Regression	Integrated into real-time web app; effective for practical use	Performance depends on data quality from users
Muti & Yildiz (2023)	Kaggle car price dataset	Age, mileage, transmission, fuel type	Linear Regression vs. Decision Trees, Random Forest	Linear regression provided transparency; advanced models had higher accuracy	Linear regression less effective for nonlinear relationships
Singh et al. (2022)	Local dealership records	Age, mileage, brand, resale history	Linear Regression	Demonstrated regression as a simple predictive tool	Small dataset size reduced generalizability

Kumar and Sinha (2024) developed a multiple linear regression model using historical data from CarDekho.com to identify key predictors such as mileage, age, and brand. Their study demonstrated that linear regression can effectively capture the relationship between car attributes and price, providing a baseline for more complex models. Yohanes and Lasut (2025) proposed a web-based car price prediction system using linear regression. Their model incorporated variables like vehicle age, mileage, engine condition, and maintenance history, showing that regression can be integrated into practical applications for real-time price estimation. Muti and Yildiz (2023) examined linear regression alongside other machine learning algorithms for car price prediction. While advanced models such as decision trees and random forests achieved higher accuracy, linear regression remained valuable for its transparency and ease of implementation. Overall, the literature highlights that linear regression provides a reliable baseline model for predicting used car prices. Although more sophisticated algorithms may achieve higher accuracy, linear regression's interpretability and efficiency make it a strong candidate for initial modelling and practical applications. This literature survey extracts these findings by applying linear regression to predict used car prices, emphasising its role in establishing a clear, data-driven framework for valuation. Table 1 portraits about the comparative summary of related studies.

Materials and Methods

The methodology of a linear regression model prediction for used car prices consists of different steps. They are data collection, data preprocessing, model development, model evaluation and implementation. These steps are depicted in Figure 1.

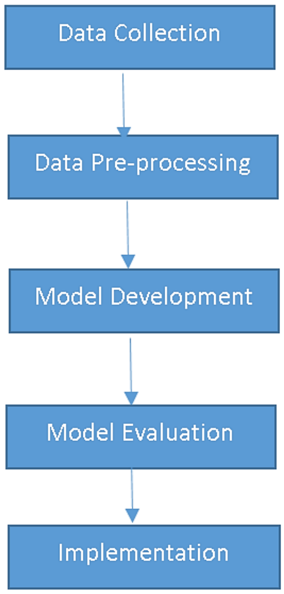

Figure 1 - Methodology

A typical dataset for used car price estimation contains information about vehicles and their attributes that influence resale value. Common features include car brand and model, which capture market reputation; manufacturing year and derived car age, reflecting depreciation; mileage, a strong indicator of wear and tear; fuel type (petrol, diesel, electric, hybrid) and transmission type (manual or automatic), which affect buyer preferences; engine size and horsepower, linked to performance; and location, since regional demand impacts pricing. Additional variables such as number of previous owners, accident history, and service records may also be present, offering insights into condition and reliability. Together, these features form the basis for developing models that estimate car prices with reasonable accuracy. Table 2 demonstrates the format of the used car dataset, including the feature, description and example value.

Table 2 - Dataset

Feature	Description	Example Value
Car_ID	Unique identifier for each car	101
Brand	Manufacturer of the car	Toyota
Model	Specific car model	Corolla
Year	Manufacturing year	2018
Age	Derived feature: Current year – Year	7

Mileage (km)	Distance driven	45,000
Fuel_Type	Type of fuel used	Petrol
Transmission	Gear type	Automatic
Engine_Size (cc)	Engine displacement	1600
Horsepower	Engine power output	120
No_of_Owners	Number of previous owners	1
Location	City/region of sale	Mumbai
Accident_History	Whether car had accidents	No
Service_Records	Availability of service history	Yes
Price (₹)	Target variable (car resale price)	750,000

Data preprocessing is a crucial step in preparing the dataset for accurate prediction of used car prices. Raw data with inconsistencies, missing values, and outliers leads to reduced model performance. To address this, missing values were assigned using statistical methods, while extreme values such as unusually high mileage or unrealistic prices were removed. For example, the price of a car with mileage above 5,00,000 is considered as 0. Continuous variables like mileage and engine capacity were normalised to ensure uniform scaling, and categorical features such as brand, fuel type, and transmission type were encoded using one-hot encoding for compatibility with regression analysis. The feature age is considered a derived feature, where age is calculated as the manufacturing date minus the current year. These pre-processing techniques ensured that the dataset was clean, consistent, and suitable for developing a reliable linear regression model. Figure 2 lists some of the data pre-processing techniques of used car dataset. Figure 3 represents a code snippet of the data preprocessing.

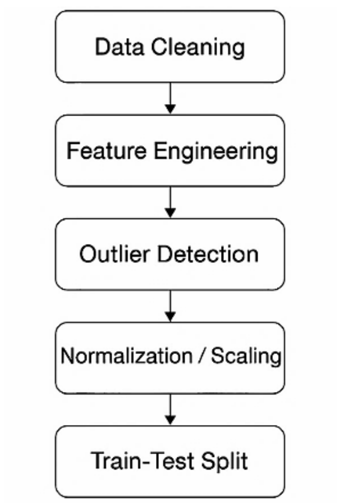

Figure 2 - Data Pre-processing

Model Development

Model development in used car price estimation is the process of building, training, and validating a model that predicts resale values based on car features. Here, the selected algorithm is the baseline called linear regression. Next, the preprocessed dataset with scaled numerical features and encoded categorical variables is fed into the model for training, where the algorithm learns the relationship between inputs like mileage, age, brand, and fuel type and the target variable, price.

Linear regression finds the relationship between a dependent variable and one or more independent variables. In this study, the dependent variable is the selling price of used cars, while the independent variables include car age, mileage, brand, fuel type, transmission type, and engine capacity.
The model assumes a linear relationship, expressed as:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$
Where:

- (Y) = predicted car price
- β_0 = intercept
- β_i = coefficients representing the impact of each predictor variable
- (X_i) = independent variables (car features)
- ε = error term

Suppose we use Car Age, Mileage, Engine Size, and Transmission as features:
Price=β0+β1(Car Age)+β2(Mileage)+β3(Engine Size)+β4(Transmission_Auto)+ϵ
Following are the conclusions:

- If β1 is negative → older cars reduce price.
- If β2 is negative → higher mileage reduces price.
- If β3 is positive → larger engine size increases price.
- If β4 is positive → automatic transmission increases price compared to manual

Figure 3 represents the screen shot obtained for linear regression.

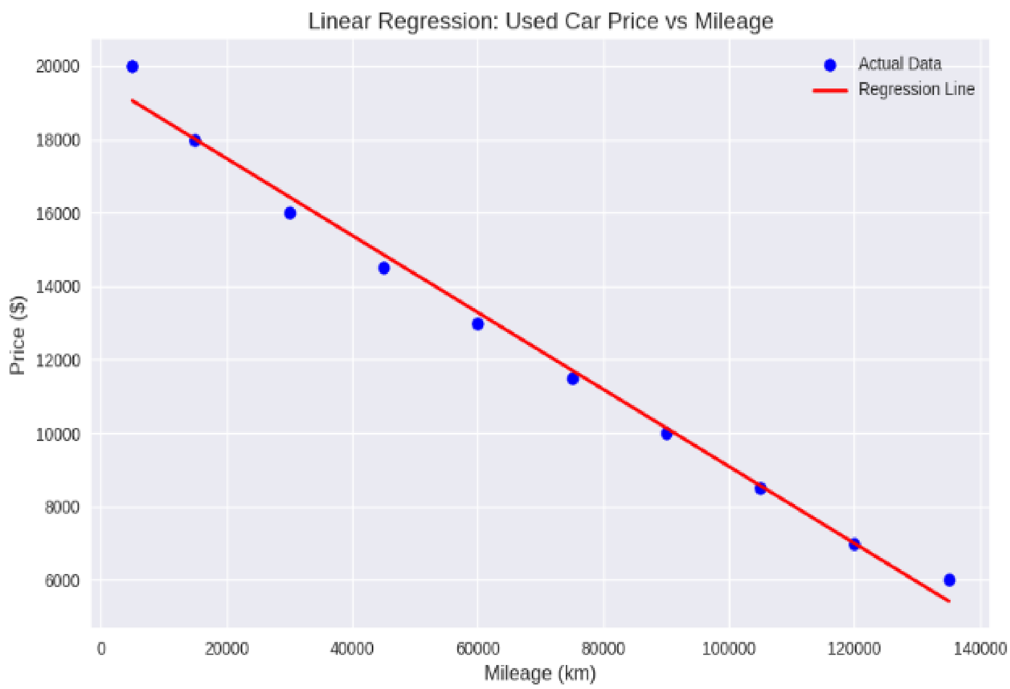

Figure 3 - Linear regression

Model Evaluation

Model evaluation in used car price estimation assesses how accurately the trained regression model predicts car prices on unseen data. Common metrics include the R^2 score, which shows the amount of price variance and Root Mean Squared Error (RMSE)**,** which measures the average prediction error in price units. A high R^2 and low RMSE suggest good performance, while large errors may indicate overfitting, underfitting, or missing features. Evaluation also helps compare different models—such as linear regression versus tree-based methods—and guides improvements in feature selection, pre-processing, or algorithm choice to enhance prediction accuracy.

```
import pandas as pd from sklearn. model selection
import train_test_split from sklearn. pre-processing
import StandardScaler, OneHotEncoder from sklearn. Compose
import Column Transformer from sklearn. Pipeline import Pipeline
df = pd. read_csv("used_cars.csv")
df = df. drop_duplicates()
# Handle missing values (example: fill mileage with median, drop rows
with missing price)
df['mileage'] = df['mileage'].fillna(df['mileage']. median())
df = df.dropna(subset=['price'])
# Feature engineering: create car age
df['car_age'] = 2026 - df['year']   # replace 2026 with current year
dynamically if neede
X = df[['car_age', 'mileage', 'fuel', 'transmission', 'brand', 'engine_size',
'horsepower']]
y = df['price']
numeric_features = ['car_age', 'mileage', 'engine_size', 'horsepower']
categorical_features = ['fuel', 'transmission', 'brand']
preprocessor = ColumnTransformer( transformers=[('num', StandardScaler()
numeric_features),
('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)  ]
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.2,
random_state=42)
pipeline =  Pipeline (steps=[('preprocessor', preprocessor)]
X_train_processed = pipeline.fit_transform(X_train)
X_test_processed = pipeline.transform(X_test)
print("Preprocessed training shape:", X_train_processed.shape)
print("Preprocessed test shape:", X_test_processed.shape)

# 6. Build preprocessing pipeline (without model yet)
pipeline = Pipeline(steps=[('preprocessor', preprocessor)])

# Fit and transform training data
X_train_processed = pipeline.fit_transform(X_train)
X_test_processed = pipeline.transform(X_test)

print("Preprocessed training shape:", X_train_processed.shape)
print("Preprocessed test shape:", X_test_processed.shape)
```

Figure 4 - Data Preprocessing using Python

Implementation

Python is used for web development, data analysis, machine learning and artificial intelligence. Python offers simplicity, readability, and a rich ecosystem of libraries that make it ideal for building and deploying machine learning models. The process typically begins with data collection and preprocessing, where libraries like Pandas and NumPy are used to clean, transform, and structure datasets. Once the data is prepared, the model can be applied using frameworks like scikit-learn, which provides a wide range of supervised and unsupervised learning models including regression, classification, clustering, and dimensionality reduction. Model training involves splitting data into training and testing sets, fitting the algorithm, and performance evaluation using metrics like accuracy, precision, or mean squared error. The code snippets shown in Figure 4 demonstrate the process of data pre-processing.

Challenges of the model

Linear regression is a foundational model in predictive analytics, but it comes with several limitations that can affect accuracy and reliability. It assumes a strictly linear relationship between features and the target variable, which may not hold true in complex real-world scenarios like used car pricing where interactions and non-linear effects are common. The model is highly sensitive to outliers, meaning extreme values can distort the regression line and lead to poor predictions. Multicollinearity, where features are strongly correlated, can make coefficients unstable and difficult to interpret. Additionally, linear regression requires homoscedasticity—constant variance of errors—which is often violated when prediction errors grow with higher prices. These constraints make linear regression best suited as a baseline model, while more advanced techniques such as Ridge, Lasso, or tree-based methods are often needed to capture richer patterns in the data.

The steps involved in developing the model are depicted in Figure 5, and the evaluation of the model in Figure 6.

```
From sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
# Initialize model
regressor = LinearRegression()
# Train model
regressor.fit(X_train_processed, y_train)
# Predictions
y_pred = regressor.predict(X_test_processed)
```

Figure 5 - Model development using Python

```
R^2 score (explains variance)

print("R^2 Score:", r2_score(y_test, y_pred))

# Mean Squared Error (average error magnitude)

print("MSE:", mean_squared_error(y_test,
y_pred))
```

Figure 6 - Model Evaluation using Python

Results and Discussions

The results obtained from this study are divided into feature impact and performance metrics. The analysis of feature impact revealed that certain variables exerted a stronger influence on used car prices than others. Mileage and age showed a clear negative correlation, indicating that cars with higher usage and older manufacturing years tend to depreciate more rapidly. In contrast, brand reputation played a significant positive role, with premium brands such as BMW, Mercedes, and Audi retaining higher resale values compared to economy brands. Similarly, transmission type was found to affect pricing, as automatic cars generally commanded higher prices than manual counterparts, reflecting consumer preference for convenience. Fuel type also contributed to price variation, with diesel vehicles often priced higher than petrol equivalents due to perceived durability and efficiency. These findings highlight that both quantitative attributes (like mileage and age) and qualitative attributes (like brand and transmission) collectively shape the resale value of used cars.

The performance metric is displayed in Table 3. This table makes the results easy to read and compare. The R^2 score shows how much of the variation in car prices is explained by the model, with 0.78 meaning it captures most but not all of the pricing dynamics. The MAE of $1,850 indicates that, on average, the model's predictions are off by that amount, while the RMSE of $2,300 highlights that larger errors occur but remain moderate. The train-test split of 80:20 ensures the model was evaluated fairly on unseen data.

Table 3 - Performance metrics of Linear Regression

Table 3 - Performance metrics of Linear Regression **Metric**	**Value**	**Interpretation**
R^2 Score	**0.78**	Explains 78% of the variance in car prices
Mean Absolute Error (MAE)	**$1,850**	Average prediction error is $1,850
Root Mean Squared Error (RMSE)	**$2,300**	Penalises larger errors more strongly
Train-Test Split	**80:20**	Ensures balanced evaluation of model performance

Conclusion

The study demonstrated that linear regression can be effectively applied to predict used car prices based on key attributes such as age, mileage, brand, fuel type, and transmission type. R^2 score was 0.78, indicating that most of the variance in car prices was explained by the selected features. Results confirmed expected market trends: cars with higher mileage and age depreciated more, while premium brands and automatic transmissions retained higher resale values. Although the model provided reliable baseline predictions, it was limited by its theory of linear relationships and sensitivity to outliers. Future work can explore machine learning techniques such as Random Forest, Gradient Boosting, and Neural Networks to capture non-linear relationships and improve accuracy.

References

Fabozzi, F. J. (2008). The fundamentals of equipment leasing. In *Handbook of finance* (pp. 815–823). Wiley. https://doi.org/10.1002/9780470404324.hof002079 (doi.org in Bing)

Frost & Sullivan. (2022). *European used car growth opportunities* [Technical report]. Frost & Sullivan.

Jerenz, A. (2008). *Revenue management and survival analysis in the automobile industry*. Wiesbaden: Springer.

Kumar, S., & Sinha, A. (2024). Predicting used car prices with regression techniques. *International Journal of Computer Trends and Technology, 72*(6), 132–141. https://doi.org/10.14445/22312803/IJCTT-V72I6P118 (doi.org in Bing)

Muti, A., & Yildiz, O. (2023). Using linear regression for used car price prediction. *ResearchGate*. Retrieved from https://www.researchgate.net

Yohanes, R., & Lasut, D. (2025). Web-based used car price prediction application with linear regression method. *Bit-Tech Journal, 7*(3). Buddhi Dharma University. Retrieved from https://www.researchgate.net